Université de Montréal

Traitement neuronal des voix et familiarité :
Entre reconnaissance et identification du locuteur

Par

Julien Plante-Hébert

Département de linguistique et de traduction, Faculté des arts et des sciences

Thèse présentée en vue de l'obtention du grade de *Philosophiae doctor* (Ph. D.)

en linguistique, option neuropsychologie

Décembre 2020

Université de Montréal

Département de linguistique et de traduction, Faculté des arts et des sciences

---

Cette thèse intitulée

Traitement neuronal des voix et familiarité

Entre reconnaissance et identification du locuteur

Présenté par

Julien Plante-Hébert

A été évalué(e) par un jury composé des personnes suivantes

Simone Falk
Président-rapporteur

Victor J. Boucher
Directeur de recherche

Boutheina Jemel
Codirectrice

Paul Foulkes
Examinateur externe

Ingrid Verduyckt
Membre du jury

Résumé

La capacité humaine de reconnaitre et d'identifier de nombreux individus uniquement grâce à leur voix est unique et peut s'avérer cruciale pour certaines enquêtes. La méconnaissance de cette capacité jette cependant de l'ombre sur les applications dites « légales » de la phonétique. Le travail de thèse présenté ici a comme objectif principal de mieux définir les différents processus liés au traitement des voix dans le cerveau et les paramètres affectant ce traitement.

Dans une première expérience, les potentiels évoqués (PÉs) ont été utilisés pour démontrer que les voix intimement familières sont traitées différemment des voix inconnues, même si ces dernières sont fréquemment répétées. Cette expérience a également permis de mieux définir les notions de reconnaissance et d'identification de la voix et les processus qui leur sont associés (respectivement les composantes P2 et LPC). Aussi, une distinction importante entre la reconnaissance de voix intimement familières (P2) et inconnues, mais répétées (N250) a été observée.

En plus d'apporter des clarifications terminologiques plus-que-nécessaires, cette première étude est la première à distinguer clairement la reconnaissance et l'identification de locuteurs en termes de PÉs. Cette contribution est majeure, tout particulièrement en ce qui a trait aux applications légales qu'elle recèle.

Une seconde expérience s'est concentrée sur l'effet des modalités d'apprentissage sur l'identification de voix apprises. Plus spécifiquement, les PÉs ont été analysés suite à la présentation de voix apprises à l'aide des modalités auditive, audiovisuelle et audiovisuelle interactive. Si les mêmes composantes (P2 et LPC) ont été observées pour les trois conditions d'apprentissage, l'étendue de ces réponses variait. L'analyse des composantes impliquées a révélé un « effet d'ombrage du visage » (*face overshadowing effect*, FOE) tel qu'illustré par une réponse atténuée suite à la présentation de voix apprise à l'aide d'information audiovisuelle par rapport celles apprises avec dans la condition audio seulement. La simulation d'interaction à l'apprentissage à quant à elle provoqué une réponse plus importante sur la LPC en comparaison avec la condition audiovisuelle passive.

De manière générale, les données rapportées dans les expériences 1 et 2 sont congruentes et indiquent que la P2 et la LPC sont des marqueurs fiables des processus de reconnaissance et d'identification de locuteurs. Les implications fondamentales et en phonétique légale seront discutées.

Mots-clés : Identification du locuteur, identification de la voix, reconnaissance du locuteur, perception multimodale, apprentissage de la voix, acoustique de la voix, potentiels évoqués (PÉ), P2, N250, LPC, phonétique légale

Abstract

The human ability to recognize and identify speakers by their voices is unique and can be critical in criminal investigations. However, the lack of knowledge on the working of this capacity overshadows its application in the field of “forensic phonetics”. The main objective of this thesis is to characterize the processing of voices in the human brain and the parameters that influence it.

In a first experiment, event related potentials (ERPs) were used to establish that intimately familiar voices are processed differently from unknown voices, even when the latter are repeated. This experiment also served to establish a clear distinction between neural components of speaker recognition and identification supported by corresponding ERP components (respectively the P2 and the LPC). An essential contrast between the processes underlying the recognition of intimately familiar voices (P2) and that of unknown but previously heard voices (N250) was also observed.

In addition to clarifying the terminology of voice processing, the first study in this thesis is the first to unambiguously distinguish between speaker recognition and identification in terms of ERPs. This contribution is major, especially when it comes to applications of voice processing in forensic phonetics.

A second experiment focused more specifically on the effects of learning modalities on later speaker identification. ERPs to trained voices were analysed along with behavioral responses of speaker identification following a learning phase where participants were trained on voices in three modalities : audio only, audiovisual and audiovisual interactive.

Although the ERP responses for the trained voices showed effects on the same components (P2 and LPC) across the three training conditions, the range of these responses varied. The analysis of these components first revealed a *face overshadowing effect* (FOE) resulting in an impaired encoding of voice information. This well documented effect resulted in a smaller LPC for the audiovisual condition compared to the audio only condition. However, effects of the audiovisual

interactive condition appeared to minimize this FOE when compared to the passive audiovisual condition.

Overall, the data presented in both experiments is generally congruent and indicate that the P2 and the LPC are reliable electrophysiological markers of speaker recognition and identification. The implications of these findings for current voice processing models and for the field of forensic phonetics are discussed.

Keywords : Speaker identification, voice identification, speaker recognition, multimodal perception, voice learning, voice acoustics, speech acoustics, event-realted potentials (ERP), P2, N250, LPC, multimodal, forensic phonetics

Table des matières

Liste des tableaux

Notez que certaines légendes sont en anglais. Celles-ci proviennent des articles composant le cœur de la thèse. Ces articles ont été rédigé en anglais en vue de publication.

Liste des figures

Liste des sigles et abréviations

Notez que certains sigles proviennent de termes en anglais. Les versions originales anglaises sont entre parenthèses.

A :	Audio
AV :	Audiovisuel
AVI :	Audiovisuel interactif
CIUSSS :	Centre intégré universitaire de santé et de services sociaux
dB(A) :	Décibel (*A-weighing*)
EEG :	Électroencéphalographie
ERP :	*Event-related potential*
F_0(mp) :	Fréquence fondamentale (moyenne parlée)
FQ:	Français québécois
FOE :	*Face overshadowing effect*
FRU :	Unité de reconnaissance des visages *(Face recognition unit*)
FV :	*Familiar voice*
Hz :	Hertz
IAC :	*Interactive activation and competition*
IFV:	*Intimately familiar voice*
IRMf:	Imagerie par résonance magnétique fonctionnelle
ISI :	Intervalle inter-stimuli (*Inter-stimuli interval*)

LCF : *Left centro-frontal*

LCP : *Left centro-parietal*

LPC : Composante positive tardive (*Late positive component*)

MCF : *Middle centro-fontal*

MCP : *Middle centro-parietal*

MEG : Magnétoencéphalographie

MMN : *Missmatch negativity*

ms : Milliseconde

NRU : Unité de reconnaissance des noms (*Name recognition unit*)

PÉ : Potentiel évoqué

PIN : Foyer d'indentification individuel (*Personnal identity node*)

RCF : *Right centro-frontal*

RCP : *Right centro-parietal*

RT : *Reaction time*

s : Seconde

SF_0 : *Speaking fundamental frequency*

GTS: Gyrus temporal supérieur (*Superior temporal gyrus*)

STS : Sulcus temporal supérieur

syll. : Syllabe

TEP : Tomographie par émission de positrons

TR : Temps de réaction

TV : Trained voice

TVA : Aire temporale de la voix (*Temporal voice area*)

UV : *Unknown voice*

VE : Voix entrainée

VF : Voix familière

VR : Voix rare

VRU : Unités de reconnaissance de la voix (*Voice recognition unit*)

µV : Microvolt

σ : Écart type

Ω: Ohm

À Jasmine (1985-2007),

La route nous a fait prendre quelques détours

Les chemins qu'on a pris n'ont pas toujours été les meilleurs

Mais on s'est rendu là où tu voulais qu'on soit

Je te l'avais promis

Tu serais fière

Remerciements

J’aimerais commencer par remercier la personne qui a sans doute vécu le plus intensément ces années de doctorat avec moi, mon conjoint Jean-Pierre. Merci pour ton support, tes encouragements et ta si grande fierté lorsque tu expliques à ma place ce sur quoi portent mes travaux.

Je veux aussi remercier ma famille, qui a été présente à toutes les étapes de ce parcours, les plus belles comme les plus difficiles. Merci de m’avoir toujours encouragé dans les chemins que j’ai choisis, aussi singuliers puissent-ils être !

À mes amis, qui ont suivi de près mes péripéties sans toujours comprendre ce que je faisais de mes journées, je pense particulièrement à Caro, Eva, Nathalie et Kadi, mais aussi à tous les autres que je ne peux nommer individuellement, merci pour vos oreilles et vos épaules. Parmi ceux-ci, une mention spéciale est de mise pour souligner tous ceux et celles qui ont joué le jeu en acceptant de me prêter votre voix, vos oreilles et vos neurones pour mes nombreux tests de toutes sortes. Vous faites partie intégrante de cette thèse.

Les derniers miles de cette aventure se sont déroulés dans un contexte de confinement sans précédent. Je tiens à remercier tous mes collègues passionnés du langage avec qui j’ai développé un réseau de motivation et de soutien virtuel. Cette aide s’est avérée essentielle et je vous en suis fort reconnaissant. Lâchez pas !

Un clin d’œil à Patrick Drouin et François Lareau, avec qui j’ai partagé un bout de corridor et plusieurs comités. Votre humour est un vent de fraîcheur dans le département, ne le perdez jamais !

Finalement, je veux remercier mes directeurs, sans qui ce projet n’aurait sans doute pas été possible. Merci à vous de m’avoir transmis tant de connaissances.

Chapitre 1 : Introduction

Brèves onto- et phylogenèses

La capacité qu'a l'humain d'identifier les individus qui l'entourent par leur voix est à la fois anodine, puisqu'utilisée au quotidien par tout un chacun, et exceptionnelle dans sa précision et sa complexité. Malgré son caractère commun, ses mécanismes demeurent méconnus et les paramètres qui la régissent flous. Ces connaissances limitées sont entre autres expliquées par l'intérêt scientifique relativement récent qui y est porté ainsi que par les comparaisons parfois hasardeuses faites entre le traitement de l'identité des individus par leur visage et par leur voix.

Pourtant, cette capacité de traiter avec précision les voix entendues est cruciale chez l'être humain dès sa naissance. Plusieurs études ont en effet rapporté des réactions spécifiques à la voix de la mère chez les nouveau-nés, mais également chez des fœtus en stade avancé de développement (DeCasper et Fifer, 1980; deRegnier, Nelson, Thomas, Wewerka et Georgieff, 2000; Hepper, Scott et Shahidullah, 1993; Kisilevsky et al., 2009; Kisilevsky et al., 2003; Lee et Kisilevsky, 2014; Mehler, Bertoncini, Barriere et Jassik-Gerschenfeld, 1978; Moon et Fifer, 1990). Comme pour plusieurs autres espèces, la présence de cette capacité dès la naissance est entre autres nécessaire à la formation d'un lien entre la mère et sa progéniture, lui-même nécessaire à la survie des nourrissons (Locke et Bogin, 2006; Sidtis et Kreiman, 2012).

Comme le soulignent aussi Sidtis et Kreiman (2012), la capacité de l'être humain de distinguer les voix familières des voix inconnues a également mené au développement d'un sentiment d'appartenance à un groupe, à la différenciation entre les membres de ce groupe et ceux de groupes rivaux et à la sélection de partenaires aptes à la reproduction.

Reconnaitre ses proches à l'aide de leur voix s'avère ainsi avoir contribué de plus d'une manière à la survie et à l'émancipation de l'espèce humaine. En dehors de ces bénéfices sur le plan de l'évolution, l'étude du traitement des voix par le cerveau humain possède un certain nombre d'applications plus contemporaines dont les principales feront l'objet des prochaines sections.

Phonétique légale

L'utilisation des technologies de communication que sont le téléphone et la radio a connu un essor planétaire au courant du 20e siècle. Il va de soi que le traitement de l'information auditive par le cerveau humain, entre autres sa capacité de reconnaitre et identifier les locuteurs, s'est avéré essentiel dans l'utilisation de ces technologies. Les enregistrements acoustiques de toutes sortes qui ont, quant à eux, transformé l'étude de la parole, sont aussi venus avec ces avancements. La phonétique légale, c'est-à-dire l'application de la phonétique au domaine de l'enquête et du droit, figure parmi les domaines d'étude qui ont émergé à la suite de l'avènement de ces technologies. Bien que les premières études scientifiques publiées au sujet de l'identification d'individus par la voix dans un contexte légal remontent à il y a près d'un siècle (voir McGehee, 1937, 1944), ce n'est qu'au tournant des années 1980 que le domaine de la phonétique légale s'est établi en tant que discipline.

La phonétique légale, en tant que domaine d'application, regroupe un nombre d'expertises liées à la parole. Certaines de ces spécialisations, comme l'amélioration de la qualité du signal ou l'authentification de l'intégralité d'enregistrements, sont d'ordre plus acoustique. D'autre, telles que la retranscription de passages litigieux, l'authentification d'accent ou de dialectes et le profilage de locuteurs, relèvent plus directement de la phonétique. C'est également le cas pour les expertises liées à la reconnaissance et à l'identification de locuteurs.

Réussir à démontrer que l'identification d'un individu par sa voix est valable demeure un problème épineux lorsqu'il est question d'une enquête ou d'un procès. Comme un certain nombre de revues de littérature à ce sujet au Canada et aux États-Unis le démontre, l'identification de suspects par leur voix est utilisée en cour depuis longtemps sans pour autant que les procédures et recommandations qui l'entourent ne soient respectées, voire même connues (Clifford, 1980; Laub, Wylie et Bornstein, 2013; Solan et Tiersma, 2002; Yarmey, 2014). Comme le souligne pourtant Yarmey (2014), il est d'un commun accord dans la communauté scientifique spécialisée, que l'identification d'individu par la voix est à prendre avec grandes précautions.

Un des problèmes particulièrement épineux que pose la voix dans une perspective biométrique est qu'elle varie sur de nombreux aspects chez un même individu. On peut souvent remarquer des différences marquées dans la voix d'une même personne selon le moment de la journée, l'activité en cours, son humeur ou encore son état de santé. Réussir à cerner ce qui est spécifique à une voix devient donc complexe et ne peut se résumer à une mesure unique ou à un ensemble de mesure fixes.

Il faut aussi souligner que l'identification de locuteurs dans un contexte légal peut s'avérer bien différente de celle dont nous faisons l'expérience au quotidien. Dans la vie de tous les jours, un certain nombre d'indices contextuels facilitent l'identification d'un individu par sa voix. On peut penser à des indices plus explicites, tels que l'affichage de l'appelant sur un téléphone, mais aussi à d'autres, bien plus subtils. Si quelqu'un marche dans les corridors de son lieu de travail, il sera par exemple plus enclin à identifier la voix d'un superviseur ou d'un collègue de travail. L'environnement joue ici le rôle d'un indice contextuel favorisant l'identification de certaines voix et non d'autres. Dans un contexte légal, il peut cependant n'y avoir aucun indice contextuel permettant de faciliter l'identification d'un locuteur, ce qui peut complexifier la tâche puisqu'elle ne relève alors que de l'information acoustique.

Deux approches bien distinctes sont présentes pour tenter de résoudre le problème de l'identification de locuteurs dans un cadre légal. Dans un premier temps, il y a les techniques automatisées, en constante évolution depuis les années 1970. Bien qu'en raison de leur potentiel commercial un grand nombre d'études y soit consacré, ces techniques ne sont pas l'objet du présent travail et ne seront par conséquent pas décrites en détail. La seconde approche, qui nous intéresse davantage, est celle qui implique un identificateur humain. Au Royaume-Uni, où le système juridique en permettent plus couramment l'usage, on qualifie généralement cette approche d'« auditive-perceptuelle » *(aural-perceptual approach; Hollien, 1990)*. Au sein même de cette approche, un nombre de guides de recommandations a été publié par des experts et chercheurs. Il est généralement admis que le recours à la technique de parade vocale (*voice line-ups*) est à privilégier par rapport aux techniques n'impliquant qu'un seul locuteur (Broeders et van Amelsvoort, 1999; de Jong-Lendle, Nolan, McDougall et Hudson, 2015; Hollien, Huntley Bahr et Harnsberger, 2014; Hollien, Huntley Bahr, Künzel et Hollien, 1995; Jessen, 2008; Nolan, 2003;

Nolan et Grabe, 1996; Yarmey, 2014). Malgré ces recommandations, les nombreuses études portant sur l'identification auditive-perceptuelle du locuteur ont mené à des résultats très variables quant à la capacité de l'être humain d'identifier un individu uniquement par sa voix. À cet effet, les travaux de Yarmey (2014), Atkinson (2015) et Braun (2016) mettent en évidence les nombreuses variables susceptibles d'influencer les performances d'un individu ou d'un groupe d'individus lors d'une tâche de reconnaissance ou d'identification du locuteur. Celles-ci peuvent être regroupées en fonction de leur caractère interne, c'est-à-dire propre aux individus auxquels incombe la tâche de reconnaissance ou d'identification, ou externe, en lien avec le matériel présenté.

Variables externes

La qualité du signal acoustique comprenant les échantillons de voix utilisés affecte sans grande surprise les performances des participants. Par exemple, un certain nombre d'études s'est penché sur les effets de la variation du signal acoustique due à la transmission téléphonique, que ce soit en raison des appareils eux-mêmes ou des réseaux. Sans entrer dans les détails des différentes altérations d'ordre acoustique rapportées dans ces études, il importe néanmoins de vérifier si elles affectent les performances aux tâches de reconnaissance et d'identification du locuteur. L'étude de Nolan, McDougall et Hudson (2013) suggère par exemple que les voix semblent plus similaires aux oreilles des participants lorsque la voix entendue est transmise par voie téléphonique. Cet effet s'explique entre autres par la largeur de bande limitée des transmissions téléphoniques qui engendrent la perte d'une certaine quantité d'information acoustique qui pourrait s'avérer essentielle à la distinction d'une voix par rapport à d'autres. Les auteurs rapportent cependant que l'effet négatif de l'altération du signal acoustique par la transmission téléphonique est observé pour certaines voix plus que d'autres. Autrement, les études de Öhman, Eriksson et Granhag (2010); Perfect, Hunt et Harris (2002) et Kerstholt, Jansen, van Amelsvoort et Broeders (2006) ne rapportent pas d'effet significatif d'un signal altéré par la communication téléphonique. En raison de l'amélioration des technologies, entre autres dans le milieu des communications, il y a lieu de croire que si aucun effet n'a été rapporté à l'époque de ces études, de nouvelles expériences seraient peu susceptibles de présenter des résultats contradictoires.

Dans le même ordre d'idées, la qualité de la voix dans les échantillons utilisés doit aussi être prise en compte, tout particulièrement dans un contexte d'applications légales. À la différence de la qualité du signal, la qualité de la voix fait référence aux altérations vocales produites par le locuteur lui-même. On parle ici par exemple de chuchotement, d'imitation, de déguisement vocal ou encore de parole forte ou criée. Comme le rappelle Braun (2016, p. 54), le Bureau de la police criminelle fédéral allemand (*Bundeskriminalamt*) indique que près du quart des cas impliquant une expertise en phonétique légale répertoriés en Allemagne implique une forme de déguisement de la voix. À cet effet, l'étude de Hollien, Majewski et Doherty (1982) fait état de scores de réussite significativement plus faibles à une tâche d'identification du locuteur lorsque la voix de ces derniers était déguisée (notons que le déguisement vocal était au libre choix des locuteurs). Comme l'explique Braun (2016, p. 55), le chuchotement est l'un des déguisements vocaux les plus simples, mais aussi les plus efficaces. Les études sur l'identification de locuteurs d'Orchard et Yarmey (1995), de Yarmey (2001) et de Smith, Foulkes et Sóskuthy (2017) rapportent d'ailleurs toutes trois des résultats significativement supérieurs lorsque la tâche d'identification implique des voix en conversation normale par rapport à des voix chuchotées. De leur côté, Blatchford et Foulkes (2007) et Brungart, Scott et Simpson (2001) se sont intéressés à la voix criée par rapport à la voix normale ou à la voix chuchotée. Dans un premier temps, les données de Brungart et al. (2001) soutiennent qu'un locuteur est significativement moins bien identifié en parole chuchotée qu'en parole normale, mais, au contraire, mieux identifiée en parole criée qu'en parole normale. Les auteurs rapportent aussi que les taux de succès étaient significativement plus faibles lorsque la qualité de la voix (chuchotée, normale ou criée) était différente entre l'entrainement et la tâche expérimentale. L'étude de Blatchford et Foulkes (2007) met quant à elle en évidence que la voix criée peut être identifiable avec d'assez hauts taux de réussite, mais qu'elle est cependant très sensible à plusieurs autres facteurs tels que la durée des échantillons vocaux présentés et l'aptitude individuelle à identifier les voix familières.

Il va de soi que la longueur des échantillons de voix présentés a un effet sur la capacité de reconnaitre ou d'identifier le locuteur. Cependant, il n'y a pas de consensus sur la durée minimalement requise pour favoriser les performances ni sur l'unité de mesure de cette longueur. Les conclusions de certaines études mettent de l'avant qu'une durée plus grande, mesurée en

secondes, permet un meilleur rappel (Cook et Wilding, 1997b; Legge, Grosmann et Pieper, 1984; Orchard et Yarmey, 1995; Yarmey, 1991). D'autres études avancent que cet effet observé lors d'une variation dans la durée n'est que le reflet d'un effet véritable de la variation concomitante du nombre de configurations articulatoires contenues dans l'échantillon sonore (Bricker et Pruzansky, 1966; Plante-Hébert et Boucher, 2014; Pollack, Pickett et Sumby, 1954; Roebuck et Wilding, 1993). En d'autres termes, plus un extrait de parole est long, plus il est probable que le nombre de phonèmes produits soit grand, ce qui procure une plus grande quantité d'information spectrodynamique reflétant les propriétés des cavités de résonances propres à un locuteur spécifique. En appui à cette perspective, les résultats de certaines études portant spécifiquement sur le rôle que joue l'utilisation des cavités nasales dans la reconnaissance et l'identification de locuteurs ont d'ailleurs observé de meilleures performances lorsque ces cavités étaient impliquées (Amino et Arai, 2009; Plante-Hébert et Boucher, 2015a; Su, Li et Fu, 1974). Il est important de souligner ici que ces études n'indiquent pas que les cavités nasales sont plus pertinentes que les cavités de résonances orales lorsqu'il est question d'identifier un locuteur. Elles soulignent plutôt la valeur ajoutée de la présence de sons nasaux dans de telles situations. Comme la voix d'un individu résulte entre autres de sa physiologie, l'utilisation de cavités de résonances supplémentaires vient compléter le portrait acoustique déjà entamé par les autres cavités et structures impliquées dans la production de la parole. Pour en revenir à la longueur des échantillons de voix utilisés, il demeure difficile d'asseoir avec certitude une « longueur » clé nécessaire à l'identification de locuteurs familiers. Les études indiquent cependant que plus d'une syllabe serait nécessaire afin que les mouvements articulatoires génèrent un minimum d'information spectrodynamique.

Le caractère distinctif d'une voix par rapport aux autres, bien que son effet sur la reconnaissance et l'identification de locuteurs demeure encore peu étudié, nécessite tout de même une attention particulière. À l'exception de l'étude de Schmidt-Nielsen et Stern (1985), la grande majorité des études qui ont observé l'impact de cette variable soutiennent que plus une voix est distinctive, c'est-à-dire que plus elle détonne par rapport à l'ensemble, plus elle est facile à reconnaitre ou à identifier (Mullennix, Pisoni et Martin, 1989; Orchard et Yarmey, 1995; Papcun, Kreiman et Davis, 1989; Skuk et Schweinberger, 2013; Stevenage, Neil, Parsons et Humphreys, 2018). Dans ces

études, les voix présentées aux participants étaient préalablement classées comme étant plus ou moins distinctives en fonction du jugement de volontaires. Dans une autre étude, menée par Foulkes et Barron (2000), les auteurs se sont intéressés aux caractéristiques acoustiques corrélées avec une plus ou moins grande propension d'une voix à être reconnue à la suite des résultats d'une tâche d'identification de locuteurs intimement familiers. C'est plus précisément en observant la moyenne de fréquence fondamentale parlée (F_0mp) de chaque participant et les écarts à la moyenne que les auteurs ont observé que les voix les mieux reconnues étaient plus distantes de la moyenne. Sørensen (2012) rapporte des résultats similaires en ayant également mesuré le caractère distinctif des voix en fonction de la F_0mp. Bien que d'autres facteurs acoustiques puissent, selon toute vraisemblance, entrer en ligne de compte dans la caractérisation des voix distinctives, la seule mesure objective à cet effet rapportée à ce jour demeure ainsi la F_0.

Malgré les différentes variables présentées ci-dessus, donc certaines portent explicitement sur les voix présentées, il importe de souligner qu'il est ardu, voire impossible, de cibler une ou un groupe restreint de caractéristiques vocales qui permettent à un individu d'en reconnaitre un autre par la voix. Dans un premier temps, chez un même identificateur, les caractéristiques utiles à l'identification d'une voix ne sont pas fixes, elles sont appelées à varier en fonction du locuteur et des caractéristiques qui y sont propres. À l'inverse, pour un même locuteur à identifier, les caractéristiques vocales utiles à un identificateur ne sont pas forcément les mêmes que pour un autre identificateur. Il faut ainsi tenir compte des propriétés spécifiques à voix, mais également à chaque système auditif.

Variables internes

Parmi les variables internes, l'âge des participants est incontournable. Les résultats rapportés par Mann, Diamond et Carey (1979) démontrent effectivement que les résultats obtenus à la suite d'une tâche de reconnaissance de locuteurs inconnus varient grandement avant l'âge de 14 ans. Dans une étude plus récente, Calderwood, McKay et Stevenage (2019) rapportent des résultats qui sont similaires à ceux d'adultes dès l'âge de 8 ou 9 ans, mais beaucoup plus faibles chez les enfants âgés de 6-7 ans. Les résultats de Levi (2018) mettent eux aussi en évidence une

amélioration significative avec l'âge en comparant les résultats à une tâche de discrimination de voix chez des enfants âgés entre 6 et 8 ans et entre 10 et 12 ans. À l'autre extrémité du spectre, les personnes âgées sont également plus susceptibles de produire de faibles résultats à la reconnaissance ou à l'identification de locuteurs. En dehors de la prépondérance de pertes auditive au sein de cette population, Johnson, De Leonardis, Hashtroudi et Ferguson (1995) ont aussi démontré qu'il était plus ardu pour des personnes âgées d'associer des propos verbaux à leur auteur. Zäske et al. (2018) rapportent quant à eux que non seulement les adultes plus jeunes performent mieux à la reconnaissance de locuteurs, mais aussi que les voix de personnes âgées sont généralement plus distinctives.

Des résultats variables ont été rapportés dans différentes études au sujet de différences de performance aux tâches de reconnaissance du locuteur en regard du sexe des participants. Si certaines études auprès d'adultes ou d'enfants n'ont rapporté aucun effet de genre sur la performance (Bartholomeus, 1973; Thompson, 1985), d'autres ont relevé un biais favorable pour la reconnaissance de voix du même sexe chez les hommes (Skuk et Schweinberger, 2013) ou chez les femmes (Roebuck et Wilding, 1993; Wilding et Cook, 2000).

Comme le soulignent Kreiman et Sidtis (2011, p. 150), les études traitant de l'ethnicité ont non seulement rapporté des résultats variés par rapport à la reconnaissance de locuteurs, mais ces différences sont également souvent liées à des caractéristiques verbales apprises plutôt qu'à des variations physiologiques. La capacité accrue d'un individu de reconnaitre des locuteurs de la même ethnicité que la sienne relèverait donc plutôt d'aspects culturels et linguistiques. Plusieurs études, par exemple celles de Thompson (1987), de Goggin, Thompson, Strube et Simental (1991) et de Perrachione et Wong (2007), soutiennent qu'il est plus facile de reconnaitre ou d'identifier une voix dans la même langue que la sienne. Selon d'autres études, il demeure plus facile de reconnaitre un locuteur lorsqu'on possède des connaissances de la langue dans laquelle il parle que s'il s'agit d'une langue complètement étrangère (Köster et Schiller, 1997; Schiller et Köster, 1996; Sullivan et Schlichting, 2007). Des résultats similaires ont aussi été constatés par Levi (2018), qui soutient d'ailleurs que le bilinguisme d'écouteurs est également bénéfique. Finalement, l'étude plus récente de Stevenage, Clarke et McNeill (2012) conclut que même une différence d'accent, dans le cas présent entre l'accent anglais et l'écossais, est suffisante pour nuire aux

tâches de reconnaissance et d'identification de locuteurs. En somme, il parait clair que la langue parlée par les personnes impliquées dans ce type de tâche influence considérablement la validité des résultats obtenus (voir aussi Perrachione, 2017).

Un facteur plus difficile à quantifier est la variabilité individuelle dans la capacité de traiter les voix. Comme le souligne Yarmey (2014), des études traitant spécifiquement de cette question sont encore à faire. À cet effet, l'étude de Sørensen (2012) a fait état d'une asymétrie dans les résultats obtenus lors d'une tâche de reconnaissance de locuteurs dont les voix étaient plus ou moins distinctives. Dans cette étude, les participants étaient premièrement exposés à un échantillon de 30 secondes de parole spontanée prononcée par 2 locuteurs inconnus. Une semaine après avoir entendu ces échantillons de parole, deux parades vocales distinctes, une pour chaque locuteur entendu, étaient présentées aux participants. Ceux-ci devaient indiquer laquelle des 5 voix composant la parade vocale avait été entendue la semaine précédente. Les résultats rapportés indiquent qu'un même sous-groupe de participant n'est pas parvenu à identifier aucune des deux voix cibles. Ainsi, certaines personnes seraient tout simplement moins aptes à reconnaitre des voix entendues. Un nombre restreint d'études se sont néanmoins intéressées sur les aspects plus tangibles liés à cette variation individuelle tel que l'effet d'une formation connexe (p. ex. dans le domaine de la phonétique ou musical). Ainsi, Eladd, Segev et Tobin (1998); Köster, Hess, Schiller et Künzel (1998) et Schiller et Köster (1998) soutiennent que les participants spécialistes de la voix ou de la parole performent significativement mieux que la population en général aux tâches de reconnaissance ou d'identification de locuteurs. Köster et al. (1998) et Xie et Myers (2015) ont observé des performances également supérieures chez les participants ayant une formation musicale. Bien que la variabilité individuelle soit encore difficile à mesurer, on peut néanmoins conclure que la sensibilité d'un individu par rapport à l'audition est un facteur pouvant s'avérer important en prévision d'une tâche liée au traitement de l'identité véhiculée par la voix.

Le facteur autour duquel est concentré le présent travail de thèse est la familiarité des voix. Comme il en sera question à la section suivante, une voix peut être familière de différente manière, surtout dans un contexte expérimental. Ainsi, trois méthodes permettent d'observer les effets de la familiarité des voix : l'utilisation de voix célèbres, l'entrainement des participants à

connaitre des voix préalablement inconnues ou, finalement, l'utilisation de voix intimement familières des participants. Cette dernière approche est la moins représentée dans la littérature en raison des complications qu'elle encourt, principalement en matière d'élaboration des stimuli. Le recours à des stimuli vocaux intrinsèquement liés au recrutement de participants complexifie grandement l'élaboration des stimuli eux-mêmes, ainsi que le recrutement de participants. Un certain nombre d'études en phonétique légale ont néanmoins tenté de relever le défi. Les études de Hollien et al. (1982) et de Yarmey, Yarmey, Yarmey et Parliament (2001) ont toutes deux rapporté des résultats significativement plus élevés à une tâche d'identification lorsque la voix à identifier était intimement familière que lorsqu'elle ne l'était pas (respectivement 98 % versus 49,8 % et 85 % versus 55 %). Une autre étude phare ayant traité de la familiarité des voix est celle de Foulkes et Barron (2000). Dans cette étude, 10 membres d'un même réseau social âgés de 20 ou 21 ans, donc tous très familiers, ont eu à s'identifier entre eux au moyen d'enregistrements vocaux. Globalement, les résultats obtenus sont inférieurs à ceux de Hollien et al. (1982) et de Yarmey et al. (2001) et atteignent une moyenne de 67 % d'identifications correctes. Dans une réplication du même paradigme expérimental, l'étude de Doromal (2016) fait par contre état de résultats plus congruents avec un taux de réussite de 94 %. Hormis les analyses acoustiques faites *a posteriori* par Foulkes et Barron (2000) et par Doromal (2016), aucune de ces études n'a fait état de la similitude, ou non, des voix présentées comme stimuli.

C'est entre autres sur cet aspect que se distingue l'étude de Plante-Hébert et Boucher (2014), puisque les voix utilisées étaient non seulement intimement familières, mais elles étaient également très similaires en termes de F_0mp. Dans cette étude, 44 participants ont chacun écouté 8 parades vocales de durée variables. Chaque parade vocale était composée de 10 voix, dont une seule était intimement familière du participant à différents degrés. Après avoir entendu une parade vocale entière, les participants devaient indiquer quelle était la voix de la personne intimement connue. Les résultats de cette étude font état de taux d'identification frôlant les 100 % d'exactitude chez les participants très familiers et avec des stimuli de plus d'une syllabe. Ce taux d'identification très élevé a été obtenu malgré l'utilisation de voix partageant une F_0mp très similaire.

Ces observations, bien qu'encourageantes quant à la précision de la capacité humaine d'identifier les individus par leur voix, rencontrent certaines limitations dans leur application au domaine légal. Si, par exemple, la seule personne en mesure d'identifier un suspect par sa voix est une personne qui lui est très familière, la possibilité de conflit d'intérêts remet en question la validité de la procédure d'identification elle-même. Le recours à une approche qui permette d'exploiter cette capacité d'identification par la voix sans pour autant reposer entièrement sur des réponses fournies intentionnellement par les identificateurs s'avère ainsi nécessaire. C'est ainsi que l'exploration des techniques neurophysiologiques s'est imposée.

Modèles du traitement neuronal des voix

Avant de se pencher concrètement sur l'utilisation de techniques neurophysiologiques dans un cadre d'application en phonétique légale, une revue des principales connaissances au sujet du traitement neuronal des voix est indispensable. Les modèles qui ont été élaborés pour rendre compte du traitement des voix dans le cerveau humain ont principalement découlé de connaissances analogues au sujet du traitement des visages, d'études de cas cliniques ou encore de données en neuroimagerie. Les principales contributions de ces domaines sont décrites dans les sections qui suivent.

Modèle clinique

Comme c'est le cas pour le domaine de la phonétique légale, c'est dans les années 1980 que les études cliniques au sujet de patients incapables de reconnaitre les voix ont commencé à foisonner. Cette condition a rapidement été désignée par le terme « phonagnosie », appellation qui fait référence à l'incapacité de reconnaitre les visages connus sous le nom de « prosopagnosie » (Neuner et Schweinberger, 2000; Van Lancker et Canter, 1982). La phonagnosie est un déficit qui peut être d'ordre perceptuel ou encore d'ordre associatif (Peretz et al., 1994). Dans le cas d'une phonagnosie perceptuelle, c'est le traitement de l'information acoustique même qui est affecté, empêchant par conséquent toute forme de reconnaissance et d'identification subséquentes. La phonagnosie associative, quant à elle, relève d'un trouble d'accès aux représentations stockées en mémoire, tandis que le traitement acoustique demeure intact. Un patient atteint de phonagnosie associative serait, par exemple, capable de faire la

distinction entre deux voix similaires, mais serait incapable d'associer l'une ou l'autre à un individu, aussi familier lui soit-il. Aussi, si les premiers cas diagnostiqués relevaient d'un trouble acquis, quelques cas de phonagnosie développementale ont aussi été décrits et étudiés plus récemment (Didic et al., 2020; Garrido et al., 2009; Roswandowitz et al., 2014; Xu et al., 2015).

L'étude de cette condition a entre autres permis d'observer une distinction hémisphérique quant à la capacité de distinguer des voix inconnues entre elles et le traitement de voix connues (Van Lancker et Kreiman, 1985; Van Lancker et Kreiman, 1987). Ainsi, selon les données cliniques recueillies, les voix familières sont traitées dans l'hémisphère droit, dans leur globalité, tandis que les voix inconnues sont traitées dans l'hémisphère gauche de manière plus analytique, par le traitement de chaque caractéristique acoustique perçue dans la voix (Kreiman et Sidtis, 2011, p. 187-188). Les autrices ont nommé ce modèle du traitement de la voix *le modèle du hérisson et du renard* (*the Fox and the Hedgehog*) en référence à l'œuvre du poète grec Archiloque dans laquelle le renard connait plusieurs petites choses tandis que le hérisson n'en connait qu'une seule très importante. Si on applique ce modèle au domaine de la voix, on peut présumer que des caractéristiques spécifiques telles que la F_0, le débit, la qualité de la voix, l'amplitude et d'autres encore seraient analysées une à une lorsqu'une voix inconnue est entendue. D'un autre côté, les voix intimement familières seraient reconnues dans leur ensemble, comme un patron fréquemment rencontré qu'il n'est pas nécessaire de décortiquer. Forcément, l'analyse détaillée de nombreuses caractéristiques acoustiques est un processus qui nécessite plus de temps et d'énergie que la reconnaissance d'un patron déjà présent en mémoire.

Globalement, les recherches auprès de patients atteints de phonagnosie ont permis de développer les connaissances au sujet des régions corticales impliquées dans les différents déficits au niveau du traitement de l'identité par la voix. Les spécialisations de l'hémisphère droit pour le traitement des voix familières et du gauche pour les voix inconnues sont également d'importantes contributions. Ces recherches ont aussi mené aux postulats que les voix inconnues sont traitées par l'analyse de chaque caractéristique acoustique tandis que les voix familières sont traitées selon leur portrait global. Comme pour la plupart des troubles d'ordre neurologique, la plasticité neuronale peut avoir entrainé une réorganisation structurelle et fonctionnelle de

certaines régions corticales chez les patients étudiés. L'étude du traitement normal des voix s'avère donc également essentielle pour pleinement saisir les processus impliqués.

Modèle de Bruce et Young

En dehors de la voix, l'identité d'une personne peut être décelée de multiples autres manières. Au quotidien, la principale modalité d'identification des individus qui nous entourent est visuelle. Il est d'ailleurs généralement admis que cette voie d'accès est privilégiée par rapport à la voix pour identifier les gens (pour une discussion complète sur ce sujet, lire Stevenage et Neil, 2014). Les études sur la reconnaissance et l'identification d'individus par le visage, suivies par des études analogues sur la voix, ont permis d'élaborer un premier modèle de traitement de l'identité des individus dans le cerveau.

Les travaux de Bruce et Young (1986) et de Burton, Bruce et Johnston (1990) sont à la base des modèles les plus répandus pour rendre compte du traitement neuronal de l'identité d'individus. Les premières versions du modèle *Interactive Activation and Competition* (IAC) ne considéraient alors que le traitement des visages. On y présente les unités de reconnaissance des visages[1] (*face recognition units*, FRU) et les foyers d'identité individuelle[2] (*person identity nodes*, PIN). Selon le modèle en question, les FRU sont des unités spécifiques à une modalité, ici visuelle, qui composent l'ensemble des représentations d'un individu en mémoire selon cette modalité. Lorsqu'un visage rencontré correspond suffisamment aux FRU en mémoire, l'association à un PIN est alors possible.

C'est Ellis, Jones et Mosdell (1997) qui ont ensuite proposé une version de ce modèle qui intégrait le traitement des voix familières en postulant l'existence d'unités de reconnaissance des voix[3] (*voice recognition units*, VRU). Dans les premières versions du modèle IAC prenant en compte d'autres modalités que visuelle, chaque modalité était traitée en parallèle des autres à chaque stade de traitement, jusqu'à parvenir aux PINs, ces derniers étant amodaux. Plusieurs versions adaptées du modèle ont été proposées en fonction de diverses données recueillies, mais

[1] Traduction libre de l'auteur.
[2] Traduction libre de l'auteur.
[3] Traduction libre de l'auteur.

l'essence du modèle demeure généralement inchangée (pour une discussion plus détaillée du modèle, voir Barton et Corrow, 2016; voir p. ex. Lucchelli et Spinnler, 2008).

Modèle neuroanatomique

Dans la foulée des études de cas de phonagnosie et du développement du modèle de l'IAC, plusieurs expériences en imagerie par résonnance magnétique fonctionnelle (IRMf) ont contribué au développement des connaissances au sujet du traitement neuronal des voix d'une perspective plus structurelle.

C'est à la suite d'une de ces études IRMf que Belin, Zatorre, Lafaille, Ahad et Pike (2000) ont adopté la désignation d'aire temporale de la voix (*temporal voice area*, TVA) pour décrire les zones corticales sollicitées spécifiquement par les stimuli vocaux, qu'ils soient verbaux ou non, en comparaison avec des stimuli sonores non vocaux (p. ex. bruits naturels divers, sons d'animaux, musique,). La TVA a ainsi été localisée dans les sulcus et gyrus temporaux supérieurs bilatéraux (respectivement STS et GTS).

Ces observations sur l'importance des régions temporales dans le traitement des voix concordent avec celles préalablement rapportées par Imaizumi et al. (1997). En utilisant la tomographie par émission de positron (TEP), les auteurs ont observé que les régions temporales étaient significativement plus sollicitées pendant une tâche de reconnaissance du locuteur que pour une tâche d'identification des émotions véhiculées par les voix. Dans une autre étude de TEP, Nakamura et al. (2001) ont démontré que le lobe temporal supérieur droit était quant à lui davantage activé lorsque des voix familières étaient présentées par rapport à des voix inconnues.

C'est en considérant ces données ainsi que les leurs que Belin, Fecteau et Bédard (2004) ont proposé un modèle de traitement des voix parallèle à celui des visages. Selon eux, le traitement des voix se divise en trois composantes qui font suite aux stades d'analyses de bas niveau : l'analyse de la parole, celle de l'affect et celle de l'identité du locuteur. Chacun de ces stades d'analyse est indépendant, mais des échanges entre les modalités auditives et visuelles sont possibles.

Les expériences IRMf subséquentes de Von Kriegstein et Giraud (2004), Von Kriegstein, Kleinschmidt, Sterzer et Giraud (2005) et Birkett et al. (2007) ont permis de raffiner les observations anatomiques précédentes en différenciant les fonctions des STS antérieur et postérieur droits. Leurs résultats, obtenus à la suite de présentations de voix familières et de voix inconnues, ont indiqué que le STS postérieur droit est plutôt spécialisé dans le traitement des caractéristiques acoustiques des voix, tandis que la portion antérieure du STS droit est spécialisée dans le traitement global de voix familières. Le rôle du STS antérieur droit dans le traitement de l'identité du locuteur avait d'ailleurs déjà été évoqué par Belin et Zatorre (2003) alors qu'une même voix était présentée à deux reprises consécutives dans une expérience en IRMf. Dans une étude d'envergure auprès de 218 individus en bonne santé neurologique, Pernet et al. (2015) ont confirmé le rôle des STS et GTS en tant que TVA, et ce, malgré la variabilité individuelle rapportée. Les auteurs ont eux aussi noté la présence de trois sous-régions, soit les TVA antérieure, médiale et postérieure.

Moutures contemporaines

Un des apports majeurs au modèle de l'IAC a été formulé par Gainotti (2014a) à la suite de revues de la littérature détaillées (Gainotti, 2013, 2014b). En réponse aux données présentées dans plusieurs rapports cliniques et expérimentaux, l'auteur propose d'intégrer les interactions entre les différentes modalités (visage, voix et noms) au stade même des unités de reconnaissance (FRU, VRU et NRU) plutôt que des voies d'accès aux PINs entièrement parallèles. Cette modification du modèle de l'IAC fait suite aux résultats d'un bon nombre d'études en neuroimagerie dont les résultats soutiennent une telle interaction (Blank, Anwander et von Kriegstein, 2011; Föcker, Hölig, Best et Röder, 2011; Gonzalez et al., 2011; O'Mahony et Newell, 2012; Robertson et Schweinberger, 2010; Schweinberger, Kloth et Robertson, 2011; Von Kriegstein et Giraud, 2006; Von Kriegstein et al., 2005). Gainotti (2018) avance d'ailleurs que ces connexions intermodales au niveau des unités de reconnaissance pourraient également représenter des voies de traitement alternatives advenant la détérioration de la voie habituelle de traitement pour une modalité donnée .

De plus, Gainotti (2014b) suggère que le sentiment de familiarité soit rencontré au stade des unités de reconnaissance tandis que l'identification relève des PINs. La distinction entre ces deux notions sera discutée plus en détail au Chapitre 2.

La version de l'IAC proposée par Gainotti (2014a) est aussi la première à synthétiser et à prendre en considération les asymétries hémisphériques observées. On y propose que les visages et les voix soient principalement traités par l'hémisphère droit tandis que les noms et l'information codée verbalement liée à l'identité sont traités par l'hémisphère gauche.

Dans leur méta-analyse, Blank, Wieland et von Kriegstein (2014) font des constats similaires à Gainotti (2014a). Pour chacune des modalités d'accès à l'identité, soit le visage, la voix et le nom, elles proposent que les régions responsables du sentiment de familiarité soient interconnectées et échangent de l'information. Pour arriver à ces conclusions, Blank et al. (2014) ont colligé les résultats de nombreux articles réunissant des études de cas, de groupes cliniques et de neuroimagerie. Leur méta-analyse démontre aussi que les régions corticales impliquées dans le traitement de voix connues varient selon le type de familiarité. Pour ce faire, les autrices ont observé les données d'études qui ont utilisé des voix intimement familières, célèbres ou encore nouvellement apprises explicitement à des fins expérimentales (p. ex en laboratoire). Malgré un certain chevauchement prévisible, les autrices montrent que les régions activées par ces trois types de familiarité ne sont pas les mêmes. Dans les conclusions de cet article phare, on invite d'ailleurs à explorer la reconnaissance de voix inconnues ou nouvellement apprises en comparaison avec les voix connues. Cet aspect n'est pas sans conséquence en vue d'éventuelles applications dans le domaine de la phonétique légale et sera exploré davantage dans le Chapitre 2.

De leur côté, Maguinness, Roswandowitz et von Kriegstein (2018) se sont concentrés davantage sur les distinctions structurelles entre le traitement de voix familières par rapport aux voix inconnues. Les auteurs de cette étude rendent compte de ces distinctions en incorporant certains éléments clés du modèle d'identification des locuteurs basé sur les prototypes de Lavner, Rosenhouse et Gath (2001) à l'hypothèse de l'IAC. Dans ce modèle, la « distance » entre les caractéristiques vocales entendues et celles d'une voix prototypique présente en mémoire est

calculée. Le prototype de voix varie d'une personne à l'autre et est composé des caractéristiques vocales les plus fréquemment rencontrées dans l'environnement de l'individu concerné (p. ex. l'accent régional). La « distance » entre une voix entendue et la voix prototypique n'est calculée que pour les paramètres pour lesquels elles divergent. Si cette « distance » est suffisamment faible, la voix est traitée comme une voix connue et peut, par la suite, être identifiée. Autrement, elle est traitée comme une voix inconnue et est stockée en mémoire en vue d'établir un nouveau prototype. En plus d'intégrer des éléments de la théorie de prototypes, le modèle de Maguinness et al. (2018) puise aussi dans le modèle de Belin et al. (2004) en proposant un traitement symétrique des visages et des voix. Comme pour le modèle de Belin et al. (2004), les deux modalités communiquent entre elles aux différents stades de traitement.

Finalement, Young, Frühholz et Schweinberger (2020) ont proposé une version mise à jour du modèle de Belin et al. (2004) qui propose un traitement multimodal de la parole, de l'affect et de l'identité plutôt qu'un traitement unimodal interactif. Ce modèle considère également la contribution pondérée de chaque modalité dans le traitement des différentes composantes vocales (parole, affect et identité). Par exemple, le modèle rend compte de la préférence générale de la modalité visuelle par rapport à la modalité auditive dans le traitement de l'identité et de la tendance inverse dans le traitement de la parole.

Les postulats des modèles présentés, bien que différents, ne sont pas entièrement irréconciliables puisqu'ils convergent également à certains égards. Premièrement, il semble généralement admis que la discrimination entre deux voix inconnues ne fasse pas appel aux mêmes processus que le traitement de voix familières. L'étude des voix inconnues reste cependant circonscrite à des tâches de discrimination, et il existe très peu de données quant à la reconnaissance de voix inconnues.

Les modèles de Bruce et Young et de Belin et al., ainsi que leurs versions plus récentes, ont aussi en commun qu'ils intègrent les connexions entre les modalités visuelle et auditive. La nature de ces échanges intermodaux varie d'un modèle à l'autre, mais il n'en demeure pas moins incontournable de considérer l'impact qu'ont les modalités l'une sur l'autre lorsqu'il est question du traitement de l'identité.

Ces questions à propos des processus impliqués dans le traitement des voix et des liens intermodaux d'accès à l'identité seront explorées dans les prochaines sections en demeurant ancrées dans une perspective d'application à la phonétique légale.

Les potentiels évoqués

Les potentiels évoqués (PÉs) sont une technique d'analyse du signal électroencéphalographique (EEG), qui est quant à lui le signal électrique représentant l'activité neuronale telle que captée par des électrodes au niveau du scalp. Les PÉs sont obtenus en calculant la moyenne d'un grand nombre de réponses EEG à une stimulation externe. En variant la stimulation externe, on peut ainsi comparer les PÉs d'une condition par rapport à une autre. Certains sommets d'activités sont bien connus pour être modulés par des processus spécifiques et portent le nom de composantes. Plus ces dernières sont précoces, plus elles représentent des processus d'analyse de bas niveau, souvent même inconscients. À l'opposé, les composantes tardives reflètent quant à elles des processus plus complexes et une analyse plus fine.

Bien que les modèles décrits dans la section précédente aient été élaborés principalement à partir de données cliniques et neuroanatomiques, les PÉs ont eux aussi été utilisés dans l'étude du traitement des voix dans le cerveau humain. C'est d'ailleurs autour de cette technique que les Chapitres 2 et 3 et les expériences qui y sont présentées sont construits. Une recension de la littérature mettant en lumière les principales contributions attribuables aux PÉs sera donc présentée plus en détail dans ces chapitres.

L'intérêt particulier des PÉs dans un cadre d'application à la phonétique légale n'est pas sans raison. Cette technique réputée pour sa grande précision temporelle permet d'observer plus facilement des distinctions fines dans les processus impliqués que d'autres techniques comme l'IRMf. Comme la voix transmet l'information de manière séquentielle au fil du temps, cette grande précision temporelle s'avère incontournable pour bien capter les nuances dans son traitement.

Contrairement à d'autres techniques neurophysiologiques, l'accessibilité et la mobilité de l'électroencéphalographe (EEG) contribuent à rendre l'analyse de PÉs intéressante d'un point de

vue appliqué en phonétique légale. En comparaison avec l'IRMf, le matériel nécessaire aux enregistrements EEG est assurément plus léger et son utilisation moins délicate. Des versions simplifiées et portatives de systèmes d'enregistrement EEG ont d'ailleurs été récemment commercialisées. Cette simplicité d'accès en fait une technique neurophysiologique de choix pour l'utilisation auprès de témoins.

Objectifs généraux

De manière générale, en plus de contribuer aux développements des connaissances en ce qui a trait au traitement des voix chez l'humain, le travail de thèse qui suit vise à explorer les possibilités qu'offrent les PÉs en vue d'application de la reconnaissance et de l'identification de locuteurs en phonétique légale.

Dans un premier temps, l'expérience 1 a comme objectif de faire une distinction claire, à l'aide des PÉs, entre les processus de reconnaissance et d'identification du locuteur. Cette même expérience examinera aussi la reconnaissance de voix intimement familières et celle de voix inconnues afin de déterminer si toutes deux relèvent ou non des mêmes processus.

L'expérience 2 se concentre quant à elle sur la mémoire épisodique et l'information contextuelle. Plus spécifiquement, les effets de la modalité audiovisuelle et de l'interaction sociale sur l'apprentissage des voix et les réponses des PÉs correspondants y sont observés. L'objectif principal de cette expérience est d'établir l'effet de l'information contextuelle multimodale sur l'apprentissage des voix et les PÉs en réponse à des voix apprises.

Par la suite, une discussion des résultats des expériences 1 et 2 permettra de mettre les observations en lien avec les modèles existants sur les accès à l'identité des individus. Cette discussion reviendra également sur les implications qu'ont ces résultats dans un cadre d'application pratique en phonétique légale.

Chapitre 2 : Reconnaissance et identification du locuteur

Ce chapitre et l'expérience qui y est présentée vise spécifiquement à clarifier les notions de reconnaissance et d'identification du locuteur, tant d'un point de vue terminologique qu'en termes de marqueurs électrophysiologiques.

Problème terminologique

En psychologie, les notions de familiarité et de rappel, respectivement *familiarity* et *recollection*, sont définies de manière claire et sans équivoque. La familiarité fait référence au fait de savoir qu'un stimulus, souvent visuel dans les études en psychologie, a déjà été rencontré précédemment, sans pour autant être en mesure de fournir des informations supplémentaires à son sujet (p. ex. à quel moment le stimulus a-t-il été vu la première fois, ce qu'il évoquait, dans quelle position il était, etc.). Le rappel réfère quant à lui aux stimuli pour lesquels il est non seulement possible d'affirmer qu'ils ont été rencontrés précédemment, mais pour lesquels il est également possible d'accéder à des informations supplémentaires d'ordre sémantique. La distinction entre les notions de familiarité et de rappel est principalement faite en fonction de la présence ou non d'information d'ordre sémantique. Ces informations sont de natures variées et multisensorielles et accompagnent ou même composent le concept en mémoire à long terme. Lorsqu'il est question d'un individu par exemple, on pourrait penser à son nom, sa taille, la couleur de ses cheveux, l'odeur de son parfum et ainsi de suite. Cette importante nuance entre rappel et familiarité trouve également écho dans la littérature, plus spécialement en électrophysiologie. Cette technique étant dotée d'une grande précision temporelle, elle permet de distinguer les fines différences entre divers processus (pour des revues de la littérature complètes à ce sujet, voir Wilding et Ranganath, 2011; Yonelinas, 2002).

Lorsque l'objet d'étude en question provient d'êtres humains, que ce soit leur visage ou leur voix, ces notions de familiarité et de rappel sont habituellement substituées par les termes « reconnaissance » et « identification » (Kreiman et Sidtis, 2011, p. 157). On peut ainsi voir ou entendre un individu tout en sachant très bien l'avoir déjà rencontré, sans pour autant être en

mesure de le nommer ou de spécifier d'où, quand et comment on le connait. Dans ce cas, on parle de « reconnaitre » quelqu'un. D'un autre côté, lorsqu'on est en capable de nommer cet individu ou de se rappeler dans quel contexte on l'a rencontré, on parle plutôt d'« identifier » quelqu'un.

Malgré l'existence de cette distinction terminologique claire, une certaine confusion demeure quant à son utilisation dans la littérature scientifique qui porte sur le traitement de l'identité du locuteur, surtout en neurosciences. On remarque que les désignations « *speaker recognition »* et « *speaker identification »* sont parfois pratiquement interchangeables, que ce soit chez un même auteur ou d'un auteur à l'autre. Certains contournent le problème en utilisant une terminologie qui leur est propre ou encore en demeurant vagues sur cette question. C'est le cas, entre autres, d'une recension exhaustive de la littérature présentée par Roswandowitz, Maguinness et von Kriegstein (2019). Bien qu'il y soit question de processus séquentiels de traitement acoustique, de reconnaissance et d'identification des voix, la terminologie demeure complexe et fait référence à ces processus sans toutefois les nommer spécifiquement. Dans un tableau synthèse présentant les études sur la phonagnosie et sur les différentes tâches utilisées dans l'étude de cette condition, les auteurs regroupent sous l'appellation « *familiar voice recognition* » des paradigmes impliquant tant la reconnaissance que l'identification de locuteurs. Les termes parapluie liés au traitement de l'identité de la voix, comme « *voice-identity processing* », sont aussi fréquents dans la littérature.

Une des principales conséquences de cette ambiguïté terminologique est qu'il est risqué d'associer avec certitude les réponses, les processus et les régions corticales rapportés suite à des expériences avec l'une ou l'autre des notions concernées. En d'autres termes, l'absence d'un consensus terminologique quant à la reconnaissance et à l'identification de locuteurs empêche l'avancement des connaissances spécifiques à chacune de ces notions.

Pour ajouter à cette incohésion, les études portant sur les voix familières font appel à différents paradigmes expérimentaux pour éviter la complexité liée au fait d'utiliser des voix intimement connues de leurs participants. Les approches les plus utilisées sont l'utilisation de voix personnalités publiques, telles que celles de politiciens et de célébrités, ou encore la familiarisation de voix inconnues en laboratoire (ou entrainement). Il n'est cependant pas

démontré que les processus impliqués sont identiques lorsque la nature de la familiarité avec une voix varie (p.ex. une voix intimement familière comparativement à une voix apprise en laboratoire).

Comme en témoignent les études directement appliquées à la phonétique légale présentées au Chapitre 1, la discordance terminologique décrite ci-dessus semble être principalement circonscrite au domaine des neurosciences. Lorsque la reconnaissance et l'identification de locuteurs sont étudiées dans un contexte d'application plus pratique, comme c'est le cas en phonétique légale, la distinction entre les deux processus est plus manifeste et soutenue puisqu'elle s'avère nécessaire.

Expérience 1

Objectifs et hypothèses spécifiques

Dans ce contexte, une première expérience avec comme objectif principal de déterminer si l'utilisation des PÉs permet de révéler la présence de marqueurs électrophysiologiques liés spécifiquement au traitement de voix intimement familières a été élaborée. Les hypothèses avancées sont que :

1. Une distinction claire, en termes de PÉs, est possible lorsque des voix intimement familières sont présentées en comparaison avec des voix inconnues.
2. Une voix inconnue, mais entendue à plusieurs reprises se distinguera de voix purement inconnues en termes de PÉs.
3. Les distinctions anticipées dans les deux premières hypothèses ne portent pas sur les mêmes composantes puisque les voix intimement familières font appel à une mémoire sémantique à propos de locuteurs, ce qui n'est pas le cas pour les voix inconnues, même si répétées fréquemment. Par conséquent, une voix intimement familière peut être reconnue et identifiée, mais une voix inconnue peut seulement être reconnue.

Méthodologie

Participants

Treize participants (8 femmes) âgés entre 21 et 43 ans (moyenne = 30,81, σ = 5,14) ont pris part à cette expérience. Ils étaient tous des locuteurs natifs du français québécois (FQ) à l'exception près d'un participant l'ayant appris dès l'âge de 4 ans. Tous étaient également droitiers, tel qu'attesté par un questionnaire standardisé (Oldfield, 1971) et avaient une audition jugée normale suite à un test audiométrique. Les tests d'empan numérique en ordre direct et indirect ont quant à eux permis d'établir que tous les participants avaient une mémoire jugée normale.

Tous les participants de la présente étude ont été recrutés suivant les recommandations des membres d'un groupe de 36 locuteurs dont les voix ont été utilisées comme stimuli (dont une partie était également utilisée dans Plante-Hébert et Boucher, 2014). Plus précisément, chacun de ces 36 locuteurs fournissait volontairement le nom d'une personne jugée très familière (p. ex. un frère, un parent, un conjoint, etc.) et le recrutement de participants était effectué auprès de ces personnes familières uniquement si la voix du locuteur l'ayant recommandé était suffisamment similaire à celles des autres locuteurs (les voix devaient entre autres être similaires en termes de F_0mp, tel que décrit dans la prochaine section). Ce lien de familiarité entre un locuteur et un participant était ensuite quantifié à l'aide d'un questionnaire élaboré et utilisé dans une étude antérieure (Plante-Hébert et Boucher, 2014). Cette méthode de recrutement unique permettant d'évaluer les PÉ lors de l'exposition de voix intimement familières explique le nombre restreint de participants ayant pris part à l'expérience.

Tous les participants ont été rémunérés après avoir lu et signé un formulaire de consentement approuvé par le Comité d'éthique à la recherche du *CIUSS du Nord-de-l'île-de-Montréal* à *l'Hôpital Rivière-des-Prairies* de Montréal. Les participants avaient également l'opportunité de poser des questions spécifiques à l'expérience avant de signer le formulaire de consentement.

Stimuli

Les stimuli vocaux étaient huit énoncés de quatre syllabes chacun, listés dans le Tableau 1, produits par 14 locuteurs natifs du FQ sans accent régional se distinguant de celui de la région

montréalaise de manière perceptible. Dans un premier temps, la F_0mp des locuteurs a été contrôlée afin que celle-ci ne varie pas plus que d'un semi-ton parmi les 14 locuteurs (voir Tableau 6 en annexe). Ce contrôle s'est fait à l'aide du logiciel *Multi Speech* (KAY Pentax) qui permet de mesure la fréquence fondamentale parlée moyenne sur des enregistrements. La longueur syllabique des énoncés a quant à elle été établie à 4 syllabes en fonction des résultats d'une étude antérieure (Plante-Hébert et Boucher, 2015a) et d'un certain nombre d'autres observations portant sur la longueur minimale des stimuli requise pour accéder à une identification du locuteur fiable (Bricker et Pruzansky, 1966; Legge et al., 1984; Pollack et al., 1954; Roebuck et Wilding, 1993; Schweinberger, Herholz et Sommer, 1997; Skuk et Schweinberger, 2013). Bien que toutes ces études fassent part d'une longueur minimale requise, celle de Plante-Hébert et Boucher (2015a) évoque un minimum de 2 syllabes afin d'offrir une certaine quantité d'information spectrodynamique à l'identificateur.

Énoncé	Transcription API	Sons nasaux
Bonjour madame.	[bɔ̃ʒuʁmadam]	3
Combien t'en prends ?	[kɔ̃bjɛ̃tɑ̃pʁɑ̃]	4
Comment qu'elle va ?	[kɔmɑ̃kavɑ]	2
De temps en temps.	[dətɑ̃zɑ̃tɑ̃]	3
Donne-moi en deux.	[dɔmwazɑ̃dø]	2
J'en connais quatre.	[ʒɑ̃kɔnɛkat]	2
Quand est-ce qu'il vient ?	[kɑ̃tɛskivjɛ̃]	2
Quelqu'un t'attend.	[kɛkœ̃tatɑ̃]	2

Tableau 1. – Énoncés de 4 syllabes utilisés comme stimuli avec transcription API et nombre de sons nasaux par énoncé.

Finalement, comme l'indique le Tableau 1, chaque énoncé composant les stimuli comportait un certain nombre de sons nasaux puisqu'il a été démontré que ceux-ci facilitent l'identification du locuteur en offrant de l'information supplémentaire sur la physiologie du locuteur liée aux cavités de résonance (Amino et Arai, 2009; Amino, Sugawara et Arai, 2005, 2006; Glenn et Kleiner, 1968; Plante-Hébert et Boucher, 2015b; Su et al., 1974).

Les stimuli vocaux ont été enregistrés dans une cabine à l'épreuve du bruit à l'aide d'un micro-casque omnidirectionnel (*C477 WRL*, AKG) et d'une carte de son externe de 16 bits à un taux

d'échantillonnage de 44,1 kHz (*Fast-track Ultra*, M-Audio). Afin d'assurer une uniformité au niveau du débit de prononciation des stimuli, les locuteurs écoutaient à plusieurs reprises un métronome créé à partir de tons juxtaposés représentant les caractéristiques prosodiques désirées. L'amplitude des signaux acoustiques numérisés a ensuite été normalisée et chaque stimulus a été sauvegardé dans un fichier *.wav* individuel. Le début de chaque fichier audio a été aligné de façon à être précédé de 200 ms le début de la première syllabe de du stimulus. La durée du signal acoustique des stimuli s'étendait de 618 ms à 1085 ms (moyenne = 818 ms, σ = 83 ms). De manière générale, les stimuli utilisés dans la présente expérience répondaient aux recommandations émises en ce qui a trait à l'élaboration de parades vocales dans le domaine de la phonétique légale (voir p. ex. Atkinson, 2015; Hollien et al., 2014; Hollien et al., 1995; Nolan et Grabe, 1996).

Les stimuli décrits ci-dessus ont été étroitement contrôlés en termes de caractéristiques acoustiques afin de que les conditions de la présente expérience puissent servir d'analogie à la parade visuelle standard couramment utilisée. Ainsi, il est nécessaire que les différents individus qui composent une parade se ressemblent, visuellement ou acoustiquement. Il serait par exemple contre intuitif d'entourer un suspect masculin mesurant six pieds 2 pouces et ayant les cheveux blonds des personnes de sexe féminin, de plus petite taille ou encore ayant les cheveux bruns. C'est donc selon cette logique que des stimuli adéquatement contrôlés ont été élaborés. Il est également à noter que la technique des PÉs est très sensible aux variations d'ordre purement acoustique. Ainsi, des stimuli possédant de trop grandes différences sur ce plan seraient propices à éliciter certaines réactions pouvant être confondues avec les effets expérimentaux anticipés.

Validation des stimuli

Afin d'assurer la validité interne des stimuli, un prétest auprès de 4 volontaires qui ne connaissaient aucun des locuteurs dont les voix étaient présentées comme stimuli a été fait. L'objectif de ce prétest était de s'assurer qu'aucune réaction électrophysiologique n'était associée aux stimuli, tant en termes de contenu linguistique des énoncés que des voix des locuteurs. Dans cette perspective, chaque volontaire a été exposé à 10 reprises à chaque énoncé pour chaque locuteur dans des conditions identiques aux conditions expérimentales décrites dans la prochaine section.

Ce prétest a permis de conclure que les PÉ observés suite à un nombre égal de présentations de chaque stimulus ne différaient pas visuellement en regard du contexte linguistique. Par contre, la voix d'un des locuteurs a dû être retirée du lot par précaution puisque les PÉs lui étant associés se distinguaient visuellement de l'ensemble de ceux des autres locuteurs. Aucune caractéristique acoustique n'a permis d'établir pourquoi les PÉ en réponse à la voix de ce locuteur différaient. En résumé, les PÉs décrits dans les prochaines sections ne peuvent qu'être associés à la manipulation des variables indépendantes manipulées, soit la familiarité du locuteur et la fréquence de présentation des voix.

Procédure

Les fichiers audios comprenant les stimuli ont été groupés en huit blocs selon les énoncés, tels qu'illustrés dans le tableau 1. Dans chaque bloc, la présentation de chaque voix était aléatoire avec comme seule restriction qu'une même voix ne pouvait être présentée à deux reprises consécutives. Quatre des huit blocs ont été utilisés pour procéder aux enregistrements EEG tandis que les quatre autres blocs, présentés en alternance, servaient à recueillir des réponses comportementales uniquement pour confirmer que la voix intimement familière était bien reconnue des participants. Les blocs EEG comprenaient chacun 240 essais (un essai étant la présentation d'un énoncé spécifique par un locuteur spécifique) et étaient ordonnés de façon à être présentés en premier, troisième, cinquième et septième. Les blocs comportementaux comprenaient quant à eux seulement 60 essais et étaient présentés en deuxième, quatrième, sixième et huitième. Ce nombre réduit d'essais était jugé suffisant pour attester de la reconnaissance de la voix intimement familière par les participants sans pour autant alourdir l'expérience et augmenter sa durée. Les énoncés associés à chaque bloc, EEG et comportemental confondus, était aléatoire entre les participants. Durant les blocs, la proportion de présentation de chaque voix variait : la voix intimement familière était présentée fréquemment (VF, 33 % des essais), une même voix inconnue était présentée fréquemment (VE, 33 % des essais) et les 12 autres voix inconnues étaient présentées rarement (VR, 2,77 % des essais chacune). Rappelons que chaque participant était recruté en fonction de sa familiarité élevée avec un seul des locuteurs constituant les stimuli et que toutes les autres voix lui étaient inconnues.

Tâche expérimentale

Les participants ont écouté les stimuli, dont l'amplitude était calibrée avec un sonomètre (Radioshack, modèle 33-2055) afin que les sommets d'amplitude atteignent 74 dBa à la sortie, en utilisant des écouteurs de type insert (E-A-Rtone 3A, EAR Auditory Systems). Les stimuli étaient joués par le logiciel *E-prime 1.0* (Psychology Software Tools). Les essais étaient séparés par un intervalle inter-stimuli (ISI) variant entre 500 et 650 ms par saut de 50 ms afin de minimiser tout effet d'anticipation. Lors de la présentation des stimuli, les participants étaient assis confortablement à 180 cm d'un écran d'ordinateur sans activité, mais avec une croix de fixation en son centre. Les participants recevaient la directive de garder leur regard sur cette croix de fixation pour toute la durée d'un bloc. Pour les quatre blocs comportementaux,ils recevaient également la directive d'indiquer, après chaque essai, le plus rapidement possible et le mieux possible si la voix entendue était la voix familière ou si elle était non familière en cliquant sur les boutons d'une souris d'ordinateur (qui étaient inversés pour la moitié des participants). Pour minimiser les mouvements des participants, ceux-ci devaient garder leur main dominante sur la souris.

Enregistrement EEG et analyses

Bien que les analyses n'aient porté que sur les PÉ des blocs EEG, le signal EEG a été enregistré pendant la durée entière de l'enregistrement. Les enregistrements ont été faits en suivant le système international 10-20 avec un amplificateur à 64 canaux *ASA lab EEG/ERP* (ANT neuro). Une référence moyenne en ligne a été utilisée et le signal EEG a été numérisé à un taux d'échantillonnage de 1000 Hz. Les mouvements oculaires ont été enregistrés à l'aide de quatre électrodes disposées au-dessus et en dessous de l'œil dominant (VEOG) et aux canthi extérieurs des yeux (HEOG). L'électrode AFz a été utilisée comme mise à terre et l'impédance de chacune de 64 autres électrodes était maintenue sous 10 kΩ toute la durée des enregistrements.

À l'aide du logiciel *ASA software* (ANT neuro), chaque enregistrement a été filtré avec un filtre passe-bande (0,1-30 Hz) et les clignements d'yeux ont été retirés. Tous les autres artéfacts qui excédaient un écart type de 20 µV à l'intérieur d'une fenêtre coulissante de 200 ms ont également été retirés avec le logiciel *Eegprobe GUI* (version 1.2.0.2, ANT Software). Le

moyennage de tous les essais pour tous les participants par blocs et par types de voix (VF, VE et VI) a été effectué avec *Fieldtrip* (Oostenveld, Fries, Maris et Schoffelen, 2011), un outil de source libre pour *MatLab* (R2017b 9,3). Depuis les enregistrements initiaux, chaque essai a été découpé en débutant 200 ms avant le début d'un stimulus jusqu'à 1000 ms après ce même début. L'intervalle de 200 ms précédant le début du stimulus a été utilisé pour la correction en fonction de la ligne de base.

L'inspection visuelle du signal moyenné a permis d'identifier aisément le complexe P1-N1-P2, suivi d'une négativité entre 300 et 350 ms correspondant à la N250 et d'ondes lentes qui s'étendaient jusqu'à la fin de la fenêtre d'analyse. En fonction des résultats précédemment rapportés dans la littérature, les analyses ont été concentrées sur la P2, situé sur des sites centro-frontaux droits, la N250, dans les régions centro-frontale gauches et centres et sur les ondes lentes observées dans les régions centro-frontale droite et centro-pariétales gauche et centre.

Le détail des analyses par fenêtre temporelle et par région est présenté dans l'article 1 avec figures à l'appui.

Résultats

Les résultats de l'expérience 1 sont présentés dans l'article suivant.

Article 1

L'article 1 a été publié dans la revue PLOS One suite à quelques modifications mineures.

Plante-Hébert, J., Boucher, V. J. et Jemel, B. (2021). The processing of intimately familiar and unfamiliar voices: Specific neural responses of speaker recognition and identification. *PLoS ONE*, 16, e0250214. doi: 10.1371/journal.pone.0250214

The processing of intimately familiar and unfamiliar voice:
Specific neural responses of speaker recognition and identification

Julien Plante-Hébert[a]*, Victor J. Boucher[a] & Boutheina Jemel[b, c]

[a]Laboratoire de Sciences Phonétiques, Département de Linguistique et de Traduction, Université de Montréal, Montréal, QC, Canada

[b]Laboratoire de Recherche en Neurosciences et Électrophysiologie Cognitive, Hôpital Rivière-des-Prairies, Montréal, QC, Canada

[c]École d'Orthophonie et d'Audiologie, Faculté de Médecine, Université de Montréal, Montréal, QC, Canada

*Correspondence: Julien Plante-Hebert; julien.plante-hebert@umontreal.ca

Abstract

Research has repeatedly shown that familiar and unfamiliar voices elicit different neural responses. But it has also been suggested that different neural correlates associate with the feeling of having heard a voice and knowing who the voice represents. The terminology used to designate these varying responses remains vague, creating a degree of confusion in the literature. Additionally, terms serving to designate tasks of voice discrimination, voice recognition, and speaker identification are often inconsistent creating further ambiguities. The present study used event-related potentials (ERPs) to clarify the difference between responses to 1) unknown voices, 2) trained-to-familiar voices as speech stimuli are repeatedly presented, and 3) intimately familiar voices. In an experiment, 13 participants listened to repeated utterances recorded from 12 speakers. Only one of the 12 voices was intimately familiar to a participant, whereas the remaining 11 voices were unfamiliar. The frequency of presentation of these 11 unfamiliar voices varied with only one being frequently presented (the trained-to-familiar voice). ERP analyses revealed different responses for intimately familiar and unfamiliar voices in two distinct time windows (200–250 ms and 450–850 ms post-onset) with late responses occurring only for intimately familiar voices. The late responses present sustained shifts, and short-time ERP components appear to reflect an early recognition stage. The trained voice equally elicited distinct responses, compared to rarely heard voices, but these occurred in a third time window (300–350 ms post-onset). Overall, the timing of responses suggests that the processing of intimately familiar voices operates in two distinct steps of voice recognition, marked by a P2 on right centro-frontal sites, and speaker identification marked by a late positive component (LPC). The recognition of frequently heard voices entails an independent recognition process marked by a differential N250. Based on the present results and previous observations, it is proposed that there is a need to distinguish between processes of voice “recognition” and “identification”. The present study also specifies test conditions serving to reveal this distinction in neural responses, one of which bears on the length of speech stimuli given the late responses associated with voice identification.

Keywords ERP, speaker recognition, speaker identification, voice familiarity, P2, N250, LPC

Introduction

The ability to recognize and identify voices stands as a remarkable and yet puzzling process of speech perception. From an evolutionary perspective, this ability is said to have been vital for the survival of humans and other species (Sidtis & Kreiman, 2012). But when one recognizes a voice, it is usually in the context of speech. No other species processes voice information in the context of fluctuating sounds of oral articulations, and the human ability to recognize or identify voices in such a context can be quite robust. In fact, in the case of an intimately familiar voice, such as the voice of a parent or sibling, there is nearly perfect recognition or identification independently of visual information (Plante-Hébert & Boucher, 2014). It is frequently assumed in voice research that such accuracy rests on the sensory processing of the spectral attributes of a voice signal as when producing such sounds as "ahhh" where oral motions are minimized (e.g., Beauchemin et al., 2006; Graux et al., 2013). However, as we established in an earlier study, when listeners are asked to pick out an intimately familiar voice amongst unfamiliar or unknown voices with similar fundamental frequency (F_0), there is a degree of inaccuracy when the stimuli are single syllables (Plante-Hébert & Boucher, 2015b). For near perfect recognition and identification to occur, two or more syllables can be required, and nasal sounds can be a factor for short sequences (Amino & Arai, 2009; Amino et al., 2005, 2006; Bricker & Pruzansky, 1966; Plante-Hébert & Boucher, 2015a; Pollack et al., 1954; Roebuck & Wilding, 1993; Su et al., 1974). This suggests that the processing of speaker-specific voice information involves dynamic spectro-temporal attributes reflecting moving resonators rather than static voice harmonics. It also indicates that, while some processing of speaker-specific information rapidly occurs over short intervals of speech, correct recognition or identification can require slightly longer temporal spans. Of course, given such findings, any attempt to circumscribe differing neural correlates of voice processing requires techniques that offer high temporal resolution (such as electroencephalography, EEG, and magnetoencephalography, MEG). It also entails, for the sake of clarity, a terminological distinction between processes that can potentially apply over different time intervals.

Indeed, the lack of a formal distinction between processes, or the variable use of terms such as voice "discrimination", "recognition", and "identification" to refer to an undefined "speaker

identity" has created a degree of confusion in the literature. The terms have been used to designate fundamentally different processes and can thus be essential in understanding the neurological mechanisms that underlie the processing of speaker-specific voice information. The following section serves to outline previous findings and general principles that support a strict distinction between voice *recognition* and *identification*, and also provides a demonstration of how this distinction relates to different EEG components in response to intimately familiar and trained (to-familiar) voices. For the sake of clarity, we use separate terms to designate these and other types of vocal stimuli, including intimately familiar voices (IFV), familiar voices (FV), frequently presented or trained-to-familiar voices (TV), and unfamiliar or unknown voices (UV).

1.1 On the neural underpinnings of voice discrimination, recognition, and identification

1.1.1 Early findings and clinical observations

Clinical reports in the 1980s provided crucial insights that have guided research on the processing of speaker-specific voices. The condition associated with an impaired ability to recognize FVs first appeared under the name of "phonagnosia" in Van Lancker & Canter (1982), a designation still widely used today. Since then, phonagnosia cases have been classified in two major categories: apperceptive phonagnosia, where the deficit is seen at the sensory or perceptual stages of voice processing, and associative phonagnosia, where the deficit lies in the association between a perceived voice and a particular speaker (Buchtel & Stewart, 1989; Hailstone, Crutch, Vestergaard, Patterson & Warren, 2010; Roswandowitz et al., 2019). It is useful to note with respect to clinical reports that, until quite recently, all cases of phonagnosia were observed in patients with brain damages. However, Garrido et al. (2009) presented a case study of developmental phonagnosia.

Early reports also focused on the ability of listeners to distinguish between FVs and UVs although "familiar" voices in these reports often referred to voices of famous individuals. The processing of these types of voices was generally investigated using tasks involving two-alternatives forced-choice paradigms (2AFC; e.g., Legge et al., 1984). Moreover, investigations of the processing of

UVs typically used designs where dyads of voices were presented in tasks requiring discriminatory same/different or old/new judgments following in-lab learning (e.g., McGehee, 1937). Using such protocols with participants presenting various brain lesions, Van Lancker, Kreiman & Emmorey (1985) and Van Lancker & Kreiman (1987) established that FV recognition can occur even when participants present an impaired ability to discriminate between pairs of UVs. This led the authors to conclude that sensory discrimination of unfamiliar voices could not be a preliminary stage of familiar voice recognition. Instead, the two abilities reflected different neural processes that were applied in parallel and not in a particular sequential order (Van Lancker & Kreiman, 1987, p. 833).

Following this line of research, later studies suggested that the processing of UVs rests on the perceptual processing of specific acoustic indices of pitch, speech rate, voice quality (etc.). According to these studies, the processing of voices involves a comparison of acoustic indices to prototypes stored in long-term memory and which come to consolidate in memory through a repeated exposure to voices (see Kreiman & Papcun, 1991; Lavner et al., 2001; Papcun et al., 1989). In this view, then, the discrimination of FVs and UVs centers on a presumed process of comparison between heard acoustic features of different voices and particular features coded in long-term store.

When it comes to familiar voices, however, an important distinction needs to be made between the feeling of knowing a stimulus and being able to explicitly recall qualitative information about the stimulus. In psychology, this principled difference is captured by general terms of “familiarity” and “recollection” (Curran, 2000; Tulving, 1985; Tulving & Murray, 1985). Such a distinction is generally admitted in memory research and supported by neurophysiological observations (Rugg & Curran, 2007; Yonelinas, 2001, 2002). Recollection, or recalling information about a stimulus as compared to judgments of its familiarity, involves episodic memory which is generally seen to entail activity in frontal cortical regions (Tulving, Kapur, Craik, Moscovitch & Houle, 1994; Wheeler, Stuss & Tulving, 1997). In this light, reviews of the clinical literature on the processing of FVs and UVs have indicated that distinct neural mechanisms underlie the feeling of familiarity as compared to the retrieval of episodes that have consolidated to form semantic representations relating to voice or speaker identity (Gainotti, 2011, 2015). Thus, the feeling of having heard a voice and knowledge of who a speaker is entail different processes, which implies that

investigations of these processes require different types of voice stimuli. Within early and current voice research, the terms voice *recognition* (familiarity) and *identification* (involving the retrieval of semantic information) should be regarded as a principled distinction by which to understand voice processing, as suggested by Kreiman & Sidtis (2011, p. 157). But this entails that stimuli consisting of previously heard or marginally familiar voices (FVs), including trained-to-familiar voices (TVs), may not necessarily involve identification processes as in *intimately* familiar voices (IFVs).

In their model of voice identity processing, Kreiman & Sidtis (2011, p. 187-188) propose that UVs are processed in terms of characteristic features while FVs are processed as whole "Gestalt-like" patterns. Hemispheric specialization, as described in Kreiman & Sidtis (2011, p. 209-212), varies specifically with voice familiarity. The view is that comparisons of features, which occurs in discriminating and recognizing UVs, links to processes in the left hemisphere whereas pattern-like recognition and identification of FVs involves functions of the right hemisphere. This distinction between FVs and UVs in terms of pattern and feature processing has also been supported by a number of recent studies reviewed by Stevenage (2018). In sum, the aforementioned differences between voice discrimination, recognition, and identification, as well as between types of voice stimuli (IFV, FV, TV, and UV) appear essential to circumscribing different neural mechanisms involved in processing vocal attributes of speech. Yet such distinctions, especially between voice recognition and identification, are not generally reflected in voice research. This can lead to a degree of confusion in interpreting observations in terms of underlying neural processes over and above differences in methodology, as illustrated below.

1.1.2 Electrophysiological observations

In considering studies that use ERPs, the following brief review sets aside a body of work relating to the interplay of visual and vocal information in voice processing, which entails varying methodologies (for summaries of this work, see Barton & Corrow, 2016; Blank et al., 2014; Gainotti, 2014a; Stevenage & Neil, 2014). Early studies involving ERPs focused on the discrimination of human voices and non-human sounds (e.g., animal cries, bell sounds, tones, etc.), which showed distinct responses to voices with an onset as late as 400 ms or the N400

(Gunji et al., 2003; Levy, Granot & Bentin, 2003; Levy, Granot & Bentin, 2001). More recent reports have revealed that the discrimination of human voices compared to generic sounds is represented by early components, around 150 ms, which have come to be termed the "fronto-temporal positivity to voice" (FTPV; Capilla, Belin et Gross, 2012; Charest et al., 2009; De Lucia, Clarke & Murray, 2010; Rogier, Roux, Belin, Bonnet-Brilhaut & Bruneau, 2010). Thus, there is evidence that the earliest neural components that relate specifically to human voices are in the order of 150 ms post-onset. Given these results, one can logically assume that any processing of voice identity information would entail later-occurring ERPs as could be revealed on stimuli of IFVs.

Few studies, however, have investigated the processing of IFVs such as the voice of a close friend or family member. One exception is a study by Beauchemin et al. (2006). That ERP study focused on responses of listeners to IFVs (close relatives or long-time friends) using an auditory oddball paradigm in reference to the MMN components (Näätänen, Gaillard & Mäntysalo, 1978). Short speech samples consisting of single vowels lasting some 200 ms produced by intimately familiar and unfamiliar speakers (IFVs and UVs respectively) were presented in conditions of passive listening. The results showed distinct responses across IFVs and UVs peaking at 200 ms post-onset, in the MMN range. Similar results of MMNs have also been reported in studies involving newborns, suggesting that the ability to recognize voices arises in early development (Beauchemin et al., 2011; Mai et al., 2012; Zinke, Thöne, Bolinger & Born, 2018).

In a different study that also involved an auditory oddball paradigm, Graux, Gomot, Roux, Bonnet-Brilhault & Bruneau (2015) compared the ERP of three sets of presented voices, including FVs, UVs, and participants' own voices (designated as "self"). The results displayed a significant MMN between 180 and 210 ms post-onset (for FVs compared to UVs) and a significant difference on the P3a between 230 and 320 ms for FVs compared to self-voice. These results confirmed a previous report of a distinct process between self and familiar voices (Graux et al., 2013). On the other hand, given that externally generated familiar voices are never heard as self-generated voices, it is difficult to extrapolate results on self-voices to processes of voice recognition and identification (and see Conde, Gonçalves & Pinheiro, 2015, 2016; Pinheiro et al., 2016; Rachman, 2018). On their side, Holeckova, Fischer, Giard, Delpuech & Morlet (2006), after exposing

participants to their own name pronounced by intimately familiar and unfamiliar speakers, reported a small effect on the P3, between 300 and 380 ms, but mostly on later-occurring ERP between 625 and 800.

Other studies have investigated ERPs to IFVs but with quite different results bearing specifically on voice identification. Of particular interest is a study by Schweinberger, Walther, Zäske & Kovács (2011) based on a 2AFC task involving paired stimuli of two IFVs that were morphed to varying degrees with one another. ERPs in this paradigm reflected changes in voice identification when increasing the proportion of one IFV in the stimuli relative to another IFV. The experiment also included congruent/incongruent speech contexts with /aba/ and /igi/ serving to examine the effects of verbal contexts on responses. Importantly, the results showed two responses occurring at different time intervals. A first response to IFVs occurred at parietal sites starting at 250 ms post-onset when speech contexts were congruent whereas, when speech contexts were incongruent, a speech-independent response to IFVs appearing, not as short-time components, but as protracted changes between 300 ms and 600 ms post-onset, in the P3 range. It is useful to note that the authors used the designation “voice identification” in commenting on their results (while also expressing reservations on their interpretations owing to the small number of participants).

Another investigation that involved ERPs and more or less FVs was that of Gonzalez et al. (2011). Their experiment used a go/no-go task with presented FVs and UVs in the context of a short phrase (the Spanish word /ola/). In this case, “FVs” referred to participants’ friends or colleagues so that it is unclear whether the stimuli could qualify as IFVs. The results showed ERP differences between FVs and UVs appearing between 280 and 840 ms post-onset, including a N250r and a P3, but again reflected protracted responses rather than short-time components as in Schweinberger, Walther, et al. (2011).

Finally, one should note that it is often assumed in voice research that pitch (given by F_0) is voice-specific. However, such aspects can relate to speech processing as in the case of “tone languages” where pitch changes serve to distinguish between words. In a study involving ERPs (and fMRI), Zhang et al. (2016) examined the varying responses obtained when listeners attend to changing

lexical tones in two Cantonese words /ji/(produced with high or rising tones) and when they attend to changing voices (UVs of a male and female speaker producing the words). The design aimed to compare ERPs of talker and speech deviants with reference to a standard. The analyses of designated components showed talker-specific changes in P2, P3a, and on frontal negativities examined over an interval of 500–800 ms (late parietal components also appeared but were not analyzed). An important methodological implication of this study is that it showed task-dependent interactions between talker and speech processing where pitch could not be taken *a priori* as a property of "voices". Moreover, the authors specified that the differences in F_0s between the male and female voices (101 Hz) exceeded differences in F_0s of speech contexts (56 Hz). There is much behavioral evidence that salient differences in voices can influence memory as opposed to less distinctive voices and such differences on distinctiveness are likely to reflect in ERPs. However, few reports specify F_0 values of voice stimuli, which may underlie the discrepancy in reported components of voice processing. But perhaps a more important source of variation is the length of the stimuli used across studies.

Generally, and in comparing various reports listed in Table 2, ERP responses to IFVs appear to involve short-time components between 150 and 320 ms but also prolonged responses with latencies up to 840 ms that have not been identified in terms of specific components. Although several methodological factors may underlie the discrepancies in reported latencies, one basic factor appears to be the duration of the stimuli, as seen in Table 2. In terms of reports involving IFVs, the stimuli length in studies by Schweinberger, Walther, et al. (2011) and Gonzalez et al. (2011) provided sufficient dynamic spectral information so as permit speaker identification, whereas it can be questioned whether single vowels offer sufficient sensory information for this process (see Plante-Hébert & Boucher, 2015a).

As for investigations that focus on stimuli classified as TVs and "famous" FVs, these stimuli involve, respectively, UVs that become familiar during a training phase of an experiment, or FVs from celebrities. Importantly, an experiment by Schweinberger (2001) using a priming paradigm established that priming voices before the presentation of FVs or UVs creates a response at 200 ms post-onset indicating a voice-recognition response. However, a speaker-identity response for famous FVs was only observed in a time window between 450–800 ms (although the author did

not label these sustained potentials identification responses). Contrasting with these results, several reports using TVs have not revealed responses in windows beyond 450 ms. Thus, the MEG study of Schall, Kiebel, Maess & von Kriegstein (2014), based on TVs, used long sentence-length stimuli. After learning six voices with corresponding names, participants were asked to indicate if a speech sample and a name were matching or not. Significant responses to speaker identity were observed at 200 ms post-onset. Zäske, Volberg, Kovács & Schweinberger (2014a) similarly reported a significant difference in ERPs using an old/new task with TVs and long stimuli. TVs that were correctly identified elicited a greater positivity than UVs starting at 300 ms post-onset, although how this reflected a speaker-identity response was unclear since the responses occurred on identical linguistic stimuli (i.e., it was unclear whether identity information was processed independently of verbal contexts). A following study reported in Humble, Schweinberger, Dobel & Zäske (2019) reported a similar old/new effect bearing on speaker identity, but this effect was observed later (500–800 ms) and was elicited following the presentation of stimuli different at learning and at test. Spreckelmeyer, Kutas, Urbach, Altenmüller & Münte (2009) also reported a voice recognition response at around 300 ms post-onset during a same/different task involving pairs of UVs. Consistent with these results, Föcker et al. (2011) reported rising negativity starting at 270 ms post-onset for person-incongruent dyads of TVs compared to person-congruent ones. Yet, in a very similar study, Föcker, Best, Hölig & Röder (2012) found a significant response to paired TVs in time windows between 200 to 250 ms and 350 to 550 ms. However, with the exception of Schweinberger (2001), it is unclear how the paradigms in the preceding reports serve to distinguish responses bearing on a processing of speaker-identity information from those that reflect a *recognition* of voices. In fact, in many of the reports the terms voice recognition and voice or speaker identification are used interchangeably or with vague definitions.

Reference	Voices	Stimuli	Component(s)	Latency (ms)
Beauchemin et al. (2006)	IFV / UV	/a/	MMN, P3	200, 240–320
Graux et al. (2015)	IFV / UV	/a/	MMN, P3a	180-210, 230-320
Gonzalez et al. (2011)	IFV / UV	/ola/	N250r, P3	280–840
Holeckova et al. (2006)	IFV / UV	530 ms name	P3, Slow waves	
Schweinberger, Walther, et al. (2011)	IFV / IFV	/aba/ /igi/	P3	250-600
Schweinberger (2001)	FV / UV	2000 ms speech	Sustained potentials	450–800
Schall et al. (2014)	TV	2-syll. words	N/A	200
Humble et al. (2019)	TV / UV	1719 ms speech	Old/new effet	500–800
Zäske et al. (2014)	TV / UV	8-syll. words	N/A	290–370
Föcker et al. (2011)	UV / UV	2-syll. words	N/A	270–530
Föcker et al. (2012)	UV / UV	2-syll. words	N/A	200–250
Spreckelmeyer (2009)	UV / UV	Sung tones	N/A	300–400

Tableau 2. – Summary of ERP studies of voice processing arranged by type of stimuli and types of voices—intimately familiar voices (IFV), famous/familiar voices (FV), trained-to-familiar voices (TV) or unfamiliar/unknown voices (UV). Only time windows in relation to voice processing are reported in the table.

Overall, neural responses that relate to the recognition of TVs appear to occur in the range between 200 and 370 ms post-onset (see Table 2). The experiments of Schweinberger (2001), using FVs, yielded much later responses that could be related to speaker identification. This also applies to the report by Gonzalez et al. (2011). In comparing these studies to others in Table 2, one notices that the reported long latencies ranging from about 500 to 840 ms post-onset appear for speech contexts consisting of at least a few syllables. In understanding the differences between responses at long latencies and those that occur at about 200–370 ms, it should be weighed that stimuli of famous FVs can associate to varying degrees with a multimodal episodic memory of speakers, whereas TVs, which are experienced in a laboratory setting or through repeated audio presentations, may not serve to constitute such multimodal representations. This is not an issue when using IFVs where sensory experiences spanning years associates with the voice of an individual. Such differences could well underlie the separate responses across 200–370 ms and 500–840 ms where the first response reflects voice recognition and a later-occurring response may reflect a processing of identity information that bears on episodic or semantic memory of a speaker. However, it remains unclear whether this is actually the case given that, except for Schweinberger (2001), studies have not compared responses to TVs and IFVs. In interpreting the time windows reported in Table 2, it is interesting to note that Schweinberger,

Walther, et al. (2011) is the only study where voice identification was associated with ERP response between 250 and 600 ms post-onset. As noted, the two other studies where voice identification possibly occurred, Schweinberger (2001) and Gonzalez et al. (2011), showed responses ranging from about 500 ms to 840 ms. One potential explanation for earlier response times reported by Schweinberger, Walther, et al. (2011) is that all presented voices were IFVs although participants did not specify if they knew the speaker and were aware that any of the voices they would hear was and IFV. This accurate prediction could have facilitated the identification process and therefore fasten the EEG response.

1.2 The present study

In terms of the above research, one can surmise that EEG/MEG investigations of voice processing have not circumscribed the time course of fundamentally different processes of voice recognition and voice identification. Moreover, as summarized in Table 2, few studies focus on IFVs using sufficiently long speech samples that support accurate speaker identification (Plante-Hébert & Boucher, 2015a). Of the studies that do use stimuli consisting of at least a few syllables, separate responses appear on different time windows. Thus, while IFVs elicit responses in a 150–320 ms window, they also associate with prolonged responses as late as 840 ms post voice onset. The above discussion suggests that one reason for these prolonged responses is that IFVs and FVs carry information that links to semantic memory of a speaker such that the late responses reflect a process of voice identification.

The present study aims to bring further evidence supporting this latter view by examining the following prediction. Specifically, it is hypothesized that IFVs, compared to TVs and UVs, elicit voice recognition responses in a window of 150–320 ms, in the range of the P2 ERP component, as well as later-occurring responses extending beyond 450 ms, encompassing slow ERP waves, suggesting a distinct process of identification. This prediction also serves to clarify the effects of different types of voice stimuli, which are often indiscriminately associated with recognition and/or identification. Studies have frequently suggested similar responses for known voices regardless of whether these are IFVs, FVs, or TVs (as outlined in Table 2). Yet, as Kreiman & Sidtis (2011) note, IFVs are distinctly processed, which should reflect in differential neural responses. It

should be noted, however, that reports confirming these differential responses point to changes over long time frames (roughly 500–840 ms) and not to particular short-time ERP component (as in Gonzalez et al., 2011). Indeed, studies of responses to IFVs that refer to components such as MMNs and FTPVs have used brief stimuli like single syllables which, as noted, may not provide sufficient information for processing voice identity (cf. Beauchemin et al., 2006). For this reason, the present research is not driven by an assumption of particular components but instead explores how IFVs, TVs, and UVs elicit differential electric brain responses reflecting distinct processes of voice recognition and identification.

Method

2.1 Participants

Thirteen participants (8 females), aged between 21 and 43 years (mean = 30.81, s.d. = 5.14) completed the study. They were all native speakers of Quebec French except one speaker who learned Quebec French at four years of age. All were dominant right handers according to a standard questionnaire (Oldfield, 1971) and had normal hearing as established by an audiometric screening test. A forward and backward digit-span test ("WMS-III", Chlebowski, 2011) confirmed normal memory performance for all participants. It should be noted that participants recruited in the present study were selected following the recommendation of a member of an original pool of 36 male volunteers from which voice samples could be recorded and analyzed to create the stimuli (partly from Plante-Hébert & Boucher, 2014). Each volunteer provided the names of a family members, close friends, or life partners. The recruitment was confined to individuals who could be matched to a target IFV that presented attributes similar to those of other voices (all the voices used as stimuli represented speakers with similar F_0, as described subsequently). The “intimate familiarity” of a target IFV was established via a questionnaire and criteria that were elaborated in a previous behavioural study (Plante-Hébert & Boucher, 2014). The fact that participants were selected by reference to an IFV restricted the recruitment to a small number of specific individuals. All participants were paid, and written informed consent was obtained

following guidelines of the Ethics Committee of *CIUSSS du Nord-de-l'île-de-Montréal* at *Rivière-des-Prairies Hospital* (Montreal, QC).

2.2 Stimuli

The voice stimuli were eight four-syllable utterances, listed in Table 3, produced by 12 native speakers of Quebec French without any discernible regional accents. The length of the stimuli (4 syllables) was decided following the results of Plante-Hébert & Boucher (2015a) and other observations relating to the length of contexts required for accurate speaker identification (Bricker & Pruzansky, 1966; Legge et al., 1984; Pollack et al., 1954; Roebuck & Wilding, 1993; Schweinberger et al., 1997; Skuk & Schweinberger, 2013). These studies, especially the one from Plante-Hébert and Boucher, refer to stimuli exceeding one syllable for correct identification. Additionally, average Speaking Fundamental Frequency (SF_0) was controlled for using *Mutli-Speech* (KAY Pentax) and was similar across the voice stimuli used in the experiment (cross-speaker differences in SF_0 for the voice samples did not exceed one semitone).

Utterance stimuli	IPA transcription	Nasal segments
Bonjour madame.	[bɔ̃ʒuʁmadam]	3
Combien t'en prends ?	[kɔ̃bjɛ̃tɑ̃pʁɑ̃]	4
Comment qu'elle va ?	[kɔmɑ̃kavɑ]	2
De temps en temps.	[dətɑ̃zɑ̃tɑ̃]	3
Donne-moi en deux.	[dɔmwazɑ̃dø]	2
J'en connais quatre.	[ʒɑ̃kɔnɛkat]	2
Quand est-ce qu'il vient ?	[kɑ̃tɛskivjɛ̃]	2
Quelqu'un t'attend.	[kɛkœ̃tatɑ̃]	2

Tableau 3. – The four-syllable utterances used as voice stimuli. Transcripts in regular orthographic Quebec French and IPA.

Finally, as indicated in Table 3, each utterance contained a number of nasal sounds, which have been shown to facilitate speaker identification, likely because they provide additional information on speaker physiology in relation to resonance cavities (Amino & Arai, 2009; Amino et al., 2005, 2006; Glenn & Kleiner, 1968; Plante-Hébert & Boucher, 2015b; Su et al., 1974). The voice stimuli were produced in a conversational fashion at steady rates and recorded in a sound-treated booth

using an omnidirectional headset microphone (*C477 WRL*, AKG) and a 16-bit external sound card set to a sampling rate of 44.1 kHz (*Fast-track Ultra*, M-Audio). While recording these stimuli, the speakers produced each utterance after listening to an audio pacer consisting of separate tones. This ensured the production of similar timing and prosody across utterances. The recorded signals were amplitude normalized and each stimulus was segmented as a separate audio file. The onsets of the speech signals in the audio files were aligned so that the perceptual-center (*P-center*) of the first syllable of all utterances was at 200 ms from the beginning of the file. Alignment in terms of *P-centers* (described in Marcus, 1981; Morton, Marcus & Frankish, 1976) insures that perceptual onsets of speech stimuli are stable and reduces jitter in EEG responses at the onset (see Gilbert, Boucher & Jemel, 2014). The overall length of the signals ranged from 618 ms to 1085 ms (mean of 818 ms, SD of 83 ms). Overall, the stimuli used in the present experiment respected generally admitted guidelines for the elaboration of voice line-ups in forensic applications (Atkinson, 2015; Hollien et al., 2014; Hollien et al., 1995; Nolan & Grabe, 1996).

2.3 Pre-test stimuli validation

As a preliminary verification of the stimuli used in the present study, we conducted a pretest involving four volunteers that did not know any of the presented voices. The purpose was to establish whether equal numbers of presentations of the different voices and utterance contexts created non-specific ERPs. The test conditions were the same as during the experiment described below and each volunteer was exposed to a total of 10 trials per voice per utterance. The pretest confirmed that, in presenting different voices an identical number of times, average ERPs did not visually differ across utterance contexts. However, one of the voices had to be removed due to an unexplained difference in ERPs compared to the other voices. The pretest also confirmed that the multiple presentations of the different utterances did not have an effect on average ERPs across voices. In sum, variations in ERPs under the present test conditions can be related specifically to familiarity and frequency of presentation rather than utterance contexts or vocal idiosyncrasies.

2.4 Procedure

Audio files containing the stimuli were arranged in eight blocks, each reflecting a specific utterance of Table 3. Within each block, the voices were randomized with the restriction that no consecutive presentation contained the same voice. Of the eight blocks of stimuli, four served to record EEG responses, and these alternated with four blocks that served to collect behavioural responses on speaker identity. Specifically, the EEG-recording blocks were ordered such that the first, third, fifth, and seventh blocks each contained 240 trials of passive listening. The four other alternating blocks each contained 60 trials where listeners identified the IFV using a key press. The latter blocks of trials were reduced in number so as to limit the overall test duration while allowing to collect behavioral confirmation of IFV identification. All blocks bore presentations of different types of voices in varying proportions: a frequently presented IFV (33.33% of trials), a frequently presented TV (33.33% of trials), and twelve rarely presented UVs (each UV was presented on 2.77% of trials). Note that the 13 participants were recruited on the basis that they were intimately familiar with only one target voice (IFV) in the presented stimuli. Thus, 12 different voices were presented in the blocks but only one voice was intimately familiar to one participant.

Participants listened to the utterance stimuli via insert earphones (*E-A-Rtone 3A*, EAR Auditory Systems) and the amplitude of the audio signal was calibrated so as to obtain peak levels of 74 dBa at the inserts. The stimuli were played back using *E-prime 1.0* (Psychology Software Tools). Trials were separated by an inter-stimulus interval (ISI) that varied randomly from 500 ms to 650 ms in steps of 50 ms to minimize anticipation effects. In listening to the stimuli, the participants were sitting at a distance of 180 cm from a blank computer screen with a fixation cross. They were asked to listen to the stimuli and keep their eyes on the fixation cross. For the four behavioural blocks, participants were also required to keep the fingers of their dominant hand positioned on a mouse and to indicate as quickly as possible if the voice heard during each trial was the familiar one or unfamiliar by pressing either the left or the right mouse key, respectively (this was reversed for half of the participants).

2.5 EEG recordings and analyses

EEG signals were recorded throughout the experiment (including behavioural blocks that were not included in the EEG analyses). The recordings were performed according to the international 10–20 system and with an *ASA-lab EEG/ERP 64 channels amplifier* (ANT neuro). An online average reference was used and signals were digitized at sampling rate of 1000 Hz. Eye movements and blinks were recorded using four electrodes placed above and below the dominant eye (VEOG) and at the outer canthus of each eye (HEOG). AFz was used as ground and all other 64 channels were kept below 10 kΩ impedance during the recordings.

Offline, the recordings were band-pass filtered (0.1-30 Hz) and blinks were removed using *ASA software* (ANT neuro). All other artefacts exceeding a standard deviation of 20 μV within a sliding window of 200 ms were automatically removed with *Eeprobe GUI* (version 1.2.0.2, ANT Software). EEG recordings were then averaged across blocks and by types of voices (IFV, TV, and UV) using Fieldtrip (Oostenveld et al., 2011), an open-source toolbox for MatLab (R2017b 9.3). Each trial in the recordings was epoched between 200 ms before and 1000 ms after each stimulus onset and the 200 ms pre-stimulus interval was used for baseline correction.

Visual inspection of the averaged signal for all conditions allowed to easily identify a P1-N1-P2 complex, directly followed by a negative deflection between 300 and 350 ms post stimulus onset, in the range of the N250, and a late positive component (LPC) extending to the end of the analysis window. Considering the data in Table 2, the P2 peak on right centro-frontal sites, the N250 peak on left fronto-central sites and the LPC on both right centro-frontal sites and left/middle centro parietal sites were of particular interest in the present study.

As displayed in Figure 2, the P2 was peaking on frontal sites between 200 ms and 250 ms post stimuli onset. A mixed design ANOVA with three within subjects factors was carried out on the mean amplitudes between 200 and 250 ms. The factors included were voice condition (repeated measures; IFVs, TVs and UVs), site (F, FC, C and CP) and laterality (right and left hemispheres)

Figure 2 also illustrates that the N250 and the slow waves ERPs had a wider scope than the P2. In order to reduce statistical analyses for those two components, pools of electrodes were created

to represent six scalp regions. The regions included the following electrodes and will be referred to as: middle centro-frontal (MCF; Fz, F1, F2, FCz, FC1, FC2, Cz, C1, C2), right centro-frontal (RCF; F4, F6, FC4, FC6, C4, C6), left centro-frontal (LCF; F3, F5, FC3, FC5, C3, C5), middle centro-parietal (MCP; CPz, Pz, POz, CP1, CP2, P1, P2), right centro-parietal (RCP; CP4, CP6, P4, P6, PO4, PO6) and left centro-parietal (LCP; CP3, CP5, P3, P5, PO3, PO5). Outlining electrodes were not included since ERP of interest were not elicited on those sites.

The following statistical analyses were carried out using 50 ms mean amplitude samples to compare the time-course of ERP activity between experimental conditions (IFV, TV, UV) on the N250 and the slow waves ERPs. The analyses window for the N250 was between 300 ms and 350 ms while successive 50 ms mean amplitude samples were used between 450 ms and 850 ms post-stimulus onset to investigate longer slow waves ERPs. The mean amplitudes were calculated using MatLab (R2017b 9.3). Repeated measures analyses of variance (ANOVA) were then carried out using the open-source software JASP (version 0.13) with two within-subjects factors: voice condition (IFV, TV and UV) and scalp region (6 levels: MCF, RCF, LCF, MCP, RCP and LCP). Huynh-Feldt correction was applied if required and the alpha level was set at $p < 0.05$.

Results

3.1 Behavioural data

On the analyses of behavioural responses, all responses exceeding 1300 ms were excluded (22.30%). For the remaining trials, the overall accuracy of identification of the IFV was 98.18%. The false alarm rate, that is when either TV or UV were falsely identified as an IFV, was 0.35%. The misses, or when IFV was designated as UV, represented 1.82% of response. Most of the time, when participants made mistakes, they spontaneously informed the experimenter that they were aware of their error. These results establish that the voice stimuli were readily identified by participants.

3.2 The P2

For the P2. the three levels repeated measures ANOVA carried out on mean amplitudes between 200 ms and 250 ms reached significance for the main effect of site, $F(1.883,22.595) = 17.348$, $p < .001$, $\eta^2 = .297$ as well as for the interaction between laterality x voice conditions, $F(2,24) = 3.868$, $p = .035$, $\eta^2 = .01$. Planned comparison with Bonferroni-corrected t-tests for the voice conditions on each sites revealed significant differences between IFV and TV on F4, $t(12) = -2.3$, $p = .04$ $d = .638$ and between IFV and UV on FC4, $t(12) = 2.498$, $p = .028$, $d = .693$, C4. $t(12) = -2.25$, $p = .044$, $d = .624$ and CP4, $t(12) = -2.716$, $p = .019$, $d = .753$. These results are summarized in Figures 1 and 3.

3.3 The N250

The second component was analysed using the mean amplitude between 300 ms and 350 ms post stimuli onset. In this window, the two-factors with repeated measures ANOVA revealed a main effect of scalp region, $F(2.78,71.397) = 12.308$, $p < .001$, $\eta^2 = .436$, and a voice condition x scalp region interaction, $F(5.95,71.397) = 3.593$, $p = .004$, $\eta^2 = .028$. Planned comparison with Bonferroni-corrected t-tests for the voice conditions within each given scalp region revealed a significant difference between TV and UV in MCF, $t(12) = -3.479$, $p = .006$ and LCF, $t(12) = -3.506$, $p = .005$, $d = -.972$ regions as well as between IFV and TV in LCF, $t(12) = 3.12$, $p = .014$, $d = .865$. No other difference was observed in this time window.

Since the main difference between TVs and UVs was the varying number of presentations in the course of the experiment, two ANOVAs with within-subject factors of voice condition and region were performed, respectively on the first and second halves of the experiment, to ascertain training effects. Only the ANOVA on the second half revealed significant interaction between voice condition and region $F(4.428,53.137) = 3.544$, $p = .01$, $\eta^2 = .047$. Again, Bonferroni-corrected planned comparison with voice conditions within each individual scalp region showed significant differences between TV and UV in MCF, $t(12) = -3.215$, $p = .011$, $d = -.892$ and LCF, $t(12) = -2.624$, $p = .045$, $d = -.728$ regions as well as between IFV and TV in LCF $t(12) = 3.068$, $p = .016$, $d = .851$.

3.4 The LPC

Finally, the LPC was investigated using a two-factors with repeated measures on a 50 ms sliding window of mean amplitudes between 450 ms and 850 ms post stimuli onset. Significant voice condition x scalp region interactions were found for 500–550 ms, $F(6.719,80.632) = 3.411$, $p = .003$, $\eta^2 = .036$, 600–650 ms, $F(7.506,90.073) = 3.956$, $p < .001$, $\eta^2 = .043$ and 650–700 ms, $F(5.98,71.754) = 2.406$, $p = .036$, $\eta^2 = .041$ time windows.

As before, planned comparison were used to compare the voice conditions within each individual scalp region using the Bonferroni correction for multiple comparisons. In the 500–550 ms time window, a significant difference between IFV and TV was found in RCF, $t(12) = -2.854$, $p = .026$, $d = -.792$ and LCP, $t(12) = -3.188$, $p = .012$, $d = .884$. For the 600–650 ms time window, there was a significant difference in mean amplitudes between IFV and both TV and UV in RCF, $t(12) = -3.056$, $p = .016$, $d = -.847$, $t(12) = -3.418$, $p = .007$, $d = -.948$, and in LCP $t(12) = 3.221$, $p = .011$, $d = .893$, $t(12) = 4.192$, $p < .001$, $d = 1.163$. A significant difference between IFV and UV was also found for this time window in MCP, $t(12) = 2.920$, $p = .022$, $d = .810$. Finally, the planned comparisons for the 650–700 ms window revealed significant differences between IFV and UV in MCP, $t(12) = 2.782$, $p = .031$, $d = .7725$ and LCP, $t(12) = 2.608$, $p = .046$, $d = .723$. As one can see by observing Figure 1, IFV was significantly different from both TV and UV in the early time window (200–250 ms) and for at least two windows within the LCP. Meanwhile, both TV and UV never differed significantly with the exception of the 300–350 ms time window during which UV was also different from IFV. The main regions involved were RCF for the P2 and the LCP, the MCF and LCF for the N250 and the MCP and LCP also for the LPC.

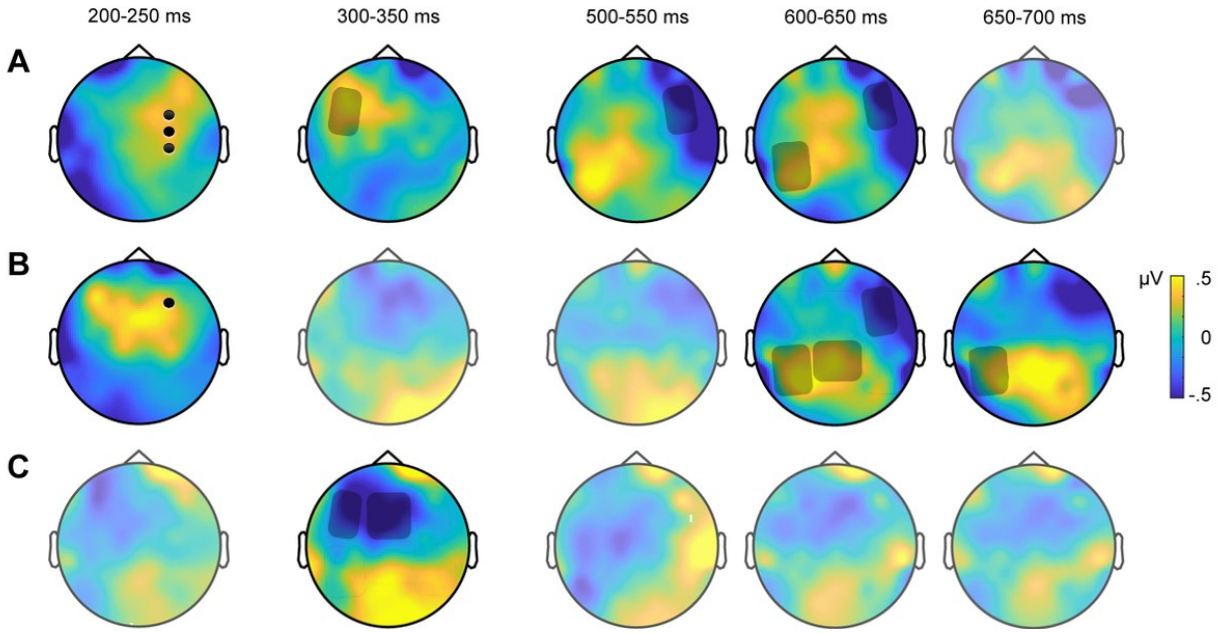

Figure 1. – Topographic representations of *the ERP differences* between (A) IFVs and TVs, (B) IFVs and UVs, and (C) TVs and UVs. Darkened areas and black dots represent regions and electrodes where voice conditions were significantly different. No significant difference were found on light-shaded topographies.

To provide a broad picture of the results across the three time windows of interest, Figure 2 offers a summary of responses across the six regions. One can see that listeners' neural responses to IFVs—where the speaker's identity is known—stand out in the early time window in the RCF region on the P2, and in the late time window in the MCP, LCP and RCF on LCP. By contrast, listeners' responses to TVs—for which the identity of the speaker is not known—appear in other regions, in this case at MCF and LCF regions in a mid-latency time window corresponding to the N250.

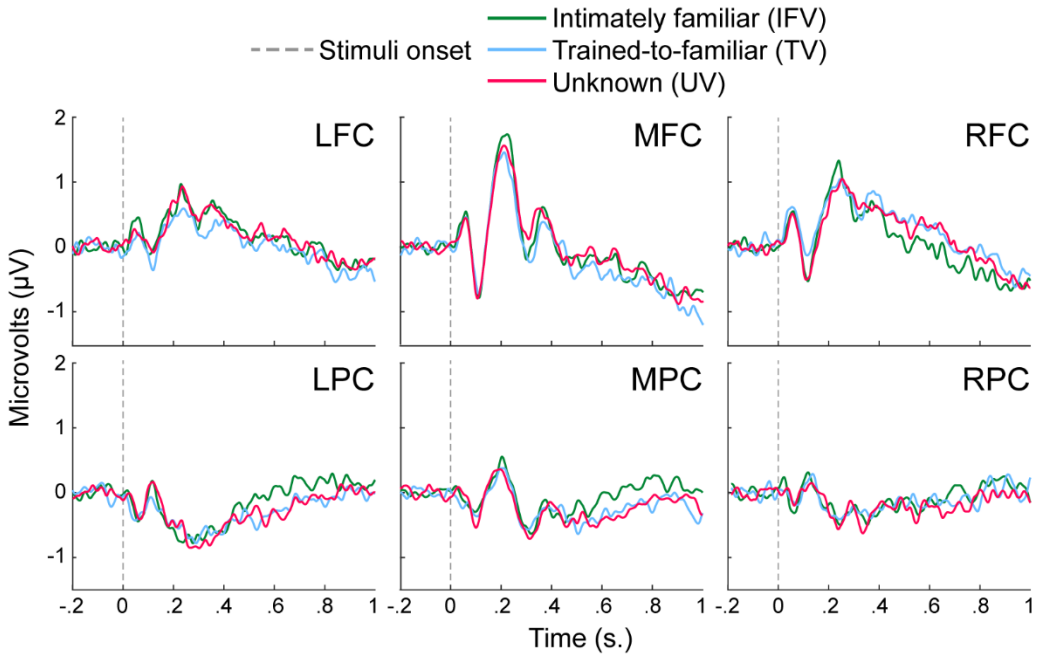

Figure 2. – ERPs on the six regions illustrating the effects of IFVs, TVs, and UVs in the three time windows of interest. Distinct responses to IFV appear in an early window of 200–250 ms and were rightward as seen changing amplitudes at RCF, and also appeared in a late window of 500–650 ms where prolonged shifts appear in parietal sites at MCP and LCP and in frontal sites at RCF. For TVs and UVs, contrasting responses were found in a mid-late window of 300 to 350 ms, as seen in the differential responses at MCF and LCF.

Discussion

When one hears the voice of a close individual or a famous voice, one can "recollect" information that has to do with the identity of the speaker (Curran, 2000). Intuitively, one knows *who* is speaking. This is inherently different from simply recognizing a voice as previously heard but where one may not recollect a particular speaker or "place" the voice. The purpose of this study was to substantiate this difference with respect to voice research where only some protocols distinguish between voice *recognition* and *identification* processes by reference to intimately familiar voices (e.g., Beauchemin et al., 2006; Gonzalez et al., 2011; Graux et al., 2015; Schweinberger, 2001; Schweinberger, Walther, et al., 2011). In circumscribing neural responses that reflect these different processes, the use of IFVs presents an advantage in that, compared to

famous voices where identity information can vary across individual listeners, there is little doubt the voice of a parent, sibling, or close-friend holds specific information on speaker identity. In this sense, the above results confirm a basic difference on processes of voice recognition and identification and suggests a time course for these processes not previously identified in the literature bearing of speaker identity processing.

Specifically, significant distinctions in ERPs were observed in three different time windows: an early response in a 200–250 ms time window associated with the P2 component, a mid-latency response at between 300–350 ms, in the N250 range, and a later occurring response between 500–700 ms. Both early- and late-latency responses were associated intimately familiar voices (in the IFV condition) compared to frequently heard or rarely heard unfamiliar voices (in the TV and UV conditions). No significant differences were observed between responses for TVs and UVs in these time windows. While some studies have also revealed such specific early ERPs and components such as MMNs for familiar voices, many have not reported later-occurring protracted responses that cannot be analyzed in terms of short-time "components". Part of the reason for this discrepancy in research findings appears to bear on the length of the stimuli. As noted previously with reference to Table 2, studies using short speech samples (single syllables) obtained results related to voice identity in a similar time range as the early responses observed in the present report. Conversely, studies in Table 2 where participants are presented with at least a few syllables have reported later-occurring responses to IFVs similar to the late responses obtained in the above results. In particular, Gonzalez et al. (2011) used stimuli lasting about 500 ms and reported responses between 280 and 840 ms post-onset. Schweinberger (2001) had stimuli of 2000 ms and obtained responses ranging from 250 to 600 ms and, using stimuli of 909 ms, Schweinberger, Walther, et al. (2011) reported responses from 450 to 800 ms post-onset. In the present results, short phrases averaging 793 ms elicited responses between 500 and 700 ms. There is, then, a degree of agreement in these reports on the fact that stimuli longer than a syllable associate with later-occurring responses to voices that bear inherent speaker-identity information. This leads to two complementary accounts of why such responses would be drawn out beyond about 500 ms post onset.

One reason may be that accurate voice identification requires more dynamic spectro-acoustic information than what is obtained in the span of single syllables. On this possibility, the results of Plante-Hébert & Boucher (2014; 2015a) showed that, although identification of intimately familiar speakers can be obtained on single syllables, quasi-perfect identification requires a few syllables. In other words, short voice samples may not provide sufficient sensory information for an associative process relating signals to a memory of speakers. In addition to this factor, a delay is likely to take place between simply recognizing sensory attributes and the associative process as the stimuli unfolds over time.

Still regarding the LPC in the present data, electrophysiological studies have previously shown that responses to known stimuli associated with semantic information stored in long-term memory occur later than responses to stimuli encountered before but not associated with additional contextual information (for more detail, see Wilding & Ranganath, 2011). Moreover, the left-parietal old/new effect, associated with the recollection of semantic information about a given stimuli, is known to occur at similar latencies (500–800 ms post stimuli onset) and, as its name hints, in the left-parietal brain region. The description and latency of this left-parietal old/new effect greatly correspond to the data observed in the present experiment.

As for the mid-latency response, frequently presented voices (TV condition) elicited significantly distinct responses from rarely presented voices (UV condition). Yet this difference at mid-latencies was absent in the beginning of the experiment and grew stronger at the end. It is interesting to note that such training effects on ERPs have also been reported in studies using familiar and unfamiliar faces, as described by Tanaka, Curran, Porterfield & Collins (2006). This evolution of responses with experience suggests specifiable neural markers of memory encoding (although heard voices in the TV condition were not accompanied by any episodic memory of particular situations involving the speakers). Thus, frequent presentations of both facial and vocal stimuli entail changing neural responses in a window of 230–320 ms post-onset. As Tanaka et al. (2006) note, there are reasons to believe that the N250 is not modality specific and can represent a developing perceptual expertise. In fact, a report by Schall, Kiebel, Maess & von Kriegstein (2013) revealed that if a listener only hears a familiar voice without seeing the speaker, cortical face-processing areas are activated. The response observed also greatly corresponds, both in

latency and in scalp distribution, to the well known mid-frontal old/new effect specifically associated with the feeling of having encountered a stimuli before without recalling detailed semantic information about it (for detailled reviews, see Curran, Tepe & Piatt, 2006; Wilding & Ranganath, 2011). With this in mind, our results, combined with those of previous experiments on both speaker identity and various memory processes, suggest that given sufficient speech material, the distinction between the well established “remember” and “known” processes also applies to voice identity in terms of elicited components.

Although EEG is ill-suited to a localization of processes, the above analyses of ERP responses offer some parallels with the model of voice perception presented by Kreiman & Sidtis (2011, p. 187-189). This model suggests a right-hemisphere processing of familiar voices as opposed to a left-hemisphere processing of unknown voices. Also, most neuroanatomical models assume that the processing of familiar voices involves the right superior temporal sulcus (e.g., Belin et al., 2004; Maguinness et al., 2018; Stevenage, 2018). The issue of localization is important in understanding the transition between an episodic memory of voices and the consolidation of a semantic memory of speakers, an issue that requires further research when it comes to speaker identity. However, the above results suggest that future investigations should adopt a strict distinction between voice recognition and identification in devising protocols. These terms can serve to characterize different processes and responses relating to types of vocal stimuli, such as the above categories of IFV, TV, and UV. It should also be a central consideration that the processing of identity information in voices operates on heard speech sounds that extend beyond a single syllable and that neural responses to IFVs are relatively late and drawn out. This suggests that methods which examine neural responses over stretches of speech, such as temporo-spectral coherence analyses, may be better suited to analyzing the processing of voice information than techniques that focus on short-time ERPs and their components.

Conclusion

In short, our study offers EEG evidence supporting a distinction between processes of voice recognition and speaker identification in relation to neural markers arising at different latencies. In addition to establishing unambiguous differences between vocal recognition and identification,

the preceding findings bear implications in the applied sector of forensic earwitness testimony. Traditionally, earwitness identification of speakers relies on the perceptions of listeners, which has been shown to be highly accurate, especially in the case of familiar voices (Plante-Hébert & Boucher, 2015a, 2015b). The present data indicate that there are, additionally, neural correlates of both familiar speaker identification and the recognition of frequently heard voices as opposed to voices that are occasionally heard. Further investigations should serve to clarify the conditions by which unfamiliar voices become highly familiar and how this relates to neural encoding processes reflecting a transition between an episodic and semantic memory of vocal information.

Acknowledgement

This research was partly funded by SSHRC and FQRSC scholarships awarded to Julien Plante-Hébert and by a grant of the Fonds Québécois de la Recherche sur la Nature et la Technologie awarded to Victor J. Boucher (FQRNT No. 175811).

References

References of the present paper were included in the general bibliography

Figures

Figures and figure captions were included in the text

Discussion

L'expérience présentée dans ce premier article avait comme objectif d'observer les PÉ associés avec le traitement neuronal de voix intimement familières. Pour y arriver, 13 participants ont écouté passivement les voix de locuteurs intimement familiers (VF), inconnus, mais fréquemment présentées (VE) et inconnus et rarement présentées (VI). Des processus distincts se sont manifestés dans trois fenêtres temporelles différentes. Premièrement, une composante P2 a été analysée entre 200 et 250 ms après le début des stimuli. Cette P2 était directement suivie d'une N250, analysée entre 300 et 350 ms. Finalement, des PÉ de longue durée ont été analysés entre 450 et 850 ms.

Familiarité

L'analyse des PÉ a permis d'observer que l'amplitude de la P2 engendrés par la présentation de VF est significativement plus grande que celle observée suite à la présentation de VE et de VI. Ces observations ont été faites sur des sites fronto-centraux droits (F4, FC4, C4 et CP4). Comme l'illustre le Tableau 2 de cet article, la latence de cette P2 correspond aux résultats généralement rapportés dans des tâches de discriminations de voix ou reconnaissance de locuteur. L'analyse des ondes lentes observées plus tardivement a révélé des différences significatives dans trois fenêtres temporelles de 50 ms chacune (500-550, 600-650 et 650-700 ms). Dans les trois cas, les comparaisons planifiées ont indiqué que VF se distinguait de VE et de VU tandis qu'aucune différence n'était à noter entre VE et VU.

Ensembles ces résultats illustrent que les voix familières sont traitées différemment des voix inconnues, et ce, en deux étapes. La première, brève et relativement hâtive, correspond à la P2 et représente la reconnaissance de la voix. Plus tardivement, et d'une ampleur plus importante, l'identification se produit et affecte la LPC. Comme décrit dans les modèles présentés au Chapitre 1, les processus d'identification nécessitent assurément plus de temps puisqu'il découle de l'accès à la mémoire à long terme et à un vaste système de mémoire sémantique.

La N250

La négativité qui suivait immédiatement la P2, une N250, a quant à elle démontré une différence significative d'amplitude moyenne principalement entre les VE et les VI dans la région fronto-

centrale gauche. Rappelons qu'avant le début de l'expérience, tant les VE que les VI étaient complètement inconnues des participants. La seule distinction entre ces conditions était la fréquence de présentation plus élevée pour les VE (33 % des essais) comparativement aux VI (2,77 % des essais chacune). De plus, une analyse subséquente a révélé que cette différence entre VE et VI n'était pas présente en début d'expérience. Des données similaires ont été rapportées par Tanaka et al. (2006) qui ont observé une réponse spécifique de la N250 suite à la présentation répétée de visage. Les auteurs suggèrent d'ailleurs que la N250 ne soit pas spécifique aux visages, mais plutôt au développement d'une expertise perceptuelle, ce que soutiennent également Kaufmann, Schweinberger et Burton (2009). Bien que la N250 ait été rapportée principalement pour des stimuli d'ordre visuels, les résultats de Gonzalez et al. (2011) en font mention dans un paradigme portant entre autres sur la reconnaissance d'individus par la voix. Ces données conjuguées à celles de la présente expérience suggèrent donc que la N250 peut servir à marquer l'encodage en mémoire d'information liée à l'identité sans que celle-ci soit associée à un contenu sémantique.

En somme, l'article 1 a permis de faire une distinction claire entre les processus de reconnaissance et d'identification du locuteur supportée par des données électrophysiologiques. Cette différenciation, déjà bien documentée dans le domaine visuel, s'illustre par l'accès, ou non, à des informations d'ordre sémantique à propos de la personne qui parle. Lorsqu'il y a identification, l'identificateur serait en mesure, par exemple, de nommer le locuteur ou encore de décrire à quelle occasion ils ont fait connaissance. La reconnaissance, quant à elle, fait référence à la conviction d'avoir déjà entendu une voix, sans pour autant être en mesure de donner plus d'information au sujet du locuteur. Les distinctions observées mettent en lumière l'importance du type de paradigme expérimental et des stimuli utilisés dans la recherche à propos du traitement de l'identité dans le cerveau.

Dans une perspective légale, les résultats obtenus démontrent que certaines composantes électrophysiologiques représentent des marqueurs fiables du traitement de voix intimement familières par rapport à des voix inconnues, mais également par rapport à des voix simplement entendues auparavant sans pour autant que le locuteur soit connu.

Chapitre 3 : Multimodalité, interaction et apprentissage des voix

La première partie de cette thèse a permis d'établir que les voix inconnues, même si déjà entendues auparavant, ne suscitent pas les mêmes réponses électrophysiologiques que les voix intimement familières. La raison principale évoquée pour expliquer cette distinction au niveau du traitement neuronal est que les voix intimement familières, contrairement aux voix inconnues, font appel à des informations d'ordre sémantique au sujet du locuteur. En comparaison, la représentation mentale de locuteurs inconnus n'est élaborée qu'à partir des caractéristiques acoustiques de leur voix et ne possède que peu d'information sémantique, souvent issue d'inférences. Lorsque des informations d'ordre sémantique sont impliquées, on parle alors d'identification du locuteur. Dans le cas où aucune information sémantique sur un individu n'est accessible, mais qu'une voix est reconnue comme ayant déjà été entendue, on parle plutôt de reconnaissance du locuteur.

Le Chapitre 3 porte quant à lui sur les facteurs qui permettent d'apprendre une voix suffisamment bien pour arriver à en identifier le locuteur. Autrement, comment la familiarité se développe-t-elle ?

Effets multimodaux

Comme il a été discuté dans le Chapitre 1, plusieurs modalités peuvent influencer le processus d'identification des voix. Les deux plus courantes sont par la vision et par l'audition. Les différents modèles présentés dans ce même chapitre font d'ailleurs tous état des connexions entre les systèmes perceptuels de ces deux modalités. Bien que ces connexions soient généralement admises, leur fonctionnement demeure sujet à débat.

Comme il en sera question dans le second article de cette thèse, plusieurs travaux ont fait était d'une préférence générale envers l'information visuelle pour traiter l'identité. Selon certains, cette préférence nuit à la mémorisation des caractéristiques vocales d'un individu pour mieux

emmagasiner l'information visuelle. Cet effet est nommé l'effet d'ombrage du visage (*face overshadowing effect*, FOE).

Une autre perspective sur le rôle de l'information multimodale lors de l'apprentissage est que celle-ci permet de consolider les apprentissages sous forme de mémoire épisodique. Ainsi, les différentes informations contextuelles connues au sujet d'une personne contribuent à son identification, peu importe la modalité de ces informations. Par exemple, le visage d'une personne pourrait aider à se souvenir de sa voix.

Ces deux points de vue sont a priori contradictoires. L'étude de Zäske, Mühl et Schweinberger (2015) est néanmoins parvenue à les réconcilier en suggérant que le FOE opère au début de l'apprentissage d'une voix, mais qu'au fur et à mesure que la représentation de l'individu se consolide en mémoire, l'information multimodale devient profitable. L'expérience 2 a comme objectif de tester ces différentes hypothèses en observant les PÉs lors de l'identification de locuteurs appris selon différentes modalités.

Effets interactionnels

L'expérience présentée dans le prochain article abordera aussi un aspect de la mémoire épisodique moins souvent étudié : l'interaction entre les locuteurs. Bien que la notion d'interaction puisse comprendre de nombreux paramètres, c'est à travers deux d'entre eux qu'elle sera explorée.

Dans un premier temps, plusieurs études ont démontré que le contact visuel entre deux interlocuteurs favorise l'attention (Helminen, Pasanen et Hietanen, 2016; Hirotani, Stets, Striano et Friederici, 2009; Hood, Macrae, Cole-Davies et Dias, 2003; Vuilleumier, George, Lister, Armony et Driver, 2005). Il émane de ces études qu'un contact visuel aiderait à l'apprentissage de mots chez les bébés et également au rappel de visage tant chez les enfants que chez les adultes. Cet effet serait principalement attribué au fait que d'un point de vue social, le contact visuel permet d'établir que les deux individus sont engagés dans une activité conjointement. Dans cette perspective, il semble raisonnable de prédire que le contact visuel favoriserait l'encodage et le rappel de la voix de la même manière qu'il le fait pour les visages.

La seconde composante de l'interaction entre locuteurs qui sera abordée dans l'expérience 2 est la réponse motrice. Lors d'une discussion, les tours de parole alternent entre les individus impliqués. Il n'est donc pas uniquement question d'aspects perceptuels, mais également d'une composante motrice. Une étude récente a d'ailleurs fait état d'une meilleure mémorisation lorsque les items étaient dits à voix haute que lorsqu'ils étaient répétés mentalement (Lafleur et Boucher, 2015). Ainsi, l'agentivité présente lors d'une interaction verbale enrichirait l'épisode mnésique et favoriserait l'encodage et le rappel.

Expérience 2

Objectifs et hypothèses spécifiques

L'objectif de la deuxième expérience est d'examiner les effets de l'information contextuelle lors de l'apprentissage sur l'identification de locuteurs. Plus spécifiquement, les effets de l'information visuelle et de l'interaction seront observés. En considérant les résultats obtenus à l'expérience 1, les hypothèses sont que :

1. Les voix explicitement apprises auront une réponse spécifique sur les composantes P2 et LPC, peu importe l'information contextuelle présente à l'apprentissage
2. Les réponses observées sur la P2 et la LPC seront tout de même modulées en fonction de la présence d'information contextuelle à l'apprentissage.
3. Un effet d'ombrage du visage affectera les composantes impliquées (P2 et/ou LPC)

Méthodologie

Participants

Au total, 18 participants (9 femmes) âgées entre 21 et 30 ans (moyenne = 25, σ = 3) ont pris part à cette expérience. Ils étaient tous des locuteurs natifs du français et avaient un niveau de scolarité universitaire. Tous étaient droitiers et avaient une audition et une mémoire jugées normales par les mêmes tests qu'utilisés dans l'expérience 1.

Comme pour l'expérience 1, tous les participants ont été rémunérés après avoir lu et signé un formulaire de consentement approuvé par le Comité d'éthique à la recherche du *CIUSS du Nord-de-l'île-de-Montréal* à *l'Hôpital Rivière-des-Prairies* de Montréal. Encore une fois, les participants avaient l'opportunité de poser des questions spécifiques à l'expérience avant de signer le formulaire de consentement.

Stimuli

Les enregistrements qui ont servi de stimuli ont été faits à partir d'une liste de 500 lexèmes dissyllabiques représentant des noms communs (voir les Tableaux 7 et 8 en annexe). Ces lexèmes

ont été sélectionnés à l'aide du logiciel de rédaction Antidote 9 (version 5.1, Druide informatique) puisque, comparativement à d'autres outils lexicographiques disponibles, ce logiciel représente avec plus de justesse le FQ. Les critères de recherche qui ont permis la sélection des lexèmes étaient la catégorie grammaticale (nom commun), le nombre de syllabes (2), le niveau de langue (neutre) et la fréquence d'occurrence (fréquents ou très fréquents). De plus, les lexèmes obtenus ont été filtrés manuellement pour ne conserver que ceux composés de quatre ou cinq phonèmes afin de restreindre autant que possible les configurations syllabiques tel qu'illustré dans le Tableau 4.

Nb. De phonèmes			
1ère syll.	2e syll.	Total	Nb. d'occurrence
1	3	4	55
3	1	4	0
2	2	4	177
2	3	5	196
3	2	5	55
1	4	5	17
4	1	5	0

Tableau 4. – Liste des différentes structures syllabiques possibles pour les mots dissyllabiques en français comprenant quatre ou cinq phonèmes. Le nombre d'occurrence de chacune de ces structures dans les stimuli est également rapporté.

En suivant ces critères de sélection, un total de 1987 lexèmes a été obtenu. De cette liste préliminaire, tous les homophones ont été trouvés et seule une entrée par dyade (ou triade) a été conservée. Trois juges ayant une formation pertinente en linguistique ont ensuite été chargés de marquer tous les lexèmes jugés saillants en fonction des champs lexicaux suivants : sexualité, dégout, drogues, religions, violence, maladie et mort. Ils devaient également marquer tous les lexèmes liés à la nationalité, au lexique enfantin et aux variantes jugées régionales du français (hors Québec). Tous les mots ayant été marqués par au moins deux juges ont été retirés. Les mêmes trois juges ont ensuite fait une relecture de la liste des lexèmes restants pour les classer selon leur concrétude sur une échelle de Likert à 4 niveaux (Robinson, 2014) avec les catégories

abstrait, *légèrement abstrait*, *légèrement concret* et **c***oncret*. Seuls les lexèmes jugés légèrement concrets et concrets par au moins deux des juges ont été conservés.

À partir de la liste des 500 lexèmes restants, les stimuli audios ont été enregistrés par 14 locuteurs natifs du FQ. Tous les locuteurs étaient des hommes avec une diction jugée normale et sans accent régional marqué par rapport au FQ urbain. Ils avaient également des fréquences fondamentales moyennes parlées (F_0mp) qui variaient d'au plus 1 semi-ton entre elles. Afin d'assurer une similitude prosodique sans avoir à manipuler les enregistrements en différé, un guide rythmique semblable à celui utilisé par Gilbert et al. (2014) a été utilisé. En suivant cette approche, les locuteurs écoutaient un métronome répliquant la structure prosodique désirée à plusieurs reprises avant le début de la séance d'enregistrement et à plusieurs reprises pendant celle-ci. Trois prises jugées correctes étaient enregistrées pour chaque lexème de manière à ce que l'expérimentateur puisse sélectionner a posteriori celle qui correspondait le mieux au guide rythmique en termes de débit et d'intonation. Tous les enregistrements ont été faits depuis une salle à l'épreuve du bruit avec un microphone Audio-technica (modèle AT831b) et d'un adaptateur Shure (modèle X2u) à un taux d'échantillonnage de 44,1 kHz avec le logiciel Goldwave (version 6.31). La fonction de filtre passe-haut intégrée du microphone a été utilisée pour atténuer les basses fréquences à une fréquence de coupe de 80 Hz. Ce filtre a permis de réduire les bruits résiduels dus à l'éclairage, à la ventilation et à l'équipement électronique à l'intérieur de la salle. Tous les stimuli ont été enregistrés séparément en format *.wav*.

Stimuli d'apprentissage

Trois locuteurs dont les voix seraient à apprendre par les participants (nommés L1, L2 et L3, voir procédure) ont chacun enregistré deux fois les mêmes 20 lexèmes à l'aide du même microphone que mentionnée ci-dessus et d'une caméra Web HD Pro C920 (*Logitech*). Une attention particulière a été portée à l'apparence physique des locuteurs pour s'assurer qu'ils se ressemblent (tailles, poids, barbes, cheveux, teints de peau, etc.) et un chandail noir uniforme leur a été fourni pour la durée des enregistrements.

Pour la première séance d'enregistrement, les locuteurs devaient fixer une marque sur le bureau d'ordinateur devant eux alors qu'ils prononçaient les énoncés. De cette manière, la caméra

captait leur visage, mais pas leur regard pendant les enregistrements. Dans la deuxième séance d'enregistrement, les locuteurs devaient regarder directement la caméra pendant la production des lexèmes. Cette deuxième version des stimuli d'apprentissage avait pour but de simuler les effets connus du regard sur la mémoire qui ne sont présents que lorsque le regard est direct (Manesi, Van Lange et Pollet, 2016), mais qui se maintiennent même si l'individu n'est pas physiquement présent (Conty et al., 2010).

Pour chaque stimulus d'apprentissage, la trame sonore provenant de la caméra Web HD Pro C920 a été remplacée par celle du microphone Audio-technica. Les enregistrements ont ensuite été utilisé pour élaborer trois conditions d'apprentissage pour chaque lexème et chaque locuteur : audio seulement (A), audiovisuel (AV) et audiovisuel interactif (AVI). Au total, 180 stimuli d'apprentissage distincts ont été enregistrés.

Les fichiers *.wav* contenant les stimuli d'apprentissage ont présentés de façon à ce que le début de chaque première syllabe. De cette manière tous les fichiers débutaient 1000 ms avant le début acoustique des stimuli et se terminaient 5000 ms après celui-ci.

Stimuli expérimentaux

Les 480 lexèmes restants ont été enregistrés uniquement à l'aide du microphone Audio-technica et du matériel précédemment décrit. Les locuteurs L1, L2 et L3 ont chacun enregistré 120 lexèmes différents parmi ceux-ci et un nouveau locuteur dont la voix n'a pas été apprise en a enregistré 60. Dix autres nouveaux locuteurs (désignés VR) ont chacun enregistré 6 lexèmes différents parmi les 60 déjà enregistrés par L4 ainsi que 6 nouveaux lexèmes propres à chacun d'entre eux. Au total, ces enregistrements ont produit 540 stimuli expérimentaux dont la durée moyenne, du début du signal acoustique à la fin de celui-ci était de 574 ms (σ = 105 ms).

Une fois les enregistrements des stimuli expérimentaux terminés, les fichiers *.wav* de chaque stimulus ont été alignés de façon à ce le début de la première syllabe se situe 200 ms du début du fichier sonore et 1000 ms de sa fin. Cette différence temporelle par rapport à l'alignement des stimuli d'apprentissage s'explique par la présence de contenu visuel dans lors de l'apprentissage tandis que les stimuli expérimentaux n'étaient qu'acoustiques.

L'amplitude de tous les stimuli, tant d'apprentissage qu'expérimentaux, a été normalisée par locuteur avec le logiciel *Goldwave* (version 6.31) afin que les sommets d'amplitude atteignent 90 % de la capacité maximale du système. De plus, la fonction de réduction de bruit de Goldwave a été utilisée avec le préréglage « *reduce hum* » pour minimiser le bruit de fond résiduel de l'éclairage et de l'équipement électronique non filtré par le microphone Audio-technica.

Notons qu'une partie des stimuli élaboré dans le cadre de cette expérience a été utilisée pour une seconde partie de l'expérience décrite ici. Cette seconde partie ne fait pas l'objet du présent travail de thèse.

Procédure

À leur arrivée, les participants étaient informés que l'expérience serait d'une durée approximative de trois heures, tel qu'indiqué dans le formulaire de consentement, et qu'elle serait divisée en trois parties : un entrainement, la partie 1 et la partie 2. Notons que la deuxième partie de l'expérience n'est pas rapportée dans le présent travail de thèse. Les participants étaient également informés que les directives de chacune de ces parties leur seraient transmises à chaque étape. Chaque participant a été testé individuellement.

Entrainement

Lors de la phase d'entrainement, les participants étaient confortablement assis dans une pièce sans bruit et bien éclairée à une distance de 85 cm de l'écran de 15,6" d'un ordinateur portable *Asus* (model X580VD) 64 bits. Cet ordinateur était équipé d'une carte graphique *NVIDIA* (modèle Geforce GTX 1050) et l'écran avait une résolution de 1920x1080 pixels avec une fréquence de rafraichissement de 60 Hz. La sortie audio de l'ordinateur a été calibrée avec un sonomètre de marque RadioShack (modèle Cat 33-2055) de manière à atteindre des sommets de 78 dBa à la sortie. La carte de son de l'ordinateur était une Conexant SmartAudio HD et des écouteurs à inserts EAR Auditory Systems (E-A-Rtone 3A) ont été utilisés tant pour l'apprentissage que pour les deux parties expérimentales.

Avant d'entreprendre la phase d'apprentissage, les locuteurs dont les voix étaient à apprendre étaient brièvement présentés aux participants. Pour ce faire, les 10 mêmes stimuli d'apprentissage parmi les 20 enregistrés étaient présentés pour chaque locuteur, un locuteur à

la fois. Chaque locuteur était présenté selon une des trois conditions d'apprentissage (A, AV ou AVI) et cette condition était maintenue pour toute la durée de la phase d'apprentissage. L'ordre d'apparition des locuteurs pour cette présentation ainsi que la condition selon laquelle ils étaient présentés étaient balancés entre les participants. Pendant cette présentation, les participants recevaient la directive de porter attention aux caractéristiques de chaque locuteur en vue d'apprendre leur voix. Uniquement pour la condition AVI, les participants recevaient la directive d'établir un contact visuel avec le locuteur à l'écran et de répéter le lexème à haute voix après lui.

Le reste de la procédure pour la phase d'apprentissage a été inspiré par l'étude de Zäske et al. (2014a) et fonctionnait sous forme de cycles apprentissage-test. Les blocs d'apprentissage consistaient en la présentation des 10 mêmes stimuli d'apprentissage que pour la présentation initiale pour chaque locuteur en ordre aléatoire pour un total de 30 stimuli par bloc. Comme mentionné précédemment, les conditions d'apprentissage étaient les mêmes que pour la présentation. Un symbole, propre à chaque locuteur, était affiché dans le coin supérieur gauche de l'écran pendant la présentation de chaque stimulus. Les participants recevaient la directive de mémoriser la voix de chaque locuteur et de l'associer au symbole respectif. Pour faciliter la tâche d'identification à venir dans les blocs de test, les participants devaient appuyer sur la touche d'un clavier de réponse ou les symboles associés aux locuteurs apparaissaient de façon à mémorise d'emblée les séquences motrices requises

Après chaque bloc d'apprentissage avait lieu un bloc de test présenté à l'aide du logiciel E-prime 3 (Psychology Software Tools). Les mêmes stimuli d'apprentissage étaient présentés à nouveau dans ces blocs à la différence près que tous étaient audio seulement. Les participants recevaient la directive d'indiquer, avec le clavier de réponse, le mieux possible et le plus rapidement possible de quel locuteur il était question après chaque essai. Compte tenu de la difficulté de la tâche, particulièrement au début, la fenêtre de réponse était d'une durée de 5000 ms. Les participants étaient d'ailleurs informés de cette difficulté de la tâche et pour y pallier, une rétroaction apparaissait après chaque réponse pour indiquer si elle était bonne ou mauvaise et quelle était la bonne réponse si nécessaire. Pour que l'apprentissage soit considéré complété, les participants devaient obtenir un score de 24 sur 30 ou plus pour trois cycles consécutifs. Les électrodes étaient ensuite installées et un dernier cycle apprentissage-test comportant les 10 stimuli

d'apprentissage n'ayant pas encore été présentés était alors fait. Ce dernier bloc avait comme objectif de valider que l'apprentissage se maintenait dans le temps et n'était pas dépendant du contenu linguistique. Le même score était nécessaire à la poursuite de l'expérience et tous les participants ont réussi ce bloc avec succès d'un seul coup.

Tâche expérimentale

La partie expérimentale impliquait des enregistrements EEG et a eu lieu dans une autre salle dont l'éclairage était légèrement tamisé. Les participants y étaient assis à 130 cm d'un écran d'ordinateur Lenovo Thinkvision. Les stimuli étaient joués en version audio uniquement et avec les mêmes écouteurs à inserts que pour l'apprentissage. L'expérience se déroulait à l'aide du logiciel MatLab (R2015b 8,6) sur un poste de travail HP Z210 CMT Workstation muni d'une carte de son interne et d'une carte graphique 64 bits. L'amplitude à la sortie était calibrée avec le même sonomètre que pour l'apprentissage de manière à atteindre des sommets de 78 dBa.

Enregistrements EEG

Les enregistrements EEG ont été faits avec le même matériel et selon les mêmes paramètres que l'expérience 1. Le nettoyage des données (artéfacts, clignements d'yeux, filtrage) s'est fait en différé lui aussi avec le même matériel et selon les mêmes paramètres que pour l'expérience 1. Les analyses ont-elles aussi été complétées avec l'outil *Fieldtrip* pour *MatLab*. Le signal EEG a été filtré (.01-30 Hz) et seuls les essais associés à des identifications correctes ont été conservés.

Chaque essai a été découpé en débutant 200 ms avant le début acoustique d'un stimulus jusqu'à 1000 ms après ce même point. L'intervalle de 200 ms précédant le début du stimulus a été utilisé pour la correction en fonction de la ligne de base. Le moyennage des données a été fait par conditions.

Résultats

Les résultats de l'expérience 2 sont présentés dans l'article suivant.

Article 2

Effects of speech modalities in acquiring neural markers of voice recognition: An ERP experiment using voice lineups

Julien Plante-Hébert[a]*, Victor J. Boucher[a] & Boutheina Jemel[b, c]

[a]Laboratoire de Sciences Phonétiques, Département de Linguistique et de Traduction, Université de Montréal, Montréal, QC, Canada

[b]Laboratoire de Recherche en Neurosciences et Électrophysiologie Cognitive, Hôpital Rivière-des-Prairies, Montréal, QC, Canada

[c]École d'Orthophonie et d'Audiologie, Faculté de Médecine, Université de Montréal, Montréal, QC, Canada

*Correspondence: Julien Plante-Hebert; julien.plante-hebert@umontreal.ca

Abstract

Interactions between modalities during identity processing have been reported by previous studies. This suggests that information stored in a specific modality, for example visual, can influence the processing of a speaker's voice.

In the present study, ERPs were used to investigate the effects of contextual information at learning on speaker identification. During a training phase, 18 participants had to learn the voices of three speakers. Each speaker was presented in a specific modality: audio (A), audiovisual (AV) and audiovisual with simulated interaction (AVI). Once the training was successfully completed, EEG recordings were made during a speaker identification task. The task included the voices of the three speakers and that of a new and unknown speaker which served as baseline. ERP analyses indicated that all three conditions led to significant differences on the same components when compared to the unknown speaker (P2 and PC). The time range of the LPC showed a Face overshadowing effect (FOE) in the AV condition compared to A. This FOE was, however, cancelled by the addition of simulated interaction.

Combined with previous observations, these results indicate that speaker recognition might not be affected by contextual information whereas speaker identification is. The implications of our results are discussed in regard to the literature on the FOE and their application in forensic phonetics.

Keywords ERP, speaker identification, voice familiarity, audiovisual, interaction, mid-frontal old/new effect, left partietal old/new effect.

1 Introduction

In hearing people speak, listeners can normally recognize and identify a voice from sounds alone. This remarkable ability to learn and identify voices arises at the earliest stages of human development, even before birth (Kisilevsky et al., 2003). Of course, in later development, voices are usually learned in the context of speech communication where sounds can be accompanied by visual information of a speaker and other multisensory cues from face-to-face interactions. It is acknowledged, following the terminology of Tulving (1972), that these multimodal *episodes* of vocal communication come to constitute one's *semantic* memory of voices, and that this encoded information can be activated upon hearing individuals speak (e.g., Barsalou, 2016; Matheson & Barsalou, 2018). The encoding, however, implies distinct processes in that learning to recognize and identify a voice not only entails an ability to perceptually discriminate between sounds. It also requires an associative process serving to link sounds to learned semantic attributes of speaker identity, which is the focus of the present report.

The distinction between perceptual and associative processes is broadly reflected in current neuroscientific models that posit a two-system architecture of voice recognition/identification (Blank, Kiebel & von Kriegstein, 2015; Gainotti, 2015; Maguinness et al., 2018; Perrodin, Kayser, Abel, Logothetis & Petkov, 2015). This standard view, partly inspired by an approach to facial recognition (Haxby, Hoffman & Gobbini, 2000), suggests that there is a core system which processes modality-specific perceptual information, and a connected "extended" system that activates associated semantic representations of speaker identification, such as a face, a situation, a name (etc.) in a putative modality-independent manner (Belin, Bestelmeyer, Latinus & Watson, 2011; Belin et al., 2004; Gobbini & Haxby, 2007; Haxby et al., 2000). Perhaps the most convincing evidence of these two systems is provided by case studies of phonagnosia, as in a recent report by Roswandowitz, Schelinski & von Kriegstein (2017); and for reviews of other cases:(De Renzi, Faglioni, Grossi & Nichelli, 1991; Garrido et al., 2009; Hailstone et al., 2010; Luzzi et al., 2017; Roswandowitz, Kappes, Obrig & von Kriegstein, 2017; Roswandowitz et al., 2019; Roswandowitz et al., 2014; Stevenage, 2018; Van Lancker, Cummings, Kreiman & Dobkin, 1988). That study, involving fMRI, described two rare cases of developmental apperceptive and

associative phonagnosia. Behavioral tests (Roswandowitz et al., 2014) showed that both individuals presented severe deficits in learning to recognize voices when the learning required associating a face, a name, or a colour with a voice, as well as a reported deficit in discriminating pitch variations. However, for the individual with an apperceptive phonagnosia, tests requiring a judgment of whether voice samples were from the same or different speakers revealed that voice discrimination was impaired. For voices that could be discriminated, semantic association was intact. By contrast, the other individual presented an associative phonagnosia where voice discrimination was intact but semantic association was impaired. The main fMRI observations, which were based on a differentiation of brain activity on tasks of voice and speech recognition, showed distinct activity clusters compared to those of the control group. Apperceptive phonognasia corresponded to decreased activity in auditory regions (in Heschl's gyrus, planum temporale, and superior temporal gyrus), whereas associative phonagnosia corresponded to lower activity levels in both the right posterior areas of the middle/inferior temporal gyrus and the amygdala. There is little question that such case studies, irrespective of localisation issues that prevail, offer strong support for distinguishing between perceptive and associative processes of voice recognition and identification. On the other hand, applying the view that an amodal associative system underlies the learning and recognition/identification of voices still presents several challenges.

In particular, given that associations serving to identify voices develop in the multisensory context of speech, neural markers of voice processing may reflect the effects of various sensory signals that carry identity information on different timelines. For example, it has been established that visual presentations of familiar faces evoke a P250 response subsequent to the "face-sensitive" N170 (Bentin & Deouell, 2000; Caharel, Poiroux & Bernard, 2002; Marzi & Viggiano, 2007). By comparison, audio presentations of familiar voices evoke later-occurring responses, and reported latencies vary widely both on early evoked components (ranging from 200 to 350 ms post-stimuli) and late sustained potentials (extending from about 450 ms to 800 ms or longer; for a review, see Plante-Hébert, Boucher & Jemel (2021)). Indeed, there are no agreed-upon neural markers of voice recognition/identification, which creates a degree of confusion in the literature. The difficulty partly owes to the fact that recognition of familiar voices, contrary to the recognition of

familiar faces, operates on unfolding acoustic information over an extent of speech. On this point, it is important to note that behavioral tests have shown that accurate voice identification in highly controlled voice lineups—even in the case of intimately familiar voices such as those of a parent or life partner—requires a span of speech reflecting a few syllables and, even then, particular articulatory motions like those of nasal sounds can have an effect (Amino & Arai, 2009; Amino et al., 2005, 2006; Plante-Hébert & Boucher, 2014, 2015a; Su et al., 1974). Accurate voice identification is *generally not obtained* with monosyllables like "aaah". It should also be weighed that pitch is not wholly voice-specific but varies with different vowel motions (Speaks, 2017). This brings to light that the associative system underlying "voice identification" critically rests on a perceptual processing of dynamic acoustic attributes reflecting motions of speech (Bricker & Pruzansky, 1966; Pollack et al., 1954; Roebuck & Wilding, 1993). Said differently, speech and voice processing are inextricably linked and so methods which consist in differentiating or subtracting neural responses of speech recognition from those of voice recognition may blur the critical role of dynamic attributes of speech in voice identification.

But determining the neurophysiological markers of associative processes is further complicated by the fact that the responses evoked by voices vary with listeners' experience of face-to-face speech interactions. Disentangling the influence of speech modalities on voice learning is thus an essential step in defining neural markers of voice recognition and identification, and requires techniques offering a high temporal resolution on neural responses. In the present work, electroencephalography (EEG) is used in determining how speech modalities in learning voices affect listeners' evoked response potentials (ERPs) upon hearing trained and untrained voices.

1.1 The influence of audiovisual speech on voice learning

Investigations of voice recognition and identification present diverging conclusions on the effects of learning voices with and without faces (for recent reviews providing other viewpoints on voice research than the one that follows, see Maguinness et al., 2018; Stevenage, 2018). Several results, some of which relate to controversies on legal applications of voice and face recognition, question received views of the benefits of audiovisual information on voice recall. Specifically, reports involving various behavioral paradigms have shown that, in the learning of voices, presentations

of faces can negatively impact subsequent voice recognition (Cook & Wilding, 1997a; Cook & Wilding, 2001; McAllister, Dale, Bregman, McCabe & Cotton, 1993; Stevenage, Hale, Morgan & Neil, 2014; Stevenage, Howland & Tippelt, 2011, and the review of Stevenage 2018). This has been called the "face overshadowing effect" (FOE). Related studies also indicate that within-modality visual associations can have a stronger effect on identity judgements than cross-modal audiovisual associations. For example, using an associative priming paradigm, Stevenage et al. (2014) found that the vocal primes have a weaker effect on face recognition than when both the prime and target are two celebrity faces. Similar effects have been obtained with videos of dynamic faces, and dynamic faces hidden with a balaclava (Heath & Moore, 2011; Tomlin, Stevenage & Hammond, 2016).

Yet, contrary to these reports, many investigations have confirmed that presentations of voices with associated faces have a stronger effect on subsequent recognition responses than voice-only presentations (Armstrong & McKelvie, 1996; Legge et al., 1984; O'Mahony & Newell, 2012; Robertson & Schweinberger, 2010; Schweinberger, Kloth, et al., 2011; Schweinberger, Robertson & Kaufman, 2007; Zäske et al., 2015). These diverging results on the effects of audiovisual information can well reflect differences in methodology across sectors of voice-recognition research, especially with respect to types of voice stimuli. Some studies oriented by questions of the accuracy of eye- and ear-witness testimony (e.g. Cook & Wilding, 1997a; Cook & Wilding, 2001; McAllister et al., 1993) refer to stringent protocols of face and voice lineups where, for instance, voice samples can have similar F_0 to within a semitone and involve a control of speaker dialect and idiosyncrasies (Hollien et al., 2014; Hollien, Huntley, Kunzel & Hollien, 2013; Nolan, 1997). Although such stringent control is not often applied in voice research, one could question whether audiovisual presentations of highly similar voices might not as such favor a processing of identifying facial cues in an associative memory of voices. Moreover, a unique study by Zäske et al. (2015) has served to clarify that the *amount* of audiovisual experience can be a pivotal factor. On this important effect, it should be mentioned that, while Cook & Wilding (1997a) and Cook & Wilding (2001) reported a face overshadowing effect on voice recognition for a heard utterance, the effect dissipated when three utterances were presented, suggesting that experience in

hearing voices rapidly alters the influence of visual information (as noted by Maguinness et al., 2018).

This latter effect was made clear in the study by Zäske et al. (2015) by comparing listeners' ability to recognize voices across 12 "study-test cycles". Each cycle involved an initial learning phase, where voices were presented alone or with static or dynamic faces, before a testing phase. Comparisons across the cycles indicated that, while presented faces initially drew attention and thereby hampered voice recognition (FOE), audiovisual presentations had significant facilitating effects in subsequent learning cycles. In sum, effects of face overshadowing decreased as semantic associations of voices became more robust. But another pivotal finding was that the beneficial effects of audiovisual information appeared specifically for voices presented with congruent facial motions in contrast to static images, which suggests an enhancement of voice memory by way of visualized speech motions (Sheffert & Olson, 2004).

With respect to voice recognition, several reports have shown that audiovisual experience of a speaker rapidly benefits subsequent identification of heard voices (Schall et al., 2013; Schelinski, Riedel & von Kriegstein, 2014; Von Kriegstein, Dogan, Giraud, Kleinschmidt & Kiebel, 2008; Von Kriegstein, Kleinschmidt & Giraud, 2006). These investigations have involved various controlled conditions of voice learning or familiarisation including conditions where voices are presented alone, with dynamic facial information, or with accompanying visual cues such as symbols or names. The main finding across studies is that training involving facial information has a superior effect on voice identification (Maguinness et al., 2018). However, these reports do not specify the acoustics of presented of voices such as F_0 or other attributes of voice discriminability (and few details are provided on facial stimuli). Recent studies on voice recognition have demonstrated that such factors as voice discriminability have a important impact on the recognizability of a given voice (Bülthoff & Newell, 2015; Foulkes & Barron, 2000; Schweinberger, Kawahara, Simpson, Skuk & Zäske, 2014; Stevenage et al., 2018). In failing to control these attributes, one is left wondering what it is in heard voice stimuli that associates with facial information and serves to identify the voices. On this question, the frequent assumption in voice research that speech and voice recognition are separate, conflicts with the general finding that voice recognition is enhanced by dynamic faces (not static images) and that such effects can arise precisely because

facial information couples to the acoustics elements of speech reflecting *motions*. Again, accurate voice recognition is generally not obtained with near-motionless voice sounds like “aaah”, as noted earlier.

Of course, other aspects of dynamic faces unrelated to speech motions have memory-enhancing effects. For instance, a body of work attests to the influence of eye gaze and eye-related “directed attention” on the encoding of faces and verbal expressions (Conty, N’Diaye, Tijus & George, 2007; Conty, Tijus, Hugueville, Coelho & George, 2006; Farrant & Zubrick, 2012; Helminen et al., 2016; Hirotani et al., 2009; Hood et al., 2003; Jiang, Borowiak, Tudge, Otto & von Kriegstein, 2017; Macrae, Hood, Milne, Rowe & Mason, 2002). Very few studies have examined the effects of such factors on voice encoding and speaker recognition and identification. However, one intriguing study by Hammersley & Read (1985), which focused on ear-witness testimony, revealed that participation in a conversation with a speaker has a superior effect on subsequent voice identification compared to a passive listening of conversations. The enhancing effect of talking to someone can also extend to a memory of spoken items (Lafleur & Boucher, 2015). Such observations tend to support the idea of a coupling of production and perception in voice recognition/identification. In fact, it is well established that hearing, or hearing and seeing someone speak, activates many of the cortical regions that are involved in producing speech (Watkins, Strafella & Paus, 2003; Wilson, Saygin, Sereno & Iacoboni, 2004). In sum, speech production and perception both entail activation of coupled, sensorimotor representations that may extend to representations of voices in their dynamic aspects. Compared to passive listening, active speaker-listener interaction engages attention, the neurophysiological response reflecting the selective processing of sensory information, which can benefit a memory of voices. Yet such effects have not been the object of voice research. The experiment described below examines how voices learned in audio and visual modalities including modalities of speaker-listener interaction can affect neural responses of voice recognition. However, the degree of control of faces and voices that are presented in the training phase of the experiment differ from those of previous studies and refer to standards of voice lineups. This approach offers specific advantages in circumscribing modality effects, as explained subsequently.

1.2 The present study

As outlined in the above discussion, conflicting reports on the role of facial information in voice recognition have emerged from different sectors of research and present diverging results owing principally to the types of stimuli that are used and the amount of training given to listeners in experiments of voice familiarization. In terms of the stimuli, some investigations bearing on the reliability of ear- and eyewitness testimony refer to protocols such as voice lineups that greatly reduce that variability of presented faces and voices in identification tasks (e.g., Hollien et al., 2014; Hollien et al., 1995; Hollien et al., 2013; Nolan & Grabe, 1996; Wells et al., 2020; Yarmey, 2014). Such control of face and voice stimuli is not the general practice in neurophysiological investigations but is nonetheless relevant in ascertaining the effects of audio and audiovisual training on neural markers of voice recognition. For one thing, salient attributes of faces or speech can engage attention on a particular modality which can impact on listeners' associative memory of voice identity and, consequently, on neural responses of voice recognition. As for the amount of training required to obtain responses reflecting a recognition of "familiar" voices, it has been unclear that responses to famous voices or voices learned in laboratory settings resemble those evoked by intimately familiar voices. Indeed, much of the problem in defining neural markers of voice recognition has to do with the inherent variability of personal experiences with voices. To address this problem, a previous study compared the ERPs of trained and intimately familiar voices that were all highly similar as in standard lineups (Plante-Hébert et al., 2021). That study revealed differing ERP components P2, N250 (or a P2-N250 complex), and a late positive component (LPC) that varied significantly for trained and intimately familiar voices. The present study builds upon these results and aims to determine the effects of modalities of voice learning on ERPs of voice recognition and identification where voice variability is again minimized as in standard lineups. To investigate such effects, three controlled learning conditions (from now on referred to as "training conditions") are used involving different modalities of stimuli presentation. These include an audio modality (A) where listeners are trained on heard voices presented alone; an audiovisual modality (AV), in which voices are trained by viewing and hearing speakers; and an audiovisual modality with interaction (AVI), where listeners repeat phrases in response to heard and seen speakers when they gaze at the listener. The design basically involves

a training phase followed by an experimental phase as in other investigations except that, in the present study, all training stimuli involve highly similar faces and voices following the standards of lineup protocols. Such an approach offers the specific advantage of minimizing the variability of face and voice stimuli that can otherwise blur the effect of training modalities on subsequent neural responses to recognized voices.

2. Methodology

2.1 Participants

Eighteen native speakers of Quebec French (9 females) aged between 21 and 30 years (mean = 25, s.d. = 3) took part in the experiment. All were dominant right handers according to a standard questionnaire (Oldfield, 1971) and had normal hearing as established by an audiometric screening test. A forward and backward digit-span test ("WMS-III", Chlebowski, 2011) confirmed that all the individuals presented normal memory performance. All participants were paid volunteers, and written informed consent was obtained following the guidelines of the Ethics Committee of *CIUSS du Nord-de-l'île-de-Montréal* at *Rivière-des-Prairies Hospital* (Montreal, QC).

2.2 General design

The design of the present investigation included a voice-training phase and an experiment phase. In the training phase, participants were trained to identify by way of symbols on a keypad three target voices, V1, V2, V3, that were separately presented in three modality conditions: audio (A), audiovisual (AV), and audiovisual with interaction (AVI), as described in the following. Thus, there were three trained voices (TVs), individually learned in three different modality conditions. The modality condition of each speaker varied across participants as well as their order of first appearance. Once the voices were successfully acquired, participants proceeded to the experiment phase at which point EEG was recorded during audio presentations of TVs and an unknown voice (UV).

2.3 Speech and voice stimuli

The training and experiment stimuli were elaborated by reference to a list of 260 two-syllable words or lexemes representing common nouns (see Table 6 and 7 annexed). The lexemes were selected using functions in the writing software *Antidote 9* (version 5.1, Druide informatique) which provides more accurate indices of word usage in Quebec French compared to indices of other databases. Items that were selected were neutral in meaning and represented frequent or very frequent concrete words (e.g., *piano*, *banane*, etc.). Recordings of the production of these words by four native speakers of Quebec French served as stimuli. These speakers had normal pronunciation with no discernible regional accent or salient idiosyncratic articulation. It is also useful to note that these speakers were selected from an original pool of 57 male volunteers, all of whom had provided speech samples which were acoustically analyzed (Plante-Hébert et al., 2021). Following these analyzes, speakers whose voices had similar average speaking fundamental frequency (SF0) to within a 1 semitone were retained for the present study (SF_0 min: 116.90 Hz , max: 123.98 Hz, mean: 120.83 Hz). In other words, the perceived pitch of the speakers' voices across utterances was practically identical. The test stimuli involved recordings of speech audio of all four speakers, and the training stimuli involved both audio and video recordings of three of the four speakers, representing target voices V1, V2, V3. The degree of similarity of visual and facial characteristics of the speakers representing these target voices is described below.

2.3.1 Experimental stimuli

These stimuli were audio recordings of the four speakers. Three speakers representing the TVs (V1, V2, V3) and the fourth speaker (V4), who served as a baseline representing untrained UV stimuli, each produced 60 different words. Recordings of these recited words were performed in a sound-attenuating booth using an external sound card (Shure, X2u) set at a sampling rate of 44.1 kHz and 16-bit resolution, a Lavalier microphone (Audio-Technica, AT831b) and recording software (Golwave 6.31). During the recording, and to obtain stimuli with similar prosody, a rhythmic guide was used (as described by Gilbert et al., 2014). With this technique, speakers hear a pacer consisting of rhythmic tones just before they produce words, and this serves to obtain regular prosody. The files were recorded in mono.wav format and amplitude normalized relative to /a/ sounds. The average duration of the files from signal onset to offset was 574 ms (s.d. 105 ms).

2.3.2 Training stimuli

These were audio and audio-video recordings of 20 words recited by the three speakers representing the TVs (V1, V2, V3). The productions were recorded in mp4 format using a webcam (Web HD Pro C920, Logitech), a 64-bit graphic card (ex. NVIDIA Quadro K5200), and subsequently edited using DaVinci Resolve 14 (Blackmagic Design). The soundtrack of the audiovisual files was later replaced by the higher quality recording using the aforementioned audio equipment. For the audiovisual training stimuli, onsets of visual displays were followed by a delay of 1000 ms before the onset the acoustic signals, and a delay of 5000 ms followed acoustic offsets. In the videos, the head and face of the speakers were similarly positioned on the screen (1080p) and displayed against a neutral background. The facial features and the clothing of the speakers were also similar, as illustrated in Figure 3. The 20 speech contexts serving to train target voices V1, V2, and V3 were randomly assigned to three training conditions across participants (A, AV, AVI), giving a total of 180 stimuli. Finally, an important aspect of the stimuli is that, during the recording of produced lexemes in the AV condition, the speakers looked down at a mark on the table in front of them whereas, for the AVI condition, speakers' eye gaze was directed at the camera during the production of words before returning to a downward gaze (as illustrated in the screenshot of Figure 4). This was meant to capture the gaze effects which occurs when there is eye contact (Manesi et al., 2016), and it should be noted that this effect is present even if the speaker is not physically present (Conty et al., 2010). In addition to eye gaze, the AVI included oral repetitions by the participant (as described in the "Procedure").

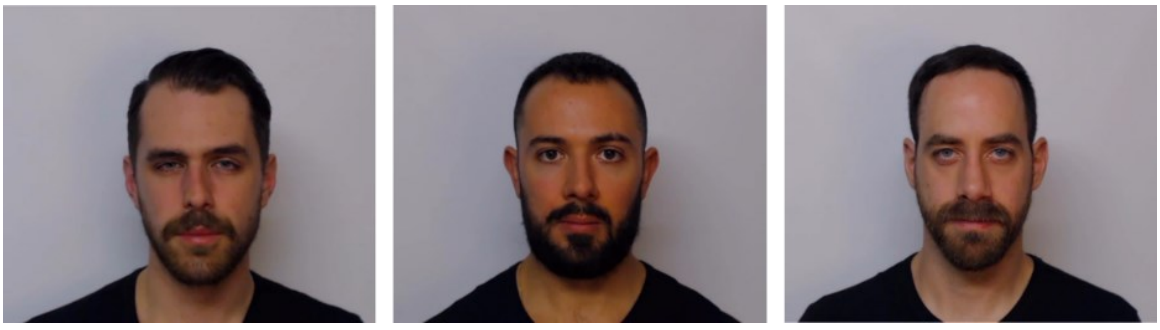

Figure 3. – Screenshots of the 3 speakers representing trained voices V1, V2 V3 (see the text for further details).

2.4 Procedure

Upon their arrival in the test room, participants were informed that the different tasks would last about three hours, as mentioned in the consent form, and that it was divided in three parts: a training phase followed by two experimental parts. There was also a second test unrelated to the present study. Participants were instructed that task-specific details would be given before each training and experimental parts. They were also told that pauses were possible upon request.

2.4.1 Training phase

The training involved an introductory presentation of the stimuli followed by repeated training-test cycles, similar to the approach described by Zäske, Volberg, Kovács & Schweinberger (2014b). The audio and audiovisual stimuli were presented via a laptop computer with a 15.6" screen, a 64-bit graphic card, and an attached keypad where keys were symbol-coded for the target TVs. Participants sat in front of the laptop in a quiet room and attended to presented audio and audiovisual recordings while wearing insert earphones (EARtone 3A). Sound output to the ears was calibrated using a sonometer and an insert adaptor so as to obtain peak levels for /pa/ of 78 dBa.

In the introductory presentation, 10 identical speech contexts representing target voices V1, V2, V3 were played back in the same order, one speaker at a time for a total of 30 trials. The three voices were, however, presented in different training conditions (A, AV, AVI) that were randomly assigned and counterbalanced across participants such that all voices and conditions were played back an equal number of times in a training block. The participants were instructed to focus on each speaker's characteristics so as to better remember their voice. Only for the AVI condition, they were also told to look at a speaker in the eyes and repeat the heard word aloud. Figure 4, provides an example of the three conditions as they were presented on trials.

In the training-test cycles, the same 10 speech stimuli used in the introduction were presented but in random order. Each cycle consisted of a training block of 30 trials followed by a test block of 30 trials presented using E-prime 3 (Psychology Software Tools). During the training blocks, participants had to learn to associate a voice with a symbol on the keypad and this symbol

was also visually displayed on slides that accompanied the presented recordings (see Figure 4). During the test blocks, only audio versions of the training stimuli were presented and participants had to identify the voices using the keypad (within 5000 ms at most). They were informed that the task could be difficult at the beginning, and to answer as best as they could. Feedback, including the correct answer, was provided on the screen. The training and test blocks were repeated until participants could correctly identify speakers on at least 23 of the 30 trials in three consecutive cycles. It took on average 4.5 cycles (min. = 3, max. = 7, s.d. 2.43) for the participants to complete the training phase. On the final test block, participants correctly identified the speakers on 84.6% of the trials (min. 73.33 %, max. 100%, s.d. 8.1%).

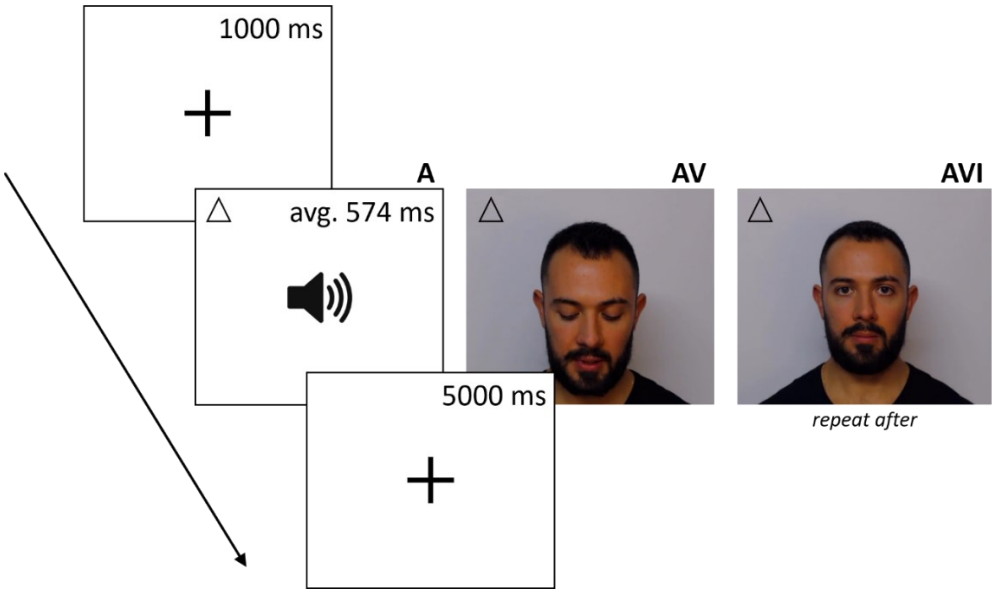

Figure 4. – Illustration of a training trial for a given speaker, with the 1000 ms pre-stimulus delay, and the different training conditions (A, AV and AVI) followed by 5000 ms post-stimulus delay. The symbols in the top left corner (a triangle in this case) represent an associated symbol of a keypad used in the identification task.

Following the above training phase, electrodes were installed on the participants' scalp, and following this installation, just before the experimental phase, a last training-test cycle was administered. Ten new training stimuli were used for this cycle so as to validate the obtained voice-identification scores. On this final test, participants' correct identification of voices across trials ranged from 73.33% to 100%, with an average of 85.34% (s.d. 8.43%).

2.4.2 Experimental phase

EEG recordings were performed at this phase. Participants sat at approximately 130 cm from a Lenovo Thinkvision computer monitor displaying a fixation cross and wore ear inserts as in the training phase. The same symbol-coded keypad was also used to record voice-identification responses. The experiment stimuli consisting of audio recordings only were played back in two continuous blocks separated with a pause. Stimuli were presented using MatLab (R2015b 8.6) and playback involved a HP Z210 CMT Workstation with its internal sound card and 64-bits graphic card. As with the training stimuli, sound output at the ears was calibrated so as to obtain peak levels of 78 dBa. The participants were required to rapidly identify, after each heard word, whether the voice was V1, V2, V3, or "other" using the symbol-coded keypad.

2.5 EEG Recordings and analyzes

EEG signals were recorded in eight continuous blocks for each participant using the international 10–20 system with *ASA-lab EEG/ERP 64 channels amplifier* (ANT neuro) with an online average reference at a 1000 Hz sampling rate. The eye movements and blinks were recorded using four electrodes placed above and below the dominant eye (VEOG) and at the outer canthus of each eye (HEOG). AFz was used as ground and all other 64 channels were kept below 10 kΩ impedance during the recordings.

Offline, the recordings were band-pass filtered (0.1-30 Hz) and blinks were removed using *ASA software* (ANT neuro). All other artefacts in the EEG exceeding a standard deviation of 20 µV within a sliding window of 200 ms were automatically removed with *Eeprobe GUI* (version 1.2.0.2, ANT Software). All subsequent analyzes including ERP averaging across individual trials and across participants as well as statistical analyses were performed using Fieldtrip (Oostenveld et al., 2011), an open-source toolbox for MatLab (R2017b 9.3). EEG recordings were then averaged across blocks according to the training conditions. Only trials associated with correct responses were included in the ERP averages. The average time window of separate epochs was set between 200 ms before and 1000 ms after each stimulus file onset. The 200 ms pre-stimulus interval was used for baseline correction.

Before analyzing the behavioral data, trials with response times (RT) exceeding 2 standard deviations from the average value of a given participant were excluded (2.08%).

In terms of ERP analyses, these were performed first via visual inspections of global field powers (GFPs) which were calculated using *Fieldtrip*. By representing the global brain activation across the scalp at each time point, GFPs served to circumscribe time windows of peak brain activity (Lehmann et Skrandies, 1980). To illustrate the approach, Figure 6 shows that a negative-going peak of activity appeared at 96 ms followed by a positive peak at 189 ms post stimuli onset. These peaks correspond respectively to components known as N1 and P2. Some later-occurring peaks were observed including a visually salient protracted peak extending beyond about 550 ms. To investigate these components, statistical analyzes were performed across two 25 ms windows surrounding the N1 peak and four 25 ms windows surrounding the P2 peak (respectively, from71 ms to 121 ms and from 139 to 239 ms post onset). As can be seen in Figure 5, two smaller peaks of activity appeared at 329 ms and 406 ms. These were also investigated using two 25-ms-windows (extending, respectively, from 304 to 354 ms, and 381 to 431 ms post stimulus sound onset). Finally, a later-occurring and prolonged component corresponding to the LPC was also investigated and this component presented a peak at 771 ms. Since this LPC was slightly skewed, the time window for the analyses was determined visually using the GFPs and analyses were carried out using 50-ms-windows extending from 550 ms to 900 ms.

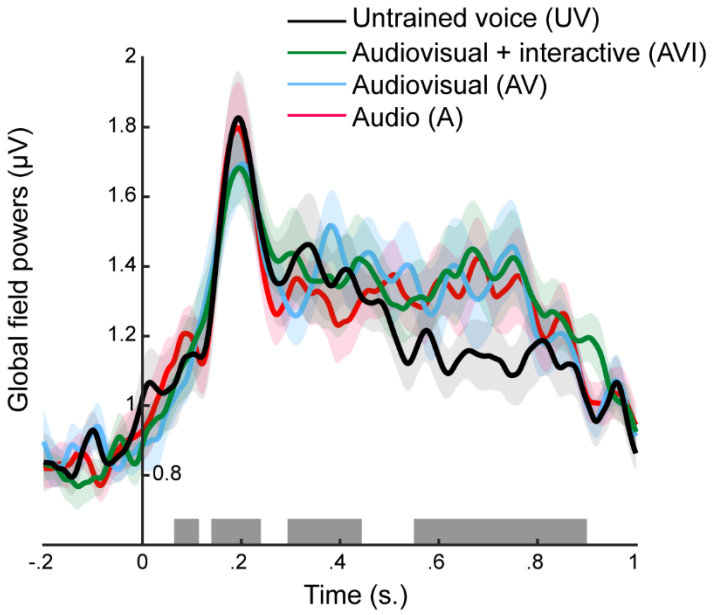

Figure 5. – Global field power (GFP) representing peaks of activity (in µV) for the duration of epochs. The darker lines represent the averages for all participants and the shaded areas their range Grey boxes over the time axis show the time windows used in statistical analyses. Note the major positive peaks at between 139 and 239 ms post-stimuli onset (P2), and between 550 and 900 ms (LPC).

For each of the aforementioned short-time-windows, statistical analyses using a Monte-Carlo method were performed with the purpose of comparing responses to each training condition to UV. The method involved using non-parametric cluster-based random permutation tests as implemented in Fieldtrip (Monte-Carlo method; Kroese, Taimre & Botev, 2013; Maris & Oostenveld, 2007). This procedure entails the following steps. First, a dependent-sample *t* test across conditions is calculated for each electrode site and clusters are formed based on neighboring electrodes where at least two adjacent electrodes' *t* values reach an α level of 0.05 (for a two-tailed test). Neighboring electrodes were specified via Fieldtrip's triangulation method (on average 5 neighbors). Then, for each cluster, a cluster-level statistic was calculated by summing the individual *t* statistics within a cluster and then comparing these with a randomized distribution of test statistics (with 5000 draws per test with randomly permuted condition labels across participants). Clusters were considered significant at $p < 0.05$.

3 Results

3.1 Behavioral results

The percentages of correct identification of the speakers presented during the experimental task (hits) and the corresponding average RTs are summarized in Table 5. A repeated measures analysis of variance (ANOVA) revealed a main effect of training condition on the RTs [$F(3, 45)=11.98$, $MSE = 47678.6$, $p<0.000$, $\eta^2=0.444$]. Post hoc tests using Bonferroni correction for multiple comparisons indicated that the strong difference largely owed to the contrast between TVs in the three training conditions and UV (all three TV conditions were significant at $p <0.001$). However, the same post-hoc tests did not reveal significant differences amongst the three TV conditions on either correct responses (Hits) or RTs. In short, the behavioral results confirmed the effects of training on participants' identification of the three voices but did not show differences between training conditions.

Condition	Hits (%)	RTs (ms)	SD (ms)
A	68.37	1856	450
AV	67.55	1823	431
AVI	64.46	1848	469
UV	33.58	2191	391

Tableau 5. – Averages and standard deviations of correct identifications (Hits), and response times (RTs) per training conditions (Audio, Audiovisual, Audiovisual interactive, and baseline UV).

3.2 ERP

As mentioned earlier, statistical analyses focused on specific time windows corresponding to activity peaks of GFPs. Figure 6 provides a broad picture of the ERPs observed across conditions using electrodes where components of interest were best seen.

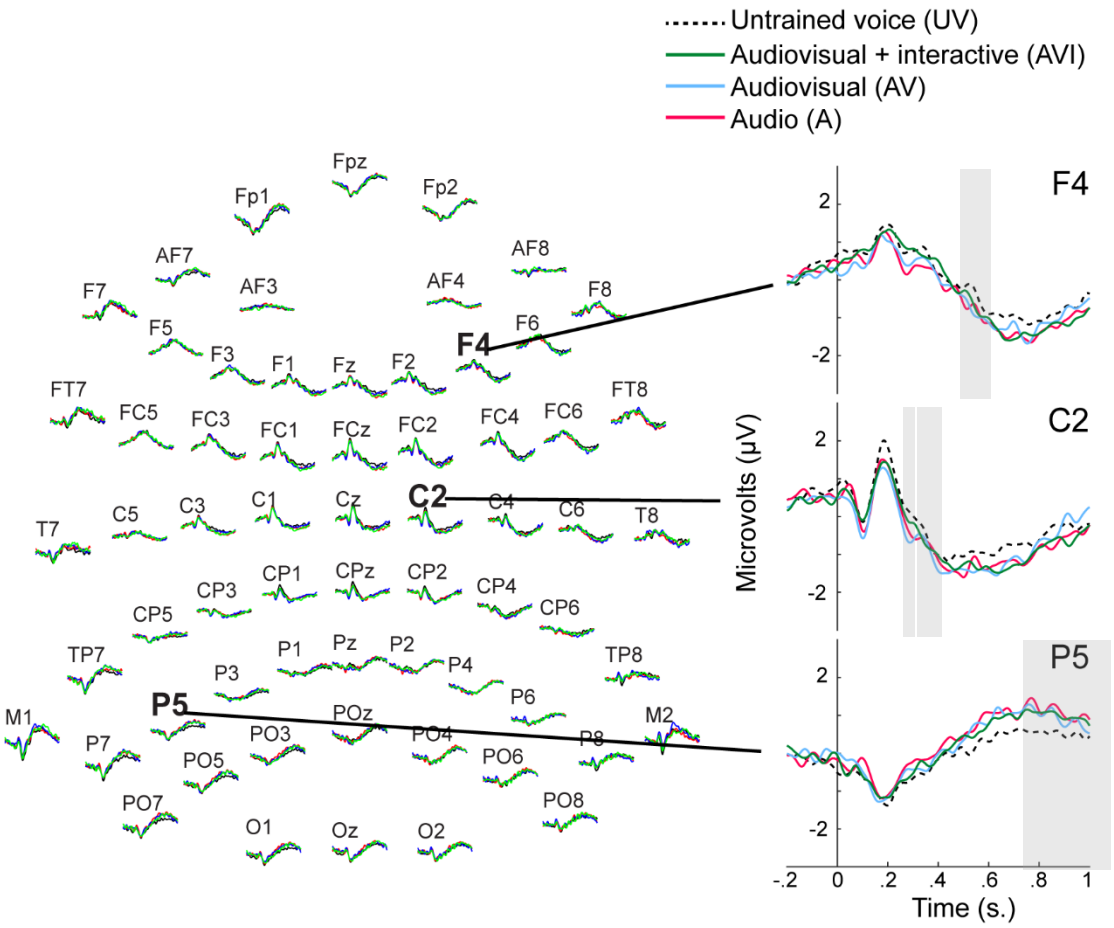

Figure 6. – ERPs across sites illustrating the effects of training conditions. Shaded areas represent the time windows of interest as shown in Figure 5.

3.2.1 Early components

3.2.1.1 N1 and P2

The Monte-Carlo analyses revealed that, between 71 and 121 post stimuli onset, differential responses to TVs relative to UV presented no significant clusters in any of the training conditions. In other words, the learning conditions had no significant effect on responses to TVs in the range of N1.

On the other hand, the analyses of the differential responses to TVs within the four 25 ms windows surrounding the P2 peak (from 139 to 239 ms post stimuli) yielded significant clusters for all three training conditions when compared individually to UV between 164 and 189 ms.

Significant clusters also appeared for the AV condition within the time window of 139 to 164 ms post stimuli onset. As can be viewed in Figure 7, all significant clusters within these time windows were located on central and parieto-central sites along the middle line.

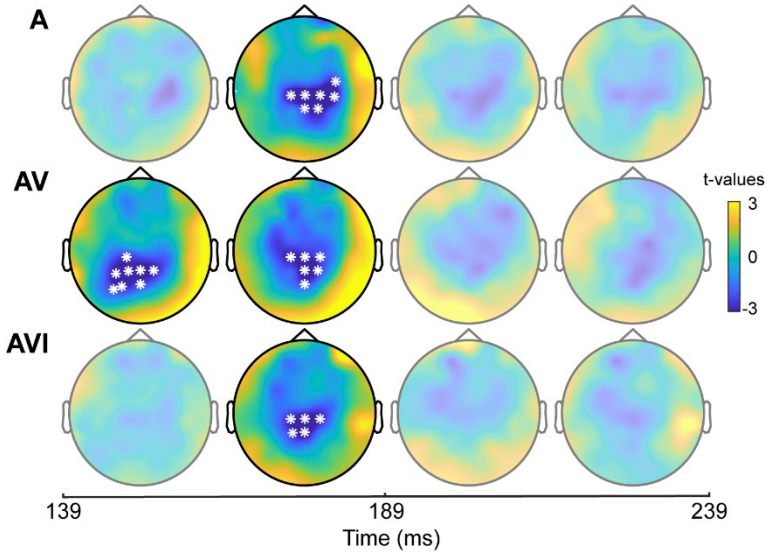

Figure 7. – Topographic representations of *t*-values obtained from cluster analyses of differential responses to TVs and UV across learning conditions for time windows of the P2 peak. Highlighted electrodes are statistically significant ($p < 0.05$).

3.2.1.2 Intermediate components

Regarding the two components with smaller peaks in GFPs at about 329 and 406 ms post onset (Figure 5), the Monte-Carlo analyses showed no significant cluster on differences in response to TVs relative to UV. Thus, in terms of early and intermediate components, the training conditions generally influenced responses appearing in the time frame of the P2.

3.2.2 LPC

As noted previously, the LPC was analyzed using 50 ms time windows in order to reduce the statistical analyses of these protracted responses. As illustrated in Figure 8, cluster analyses of the LPC responses to TVs in the three training conditions were significantly different from responses to UV across a wide time interval, although the sites of the differential activity varied across conditions. Specifically, the cluster analyses showed that the training condition A

associated with significant clusters in five of the seven time windows starting from 550 ms and extending to 900 ms. Significant clusters where mostly located on left parietal, centro-parietal, and parieto-occipital sites. The AV condition also yielded significant clusters at these sites but only between 550 and 600 ms and 750 ms 800 ms. As it seems for these LPCs, presentations of highly similar faces in learning voices (the AV training condition) did not enhance differential responses to TVs and UVs in comparison to presentations of voices alone. However, a contrasting pattern of activity in LPCs arose with the AVI training condition. In this case, significant clusters reflecting differential centro-parietal responses appeared later (starting at 650 ms post-stimuli onset) and shifted to middle and right fronto-central sites before returning to left paretal sites (at 750 ms post onset, as seen in Figure 8). This suggests that active participation in speech in learning voices of visually presented speakers bears an effect on voice processing as compared to passive conditions A and AV.

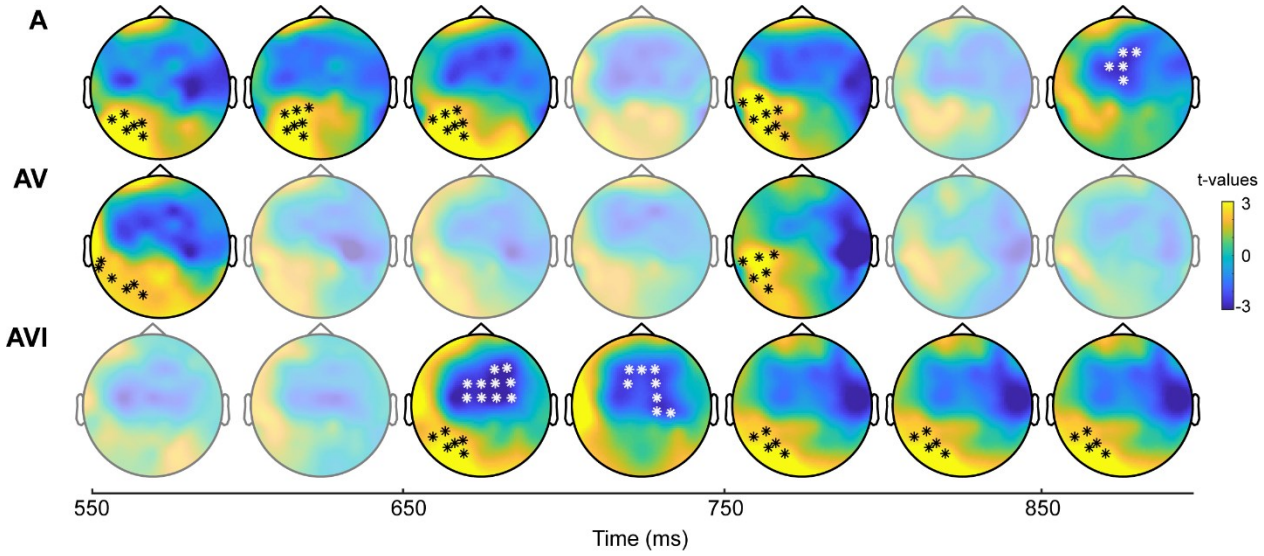

Figure 8. – Topographic representations of *t*-values obtained from the cluster analyses of differential response to TVs and UV across the learning conditions for time windows around the LPC peak. Highlighted electrodes are statistically significant ($p < 0.05$). The electrode highlight color varies for visibility purposes only.

In sum, the results revealed that the training conditions generally influenced responses to TVs in the range of P2 although this component did not show specific effects of the three learning modalities. Each of the training conditions elicited a P2 ranging between 139 and 189 ms post

stimuli onset and this differential response appeared at central and parieto-central sites. The training condition A yielded the widest ranging P2. As for late components, modality specific training effects were more apparent especially on the distribution of the LPC. Thus, while the A and AV training modalities influenced an LPC mostly at left and centro-parietal sites, the AVI training conditions corresponded to a later LPC that momentarily shifted toward middle and right centro-frontal sites.

4 Discussion

The above results can be understood in terms of a key problem facing research on neural markers of voice recognition and identification. These terms designate fundamentally distinct but interacting processes, as common experience shows. One can “recognize” a voice without knowing who is speaking, or being able “identify” the voice by associating it with a name or face. This distinction is implicit in neuroscientific models that generally posit a dual system of voice processing involving interacting perceptual and associative processes. But defining the neural markers of these processes is quite problematic. One cannot determine a priori what it is in voices that listeners recognize and associate with individuals in their personal experience. In considering this problem, the general strategy in voice research is to train listeners on given voices and provide identifying visual cues such as faces and symbols. With this strategy, however, any distinguishing mark in voices or visual cues can draw listeners’ attention and influence neural responses. For instance, presenting similar voices with dissimilar faces runs the risk of emphasizing visual information in the processing voice identity information. There is also the problem that the identification process extends to multimodal experiences that extend well beyond visual cues such as faces and symbols used in laboratory settings. The above experiment circumvented these issues through the use of narrowly controlled voice and face stimuli and by investigating effects of multimodal information on voice identification via three training modalities (audio, A; audiovisual, AV; audiovisual, AVI; audiovisual with simulated interaction). After a training phase, participants were asked to identify target speakers on multiple trials. The behavioral results showed that, for the three training conditions, trained voices (TVs) were

successfully identified, although scores showed no significant differences between the training modalities as such.

In interpreting the results obtained on ERPs, it is worth bearing in mind that the present experiment made use of highly controlled speech stimuli. In addition to general sound quality control and corrections such as amplitude normalizing and background noise filtering, all the voice stimuli reflected speakers with similar SF_0, regional accent, and similar physical appearance (Figure 3). Speech rhythm and intonation were also controlled using a prosodic guide. Moreover, disyllabic words used as training and experimental stimuli were balanced across speakers in terms of phonological content and structure so as to minimize effects of speech-related variations. It should be mentioned that the use highly similar stimuli likely contributed to the degree of difficulty of the experimental task, and this may account for the absence of differences in behavioral responses on the three training modalities.

This consistency across the stimuli is also supported by the lack of difference between any of the TVs and UV on the N1 component. As described in Näätänen & Picton (1987), the auditory N1 is an early component that is mostly sensitive to physical and temporal aspects of the stimuli. Examples of such features are sound amplitude or its frequency, the rate of repetition of the stimuli or even linguistics aspects such as the type of speech sound presented. By the absence of significant difference, the N1 observed in the present experiment confirms that all stimuli were correctly controlled and adequately similar. This further validates their rigorous elaboration.

Responses to TVs were differentiated from responses to UVs, which served as a baseline in the analyses of the ERPs. The Monte-Carlo cluster-based analyses of differential responses to TVs bore out a particular effect of training conditions A and AV. It will be recalled that A and AV only differed in terms of the presence of visual content at training. Both conditions resulted in modulations of the P2 and the LPC on similar sites, however the modulation of the LPC was wider-ranging after the training condition A. This suggests that multimodal learning involving visual face information (AV) did not facilitate voice identify encoding, but may have impeded such encoding. Such an effect concurs with research attesting to a FOE, although the effect in the present case may bear on the highly similar facial information that was provided and which not assist in identity

discrimination (Cook & Wilding, 1997a; Cook & Wilding, 2001; Heath & Moore, 2011; McAllister et al., 1993; Stevenage et al., 2011; Tomlin et al., 2016). It should also be noted that the preceding FOE occurred with dynamic facial stimuli (cf. Zäske et al., 2015).

In considering the effects of AV and AVI, it will be recalled that these two training conditions differed in that, for AV, the listeners did not see the speaker's gaze whereas, in the AVI condition, the listeners were asked to repeat the same words as the speaker while gazing at the speaker on a monitor. This mode of presentation was designed to partly simulate the effects of social interaction. As in other training conditions, the AVI condition elicited a P2 and an LPC on clusters with similar scalp distributions. Compared to the AV training, however, the AVI training resulted in a wider-ranging response in the LPC window, and also in a greater number of significant clusters that shifted momentarily toward mostly middle and right centro-frontal sites before returning to left centro-parietal sites. These results therefore provide evidence that interaction, as simulated by the effect of gaze and motor-sensory speech, enhances the ERP response at voice recall. Of course a laboratory simulation has its limits and, if anything, might underestimate the effects of actual person-to-person contact on speaker identification as reported by Hammersley & Read (1985). On the other hand, during the training cycles, some participants commented on the difficulty of the task where they had to pay attention to the words uttered to successfully repeat them while also paying attention to the vocal characteristics of the speaker. This unanticipated division of attention during training may have negatively influenced behavioral responses in the voice identification tasks at the training stage.

More importantly, a central finding of the present study is that the observed ERPs to trained voices observed narrowly conform to ERPs found in a previous study involving intimately familiar and trained-to-familiar voices (Plante-Hébert et al., 2021). Thus, both investigations involving similar stimuli have revealed responses of voice recognition and identification in the same time ranges as the P2 and LPC. It is interesting to note that in the above study, all three successfully trained voices, as validated by behavioral results, elicited a P2 that differed significantly from responses to UV. This brings new evidence confirming that the P2 is a valid marker of voice recognition. Furthermore, the present results indicate that speaker identification, or access to available semantic information on a given speaker, as measured by the LPC, was not enhanced by

the additional presentation of visual stimuli during training. This entails that a greater amount of contextual information may not necessarily facilitate voice encoding and recall. In fact, additional contextual information such as a speaker's face can, in some cases, divide the learner's attention as suggested in reports of a FOE in voice recognition tasks. On the other hand, the greater LPC response of AVI as compared to AV shows that the FOE may vary or diminish with the addition of other contextual information. In the above experiment, the added speech interaction and gaze toward a speaker during the training, even if some participants reported being distracted by the repetition of verbal forms, nonetheless created a differential LPC in response to TVs. In other words, in the same way Zäske et al. (2015) showed that the FOE decreased after an certain exposure threshold or amount of training, the present results suggest that FOE could also be reduced by the amount of contextual information that is provided.

5 Conclusion

In summary, the present study confirms that specific ERP components, more specifically the P2 and the LPC, are involved in voice recognition and identity processing. Combined with previous results, we suggest that the P2 reflects speaker recognition while the LPC represents an access to stored semantic information relating to voice identity. Such markers are especially useful not only in clarifying the neural systems of voice recognition and identification, but also in forensic applications of phonetics and neuroscience. As earwitnesses testimonies are not always reliable or robust (Clifford, 1980; Laub & al., 2013; Sherrin, 2014), methods of ERP along with standards in the elaboration of voice stimuli can be of critical importance. On such applications, further investigation should focus on the neural correlates associated with both voice memory and linguistic content recall. In other words, would ERPs be reliable in indicating if someone recalls words spoken by a given known individual?

References

References of the present paper were included in the general bibliography

Figures

Figures and figure captions were included in the text

Discussion

L'expérience présentée dans l'article 2 avait comme objectif principal d'explorer les effets de l'information contextuelle lors de l'apprentissage sur l'identification de locuteurs. Pour se faire, 18 participants ont été entrainés à identifier trois voix. Chacune de ces voix était présenté selon une modalité différente lors de l'entrainement : audio, audiovisuel ou encore audiovisuel avec interaction. Une fois l'entrainement réussi, les participants procédaient à une tâche d'identification de locuteurs impliquant les trois mêmes voix et une nouvelle voix inconnue.

Les résultats comportementaux ont confirmé que l'apprentissage avait été réussi, mais aucune des trois conditions n'a mené à des résultats significativement différents des autres.

En ce qui concerne les PÉs, les trois voix apprises ont généré des réponses significativement différentes de la voix inconnue sur deux composantes : la P2 et la LPC. Les réponses observées sur la P2 étaient similaires en termes d'amplitude et de latence pour les trois voix entrainées. Les différences entre les conditions se sont plutôt observées sur la LPC.

Dans un premier temps, la comparaison entre les LPC des conditions A et AV a indiqué une réponse électrophysiologique plus faible lorsque la voix était apprise avec l'enregistrement vidéo que lorsqu'elle était apprise uniquement à partir de matériel audio. Malgré une absence d'effets comportementaux, ces nouvelles données électrophysiologiques supportent la littérature faisant état du FOE. Les résultats comportementaux rapportés par Zäske et al. (2015), qui indiquent que le FOE tend à disparaitre au fur et à mesure que l'apprentissage d'une voix se consolide, poussent néanmoins à vouloir investiguer si cette même atténuation est observable d'un point de vue électrophysiologique dans de futures études.

La comparaison des réponses obtenues pour les conditions AV et AVI indique de son côté que l'ajout de la dimension interactive par le contact visuel et la répétition à voix haute lors de l'apprentissage a généré une réponse électrophysiologique plus grande que l'apprentissage passif. Ces données suggèrent ainsi que la mémoire épisodique est enrichie par l'interaction, ce qui facilite l'encodage des informations présentées.

Il est également intéressant de noter que les composantes sur lesquelles les réponses ont été observées, soit la P2 et la LPC, sont les mêmes que celles qui ont été observées dans la première expérience présentée dans cette thèse. Dans cette expérience, ces composantes étaient associées respectivement à la reconnaissance et à l'identification de locuteurs.

De manière générale, ces nouveaux résultats suggèrent donc que la reconnaissance, telle que reflétée par la P2, n'est pas affectée par l'information contextuelle en mémoire, mais l'identification oui. Comme l'identification fait appel à un système sémantique élargi, il n'est pas surprenant que l'influence d'autres modalités entre en jeu.

Chapitre 4 : Discussion générale

Les deux expériences réalisées dans le cadre de cette thèse portaient sur les marqueurs électrophysiologiques impliqués dans le traitement des voix. La première d'entre elles a exploré les composantes spécifiquement liées aux voix intimement familières et aux voix inconnues entendues à plusieurs reprises. Dans la deuxième expérience, ce sont les conditions favorables à la familiarisation qui ont été examinées. Pour ce faire, différentes modalités d'apprentissage ont été utilisées dans l'entrainement de participants appelés à identifier un groupe de locuteurs (A, AV et AVI).

Bien que les paradigmes expérimentaux utilisés dans les deux expériences comportaient d'importantes différences, il importe de souligner que dans les deux cas, les stimuli ont été étroitement contrôlés de manière semblable. Premièrement, les voix utilisées dans une même expérience étaient très similaires entre elles en matière de F_0mp, avec des variations qui se situaient à l'intérieur d'un semi-ton. Pour les deux expériences, les voix d'hommes sélectionnées ne comportaient pas d'accent régional perceptible ni d'idiosyncrasies marquées. Afin de restreindre les distinctions vocales liées au vieillissement, ces locuteurs étaient âgés entre 19 et 40 ans. Un guide prosodique a aussi été utilisé dans les deux cas pour assurer une intonation et un débit semblables. Les enregistrements, le montage sonore et la normalisation de l'amplitude des stimuli ont été faits dans des conditions équivalentes avec le même matériel et les mêmes logiciels. Les stimuli de l'expérience 1 ont été validés grâce à un prétest à la suite duquel l'analyse visuelle des PÉs a confirmé que les réponses observées étaient les mêmes pour chacune des voix et pour chaque énoncé. De cette manière, les réactions lors de la tâche expérimentale ne pouvaient qu'être en réponse aux manipulations des variables indépendantes. Comme une bonne partie des voix de l'expérience 1 ont été réutilisées dans l'expérience 2 et que les nouveaux enregistrements suivaient un protocole tout aussi contrôlé, le même prétest n'a pas été jugé nécessaire pour l'expérience 2.

La longueur en nombre de syllabes des énoncés utilisés n'était pas la même, principalement pour des considérations pratiques. L'expérience 2 comprenant au total plus de 500 lexèmes distincts,

l'élaboration d'un tel nombre de stimuli, mais avec quatre syllabes se serait avérée grandement complexifiée. Étant donné que la langue française comprend peu de noms communs composés d'exactement quatre syllabes, ces stimuli auraient dû inclure aussi des mots mono-, di- et trisyllabiques. Dans ce contexte, une variété de catégories grammaticales aurait aussi été nécessaire. Comme la phase d'entrainement permettait de valider que l'apprentissage des voix avec des stimuli dissyllabiques était réussi, il a été convenu qu'il était préférable d'ajuster la durée de cette phase plutôt que d'être confronté aux défis de l'élaboration de plusieurs centaines d'énoncés quadrisyllabiques.

En ce qui a trait aux participants, ils avaient tous le FQ comme langue maternelle ce qui évitait tout effet de performance en raison de la langue parlée. Sur le total des 31 participants, 17 étaient des femmes et 14 des hommes, ce qui offre une représentation quasi paritaire des sexes. L'âge des participants était lui aussi très similaire entre les deux expériences, avec un minimum de 21 ans dans les deux cas, et un maximum de 43 ans pour l'expérience 1 et de 30 pour l'expérience 2. Cette différence est attribuable aux difficultés de recrutement encourues par l'utilisation de voix intimement familières des participants dans l'expérience 1.

Finalement, les deux expériences se sont déroulées dans les mêmes locaux avec le même système d'enregistrement EEG. Seuls les logiciels de présentation des stimuli n'étaient pas les mêmes.

Les conditions générales des deux paradigmes expérimentaux étaient donc analogues tant en matière de contrôle des stimuli, de participants que de conditions expérimentales, et ce, malgré les différences de protocoles. Cette similitude permet, entre autres, de faire certaines comparaisons entre les résultats de ces deux expériences.

Voix familières

Chacune des deux expériences comportait la présentation de voix familières. Dans l'expérience 1, ces voix étaient intimement familières des participants, car elles provenaient de personnes qui leur sont proches dans la vie courante. Dans la seconde expérience, il était plutôt question de voix d'individus inconnus que les participants ont été entrainés à identifier. Dans les deux cas, l'utilisation de voix inconnues servait de ligne de base pour examiner les effets de cette

familiarité. Les résultats des deux expériences ont démontré que ces effets étaient significatifs sur l'amplitude des mêmes composantes : la P2 et la LPC. Les P2 observées étaient toutes deux situées sur des sites fronto-centraux tandis que les LPC apparaissaient sur les sites centro-pariétaux gauche et fronto-centraux droits. Les différences entre les deux paradigmes expérimentaux ne permettent pas de comparer plus en profondeurs les signaux EEG recueillis pour les voix intimement familières de l'expérience 1 et celles entrainées de l'expérience 2, mais ces similitudes pointent vers un traitement des voix familières en deux temps dans les deux cas. Dans un premier temps, la P2 reflète la reconnaissance d'un locuteur. Cette réaction brève est donc liée simplement au sentiment de connaitre un locuteur donné. C'est la LPC, plus tardive et beaucoup plus longue, qui reflèterait l'accès aux informations connues au sujet du locuteur, à son identité.

Cette réponse des PÉs en deux étapes distinctes n'est pas sans rappeler les VRU et les PINs initialement proposés dans l'IAC de Bruce et Young (1986). Dans ses versions plus contemporaines, par exemple celle de Gainotti (2014a), on propose que le sentiment de reconnaitre un individu survienne au stade des VRUs. Ces VRUs sont en quelque sorte les différents exemplaires en mémoire à propos de la voix d'un individu spécifique. Si une voix entendue correspond suffisamment aux VRUs, l'accès aux PINs est alors possible. C'est à ce stade qu'intervient l'identification de l'individu, par l'accès à un système de mémoire sémantique élargi. Les PÉs des deux expériences présentées dans ce travail ont révélé une réponse aux voix familières qui s'illustrait par une modification de l'amplitude de la P2, somme toute assez brève, et de la LPC, plus tardive et surtout plus longue. Dans le cadre de l'IAC et des modèles adaptés qui y ont fait suite, on pourrait ainsi dire que le stade des VRUs opère sur la P2 tandis que l'accès aux PINs affecte la LPC. Les données électrophysiologiques obtenues dans le cadre des deux expériences présentées sont les premières de ce type à pouvoir ainsi associer des composantes de PÉS aux stades de l'IAC.

Il importe de souligner que très peu d'études en PÉs ont utilisé des voix intimement familières comme stimuli. Dans l'expérience 1, les participants ont été recrutés en fonction de la familiarité élevée qu'ils entretenaient avec un des locuteurs présents dans les stimuli. Ce niveau de familiarité a également été attesté par les mêmes critères que ceux utilisés par Plante-Hébert et

Boucher (2015b). Le caractère inédit de cette expérience rend les résultats présentés d'autant plus uniques.

Voix inconnues

C'est principalement la première expérience qui a permis de faire des observations concluantes quant au traitement de voix inconnues. Dans cette expérience, des voix inconnues étaient présentées plus ou moins fréquemment (VE et VI). Ces deux types de voix inconnues ont démontré des réponses similaires sur les composantes liées aux voix familières, la P2 et la LPC. C'est plutôt sur la négativité immédiatement après la P2, une N250, que VE et VI ont divergé l'une de l'autre. En explorant davantage cette différence significative entre ces deux types de voix inconnues, les analyses ont révélé qu'elle était absente en début d'expérience et principalement apparente à la fin. Ces observations supplémentaires renforcent l'interprétation de ces résultats comme provenant d'un effet d'habituation à une voix inconnue.

On peut interpréter ces résultats et ceux de la reconnaissance de voix familières, observés respectivement sur le N250 et la P2, en fonction du modèle de *hérisson et du renard* de Kreiman et Sidtis (2011). Les autrices y stipulent que les voix familières sont traitées plus rapidement puisqu'elles sont perçues comme un tout *(Gestalt*) tandis que les voix inconnues sont traitées de manière plus analytique, chaque caractéristique séparément. Forcément, ce deuxième mode de traitement est plus long. Cette perspective sur le traitement différent des voix connues et inconnues expliquerait ainsi que les voix familières aient eu un impact sur la P2, tandis que les voix inconnues répétées plus fréquemment ont affecté une composante plus tardive, la N250. D'ailleurs, les autrices de ces modèles détaillent les spécialisations hémisphériques en rapport avec le traitement des voix et suggèrent que les voix familières sont traitées à droite tandis que les voix inconnues le sont plutôt à gauche. Bien que les PÉs ne soient pas précis sur le plan spatial, il est tout de même intéressant de souligner que dans l'expérience 1, la P2 observée était sur des sites fronto-centraux droits, tandis que la N250 était fronto-centrale gauche.

Ces mêmes données sont aussi interprétables selon le modèle plus contemporain de Maguinness et al. (2018). Ce modèle a été construit à la fois à partir de données neuroanatomiques et à partir du modèle des prototypes de Lavner et al. (2001). Dans ce modèle de Maguinness et al., après

l'analyse des caractéristiques vocales perçues, la distance entre ces caractéristiques et celles déjà stockées en mémoire est calculée. Tel que décrit au Chapitre 1, si cette distance est suffisamment petite, la voix est considérée comme étant connue et passe ensuite au stade d'identification tandis que si la distance est grande, la voix est considérée comme étant inconnue et est utilisée pour créer une nouvelle représentation de référence. Ainsi, cet aiguillage des voix en fonction de leur distance avec les prototypes en mémoire se verrait illustré par la P2 lorsque les voix sont classées familières et par la N250 lorsqu'elles sont classées inconnues et utilisées pour créer de nouvelles représentations.

Un point sur lequel les données observées peuvent être surprenantes est que la VE dans l'expérience 1 est, somme toute, très semblable à la condition A de l'expérience 2. Dans le premier cas, il est question d'une voix inconnue présentée fréquemment tout au long de l'expérience. Dans le second cas, c'est une voix explicitement apprise uniquement à l'aide de mots présentés auditivement. L'information encodée pour ces deux types de voix est donc pratiquement la même : de courts échantillons vocaux. Pourtant, dans l'expérience 1, les PÉs de VE se comportent comme ceux de voix inconnues sur les P2 et LPC. Parallèlement, dans l'expérience 2, la condition A a généré une réponse spécifique aux voix connues sur la P2 et la LPC. En somme, les PÉs de la voix fréquente de l'expérience 1 correspondent à ceux d'une voix inconnue, et les PÉs de la condition A de l'expérience 2 réagissent plutôt comme ceux d'une voix connue. Pourtant, les deux types de voix sont associés au même type d'information.

La réponse à cette différence à première vue étonnante provient, selon toute vraisemblance, des directives et des contextes propres à chaque expérience. Dans l'expérience 1, aucune mention de VE n'était explicitement faite. C'était tout simplement une voix qui revenait plus fréquemment que les autres dans les blocs expérimentaux. Dans l'expérience 2 par contre, A faisait l'objet d'un apprentissage explicite avec d'autres voix qui étaient, elles, présentées conjointement avec plus d'information contextuelle. Ainsi, la création d'une représentation du locuteur en A était grandement favorisée par rapport à la VE de l'expérience 1. La différence entre VE et VI sur la N250 dans l'expérience 1 reflèterait possiblement le processus d'encodage d'une nouvelle représentation. Pour la condition A, cette représentation était quant à elle déjà créée.

Modalités d'encodage

Dans l'expérience 2, les participants ont appris les voix de trois locuteurs selon des modalités différentes. Dans un premier temps, les PÉs ont généré une réponse spécifique par rapport à la VI sur la P2 qui ne variait pas d'une condition à l'autre. Cela suggère que la reconnaissance d'individus telle qu'illustrée par la P2 n'est pas modulée par l'information contextuelle à l'apprentissage. Une voix familière génère donc une réponse différente d'une voix inconnue sur la P2, peu importe qu'un visage y soit associé ou encore qu'on ait interagi ou non avec la personne.

C'est plutôt sur la LPC que des effets de modalité d'apprentissage ont été observés. Dans un premier temps, la comparaison entre A et AV a révélé une réponse plus faible dans la condition AV. Cette atténuation serait principalement due au FOE fréquemment rapporté dans la littérature. Il est intéressant de noter ici que, bien que l'apprentissage ait été réussi avec des taux supérieurs à 75 % d'identifications correctes, cet effet s'est maintenu dans les réponses électrophysiologiques. Zäske et al. (2015) rapportaient quant à eux une disparition du FOE au fur et à mesure que l'apprentissage d'une voix se consolidait. Il serait intéressant de vérifier selon quelles conditions et à partir de quand, précisément, un tel effet tend à s'estomper.

Toujours par rapport à l'analyse de la LPC, les résultats obtenus à l'expérience 2 indiquent justement que le FOE observé dans la comparaison A-AV n'était plus présent dans la condition AVI. C'est en comparant les conditions AV et AVI qu'une réponse beaucoup plus vaste pour AVI a été constatée. Sans être identique à la réponse présente lors de la condition d'entrainement A en tous points, cette LPC plus prononcée lors de la condition AVI indique que le FOE tend à être minimisé par l'ajout d'information contextuelle, ici représentée par l'interaction.

Implications en phonétique légale

Tel que rappelé tout au long de ce travail, les applications au domaine de la phonétique légale ont été une considération importante dans l'élaboration et l'interprétation des deux expériences. Les études précédentes revues au Chapitre 1, tout particulièrement celle de Plante-Hébert et Boucher (2014), indiquent qu'un niveau de familiarité élevé permet d'obtenir des taux

d'identification très fiables. La principale limite de ces conclusions dans leur application au domaine légal est que le témoignage d'un identificateur intimement familier avec un suspect repose entièrement sur la crédibilité et la fiabilité du témoin. En ce sens, le premier objectif était de vérifier si des marqueurs électrophysiologiques permettent de valider l'identification d'un locuteur intimement connu. Une telle approche permettrait d'attester l'identification sans même avoir recours aux réponses intentionnellement fournies par le témoin.

À cet effet, les résultats des deux expériences ont démontré un effet persistant de reconnaissance sur la composante P2. Cet effet s'est manifesté pour les voix intimement familières de l'expérience 1 aussi bien que pour les voix entrainées dans l'expérience 2. Ces deux types de voix ont aussi engendré une réponse spécifique sur la LPC. Cette réponse plus tardive et de grande envergure illustrait l'identification des locuteurs familiers par l'accès aux données sémantiques en mémoire.

Dans la perspective d'applications au domaine de la phonétique légale, ces réponses concluantes et spécifiques aux voix connues sont prometteuses. Non seulement les données observées dans les deux expériences portaient sur les mêmes composantes à des latences très similaires, mais elles étaient également observées dans les mêmes régions du scalp. En vue de telles applications, il faut mentionner à nouveau que les stimuli étaient rigoureusement contrôlés, rendant ainsi les tâches expérimentales plus ardues que dans la plupart des études en général. Néanmoins, les données électrophysiologiques se sont avérées statistiquement significatives en plus d'être constantes d'une expérience à l'autre.

La fiabilité de ces observations est d'autant plus renforcée par la présence d'une différence significative sur la composante N250 dans l'expérience 1. Cette différence était entre les voix inconnues fréquemment jouées et celles très rarement présentées. Ces deux catégories de voix inconnues partageaient des PÉs semblables dans l'ensemble du signal, sauf au niveau de cette composante. En vue d'application au domaine légal, ce constat indique que la reconnaissance d'une simple voix déjà entendue ne pourrait pas être confondue avec une voix familière dans l'examen des PÉs. Ces observations impliquent aussi qu'il serait possible d'attester qu'une voix est bel est bien reconnues sans pour autant qu'elle provienne d'un locuteur intimement familier.

Cette distinction est majeure dans des situations ou une victime n'est pas familière de son agresseur et n'a pas été en mesure de voir son visage lors de l'agression. Il en va de même pour des situations de témoins d'actes criminels ayant seulement entendu la vois du suspect.

En raison des différences de protocoles entre les deux expériences, il n'a pas été possible de comparer directement les signaux des VE de l'expérience 1 et des VE de l'expérience 2. Il serait toutefois intéressant d'examiner précisément où se situent les différences entre les réponses de voix explicitement et implicitement apprises. De la même manière, bien que les voix familières apprises ou intimement connues génèrent des réponses spécifiques sur les composantes P2 et LPC, un examen détaillé des réponses à ces deux types de voix au sein d'une même expérience aiderait à déterminer si les PÉs permettent de distinguer de manière fiable le traitement d'une voix volontairement apprise et celui d'une voix intimement connue. Dans le cadre d'applications légales, ces distinctions sauraient renforcer la validité de témoignages auditifs-perceptuels.

Il faut cependant noter qu'une certaine résistance à l'utilisation d'une technologie telle que l'analyse de PÉs pourrait être rencontrée. Bien que cette technique soit considérée en science comme non invasive, elle peut représenter une forme d'intrusion dans la « psychée » pour qui n'en connait pas le fonctionnement. Ainsi, dans son utilisation en contexte légal, certains pourraient être tentés de percevoir une forme d'invasion de l'intimité qui ne saurait socialement être acceptable. Un travail de démystification de la technologie utilisée s'impose donc non seulement auprès des professionnels des milieux impliqués, mais aussi auprès de la population en général.

Il n'en demeure pas moins que les résultats présentés dans le présent travail de thèse encouragent une collaboration entre les domaines de la phonétique légale et des neurosciences. Ce décloisonnement des domaines de recherche appliqués est prometteur en ce qu'il ouvre la porte à de nouvelles approches inédites et au développement de pratiques novatrices et issues de la recherche scientifique.

Limites et travaux futurs

Opérationnalisation

Il importe également de souligner que les études présentées dans ce travail de thèse, bien que conçue dans un cadre appliqué, demeurent expérimentales. En ce sens, un important travail d'opérationnalisation demeure nécessaire avant d'envisager mettre en application une technique de validation d'identification du locuteur dans un contexte légal. Cette opérationnalisation ne saurait être complète sans l'intervention et la collaboration d'acteurs d'importance en provenance mes milieux juridiques, policiers et technologiques. Dans le cadre d'un tel processus, il serait important de considérer, par exemple, les coûts humains, matériels et financiers que pourrait encourir l'utilisation des PÉs dans le milieu légal afin de les minimiser et de rentabiliser l'approche. L'équipement technique pourrait ainsi être adapté pour réduire les frais et faciliter la procédure, tout en s'assurant que les observations requises sont toujours fiables.

Variabilité inter-sujets

Un aspect majeur n'ayant pas été abordé dans le cadre du présent travail de thèse est la variation inter-sujets, tant dans la capacité à reconnaitre et identifier des locuteurs que dans les réponses électrophysiologiques.

Il est bien connu que les PÉs sont très sensibles et diffèrent d'un sujet à l'autre (Luck et Kappenman, 2011; Woodman, 2010). Il serait donc pertinent, voir nécessaire dans le cadre de l'opérationnalisation décrite ci-dessus, de prendre en considération cette variable. Il s'agit, en l'occurrence, de vérifier que les composantes impliquées réagissent de manière fiable, malgré que leur latence ou encore leur amplitude puissent varier entre les individus. Ainsi, on peut espérer établir que la composante P2 et la LPC sont systématiquement modulés par l'écoute de voix intimement familières, et ce, même si elles apparaissent plus hâtivement ou tardivement en fonction du sujet.

Interaction expérimentale

L'expérience 2 faisait appel à une condition d'apprentissage de la voix dite interactive (AVI). Le choix méthodologique d'utiliser des enregistrements dans lesquels le regard du locuteur était

dirigé vers la caméra (donc vers l'apprenant) et suite auxquels l'apprenant devait répéter le mot entendu avait son lot d'avantages, mais aussi de limitations. En termes d'avantages, cette approche a permis de contrôler avec précision l'information acoustique et visuelle présentée aux différents participants. Contrairement à une interaction en personne, où il aurait été impossible d'anticiper et de mesurer les variations d'un participant à l'autre, l'utilisation d'enregistrements proposait une information prédéfinie et constante.

Il importe par contre de souligner que cette version adaptée de ce qu'est une interaction était limitée. Bien que certaines études soulignent que l'effet du regard partagé se maintient même dans un contexte où les deux individus ne sont pas physiquement en présence l'un de l'autre, il semble raisonnable de penser que son effet est néanmoins atténué s'il est question d'enregistrements. Les participants à l'expérience 2 étaient d'ailleurs avisés qu'il était question d'enregistrements pour simuler l'interaction.

Il en va de même pour l'effet de répétition des mots entendus. Cette composante de la condition AVI servait à rendre compte de la composante motrice présente lors d'interaction. Lorsque deux individus discutent, les tours de parole alternent et celui qui écoute devient locuteurs et vice-versa. La répétition après chaque mot entendu servait donc à reproduire ce changement de tour de parole afin d'enrichir l'épisode mnésique vécu par les participants. Encore une fois, il est raisonnable de penser que la répétition d'un mot n'est pas égale à la formulation d'une réponse appropriée à un énoncé complexe.

Malgré ces inconvénients, la condition AVI a tout de même généré une réponse de PÉ bien plus vaste que la condition AV sur la LPC. Il serait donc intéressant, dans de futures études, de se pencher davantage sur les effets de l'interaction sur l'encodage d'information en utilisant différents types de paradigmes expérimentaux d'apprentissage, pouvant aller jusqu'à l'utilisation d'interactions réelles.

Familiarité et mémoire verbale

Comme évoqué dans la section méthodologique du Chapitre 3, la deuxième expérience comportait une dernière partie dont les données n'ont pas été analysées à ce jour. Cette partie expérimentale avait pour objectif d'observer les effets des modalités d'apprentissage (A, AV et

AVI) sur le rappel de mots entendus. Pour y parvenir, les mêmes 240 stimuli qu'utilisés dans la première partie ont été présentés à nouveau ainsi que 240 nouveaux mots enregistrés par les mêmes locuteurs. Les participants, qui n'avaient pas été avisés de la tâche de cette dernière partie, devaient indiquer si chaque mot entendu avait été présenté ou non dans la partie 1. Les participants n'avaient pas besoin de tenir compte de l'identité du locuteur pour cette partie. Comme pour la première partie, les données comportementales et les enregistrements EEG ont été collectés.

Une fois analysées, les données recueillies permettront d'investiguer les effets de l'information contextuelle lors de l'apprentissage sur le rappel de contenu linguistique plutôt que sur le rappel de l'identité. Présentement, le modèle de Belin et al. (2004) et sa version la plus à jour proposée par Young et al. (2020) suggèrent que l'information vocale soit traitée en trois modules indépendants correspondants à l'information linguistique, l'information affective et l'information sur l'identité. Cette partie supplémentaire de l'expérience 2 permettra non seulement de confirmer que l'information contextuelle influence aussi le rappel de ce qui est dit, mais elle apportera également des données quant à l'influence qu'a la familiarité du locuteur sur le rappel de ses propos. Bien entendu, ces données seront-elles aussi interprétées dans le contexte d'éventuelles applications à la phonétique légale.

Chapitre 5 : Conclusion

La revue de la littérature ainsi que les résultats présentés dans ce travail de thèse confirment que les PÉs pourraient s'avérer un outil efficace pour les applications à phonétique légale, principalement en ce qui a trait à l'identification d'individus par des témoins auditifs.

Tel que discuté au Chapitre 1, les PÉs représentent la technique neurophysiologique la plus adaptée à des applications légales en raison de son accessibilité et de sa mobilité. Les méthodes utilisées dans les expériences 1 et 2 demeurent néanmoins scientifiques et nécessitent assurément une opérationnalisation avant de pouvoir être utilisées. Dans ce sens, la conclusion de cette thèse encourage la collaboration entre les différents milieux de pratique, que ce soit les services de renseignement, les tribunaux ou encore les juristes eux-mêmes, et le domaine scientifique. L'identification d'individus uniquement par leur voix est moins fréquente que par le visage et par conséquent moins connue. Il n'en reste pas moins que dans certaines situations, lorsque le visage est masqué par exemple, ce mode d'identification représente l'unique avenue possible. Plutôt que d'admettre des témoignages dans des conditions souvent inadéquates qui sèment le doute sur leur validité, ne vaudrait-il pas mieux développer et faire appel à des technologies certes avant-gardistes, mais scientifiquement éprouvées ?

Références bibliographiques

Amino, K. et Arai, T. (2009). Speaker-dependent characteristics of the nasals. *Forensic science international, 185*, 21-28. doi: 10.1016/j.forsciint.2008.11.018

Amino, K., Sugawara, T. et Arai, T. (2005). *The correspondences between the perception of the speaker individualities contained in speech sounds and their acoustic properties.* Communication présentée Interspeech. Repéré à https://www.isca-speech.org/archive/archive_papers/interspeech_2005/i05_2025.pdf

Amino, K., Sugawara, T. et Arai, T. (2006). Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties. *Acoustical science and technology, 27*, 233-235. doi: 10.1250/ast.27.233

Armstrong, H. A. et McKelvie, S. J. (1996). Effect of face context on recognition memory for voices. *Journal of General Psychology, 123*, 259-270. doi: 10.1080/00221309.1996.9921278

Atkinson, N. (2015). *Variable factors affecting voice identification in forensic contexts.* (University of York). Repéré à http://etheses.whiterose.ac.uk/id/eprint/13013

Barsalou, L. W. (2016). Situated conceptualization: Theory and applications. Dans Y. Coello & M. H. Fischer (dir.), *Perceptual and Emotional Embodiment: Foundations of Embodied Cognition* (Vol. 1). London: Routeledge. doi:10.4324/9781315751979

Bartholomeus, B. (1973). Voice identification by nursery school children. *Canadian Journal of Psychology/Revue canadienne de psychologie, 27*, 464. doi: 10.1037/h0082498

Barton, J. J. et Corrow, S. L. (2016). Recognizing and identifying people: A neuropsychological review. *Cortex, 75*, 132-150. doi: 10.1016/j.cortex.2015.11.023

Beauchemin, M., De Beaumont, L., Vannasing, P., Turcotte, A., Arcand, C., Belin, P. et Lassonde, M. (2006). Electrophysiological markers of voice familiarity. *European Journal of Neuroscience, 23*, 3081-3086. doi: 10.1111/j.1460-9568.2006.04856.x

Beauchemin, M., González-Frankenberger, B., Tremblay, J., Vannasing, P., Martínez-Montes, E., Belin, P., . . . Wallois, F. (2011). Mother and stranger: An electrophysiological study of voice processing in newborns. *Cerebral Cortex*, bhq242. doi: 10.1093/cercor/bhq242

Belin, P., Bestelmeyer, P. E. G., Latinus, M. et Watson, R. (2011). Undestanding voice perception. *British Journal of Psychology, 102*, 711-725. doi: 10.1111/j.2044-8295.2011.02041.x

Belin, P., Fecteau, S. et Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8*. doi: 10.1016/j.tics.2004.01.008

Belin, P. et Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport, 14*, 2105-2109. doi: 10.1097/00001756-200311140-00019

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P. et Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature 403*. doi: 10.1038/35002078

Bentin, S. et Deouell, L. Y. (2000). Structural encoding and identification in face processing: ERP evidence for seperate mechanisms. *Cognitive Neuropsychology, 17*, 35-54. doi: 10.1080/026432900380472

Birkett, P. B., Hunter, M. D., Parks, R. W., Farrow, T. F., Lowe, H., Wilkinson, I. D. et Woodruff, P. W. (2007). Voice familiarity engages auditory cortex. *Neuroreport, 18*, 1375-1378. doi: 10.1097/WNR.0b013e3282aa43a3

Blank, H., Anwander, A. et von Kriegstein, K. (2011). Direct structural connections between voice- and face-recognition areas. *Journal of Neuroscience, 31*, 12906-12915. doi: 10.1523/JNEUROSCI.2091-11.2011

Blank, H., Kiebel, S. J. et von Kriegstein, K. (2015). How the human brain exchanges information across sensory modalities to recognize other people. *Human Brain Mapping, 36*, 324-339. doi: 10.1002/hbm.22631

Blank, H., Wieland, N. et von Kriegstein, K. (2014). Person recognition and the brain: Merging evidence from patients and healthy individuals. *Neuroscience & Biobehavioral Reviews, 47*, 717-734. doi: 10.1016/j.neubiorev.2014.10.022

Blatchford, H. et Foulkes, P. (2007). Idenfication of voices in shouting. *International Journal of Speech, Language and the Law, 13*, 241-254. doi: 10.1558/ijsll.2006.13.2.241

Braun, A. (2016). *The speaker identification ability of blind and sighted listeners: An empirical investigation*. United Kingdom: Springer. doi:10.1007/978-3-658-15198-0

Bricker, P. D. et Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America, 40*, 1441-1449. doi: 10.1121/1.1910246

Broeders, A. P. A. et van Amelsvoort, A. G. (1999). *Lineup construction for forensic earwitness identification: A practical approach.* Communication présentée 14th International Congress of Phonetic Sciences, San Francisco, CA. Repéré à https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_1373.pdf

Bruce, V. et Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305-327. doi: 10.1111/j.2044-8295.1986.tb02199.x

Brungart, D. S., Scott, K. R. et Simpson, B. D. (2001). *The influence of vocal effort on human speaker identification.* Communication présentée Seventh European Conference on Speech Communication and Technology. Repéré à https://www.isca-speech.org/archive/archive_papers/eurospeech_2001/e01_0747.pdf

Buchtel, H. A. et Stewart, J. D. (1989). Auditory agnosia: Apperceptive or associative disorder? *Brain and language, 37*, 12-25. doi: 10.1016/0093-934X(89)90098-9

Bülthoff, I. et Newell, F. N. (2015). Distinctive voices enhance the visual recognition of unfamiliar faces. *Cognition, 137*, 9-21. doi: 10.1016/j.cognition.2014.12.006

Burton, M. A., Bruce, V. et Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology, 81*, 361-380. doi: 10.1111/j.2044-8295.1990.tb02367.x

Caharel, S., Poiroux, S. et Bernard, C. (2002). ERPs associated with familiarity and degree of familiarity during face recognition. *International Journal of Neuroscience, 112*, 1531-1544. doi: 10.1080/00207450290158368

Calderwood, L., McKay, D. et Stevenage, S. (2019). Children's identification of unfamiliar voices on both target-present and target-absent lineups. *Psychology, Crime and Law*, 1-15. doi: 10.1080/1068316X.2019.1597090

Capilla, A., Belin, P. et Gross, J. (2012). The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. *Cerebral Cortex, 23*, 1388-1395. doi: 10.1093/cercor/bhs119

Charest, I., Pernet, C. R., Rousselet, G. A., Quinones, I., Latinus, M., Fillion-Bilodeau, S., . . . Belin, P. (2009). Electrophysiological evidence for an early processing of human voices. *BMC Neuroscience, 10*. doi: 10.1186/1471-2202-10-127

Chlebowski, C. (2011). Wechsler Memory Scale All Versions. Dans J. S. Kreutzer, J. DeLuca & B. Caplan (dir.), *Encyclopedia of Clinical Neuropsychology* (p. 2688-2690). New York, NY: Springer New York. doi:10.1007/978-0-387-79948-3_1163

Clifford, B. R. (1980). Voice identification by human listeners: On earwitness reliability. *Law and Human Behavior, 4*, 373-395. doi: 10.1007/BF01040628

Conde, T., Gonçalves, Ó. F. et Pinheiro, A. P. (2015). Paying attention to my voice or yours: An ERP study with words. *Biological psychology, 111*, 40-52. doi: 10.1016/j.biopsycho.2015.07.014

Conde, T., Gonçalves, Ó. F. et Pinheiro, A. P. (2016). The effects of stimulus complexity on the preattentive processing of self-generated and nonself voices: An ERP study. *Cognitive, Affective, & Behavioral Neuroscience, 16*, 106-123. doi: 10.3758/s13415-015-0376-1

Conty, L., N'Diaye, K., Tijus, C. et George, N. (2007). When eye creates the contact! ERP evidence for early dissociation between direct and averted gaze motion processing. *Neuropsychologia, 45*, 3024-3037. doi: 10.1016/j.neuropsychologia.2007.05.017

Conty, L., Russo, M., Loehr, V., Hugueville, L., Barbu, S., Huguet, P., . . . George, N. (2010). The mere perception of eye contact increases arousal during a word-spelling task. *Social neuroscience, 5*, 171-186. doi: 10.1080/17470910903227507

Conty, L., Tijus, C., Hugueville, L., Coelho, E. et George, N. (2006). Searching for asymmetries in the detection of gaze contact versus averted gaze under different head views: A behavioural study. *Spatial vision, 19*, 529-545. doi: 10.1163/156856806779194026

Cook, S. et Wilding, J. (1997a). Earwitness testimony 2. Voices, faces and context. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory*

and Cognition, 11, 527-541. doi: 10.1002/(SICI)1099-0720(199712)11:6<527::AID-ACP483>3.0.CO;2-B

Cook, S. et Wilding, J. (1997b). Earwitness testimony: Never mind the variety, hear the length. *Applied Cognitive Psychology, 11*, 95-111. doi: 10.1002/(SICI)1099-0720(199704)11:2<95::AID-ACP429>3.0.CO;2-O

Cook, S. et Wilding, J. (2001). Earwitness testimony: Effects of exposure and attention on the face overshadowing effect. *British Journal of Psychology, 92*, 617-629. doi: 10.1348/000712601162374

Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory & Cognition, 28*, 923-938. doi: 10.3758/BF03209340

Curran, T., Tepe, K. L. et Piatt, C. (2006). Event-related potential explorations of dual processes in recognition memory. Dans H. D. Zimmer, A. Mecklinger & U. Lindenberger (dir.), *Binding in human memory: A neurocognitive approach* (p. 467-492). Oxford: Oxford Univeristy Press. doi:10.1093/acprof:oso/9780198529675.003.0018

de Jong-Lendle, G., Nolan, F., McDougall, K. et Hudson, T. (2015). *Voice lineups: A practical guide.* Dans The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow. ISBN 978-0-85261-941-4. Paper number 1041. 1-9. Repéré à http://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0598.pdf

De Lucia, M., Clarke, S. et Murray, M. M. (2010). A temporal hierarchy for conspecific vocalization discrimination in humans. *Journal of Neuroscience, 30*, 11210-11221. doi: 10.1523/JNEUROSCI.2239-10.2010

De Renzi, E., Faglioni, P., Grossi, D. et Nichelli, P. (1991). Apperceptive and associative forms of prosopagnosia. *Cortex, 27*, 213-221. doi: 10.1016/S0010-9452(13)80125-6

DeCasper, A. J. et Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science*, 1174-1176. doi: 10.1126/science.7375928

deRegnier, R.-A., Nelson, C. A., Thomas, K. M., Wewerka, S. et Georgieff, M. K. (2000). Neurophysiologic evaluation of auditory recognition memory in healthy newborn infants

and infants of diabetic mothers. *Journal of pediatrics, 137*, 777-784. doi: 10.1067/mpd.2000.109149

Didic, M., Aglieri, V., Tramoni-Nègre, E., Ronat, L., Le Ber, I., Ceccaldi, M., . . . Felician, O. (2020). Progressive phonagnosia in a telephone operator carrying a C9orf72 expansion. *Cortex, 132*, 92-98. doi: 10.1016/j.cortex.2020.05.022

Doromal, M. (2016). *Bilingual and whispered speaker identification within a social network.* (University of York, United Kingdom).

Eladd, E., Segev, S. et Tobin, Y. (1998). Long-term working memory in voice identification. *Psychology, Crime and Law, 4*, 73-88. doi: 10.1080/10683169808401750

Ellis, H. D., Jones, D. M. et Mosdell, n. (1997). Intra- and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology, 88*, 143-156. doi: 10.1111/j.2044-8295.1997.tb02625.x

Farrant, B. M. et Zubrick, S. R. (2012). Early vocabulary development: The importance of joint attention and parent-child book reading. *First Language, 32*, 343-364. doi: 10.1177/0142723711422626

Föcker, J., Best, A., Hölig, C. et Röder, B. (2012). The superiority in voice processing of the blind arises from neural plasticity at sensory processing stages. *Neuropsychologia, 50*, 2056-2067. doi: 10.1016/j.neuropsychologia.2012.05.006

Föcker, J., Hölig, C., Best, A. et Röder, B. (2011). Crossmodal interaction of facial and vocal person identity information: An event-related potential study. *Brain research, 1385*, 229-245. doi: 10.1016/j.brainres.2011.02.021

Foulkes, P. et Barron, A. (2000). Telephone speaker recognition amongst members of a close social network. *Forensic Linguistics, 7*, 180-198. doi: 10.1558/sll.2000.7.2.180

Gainotti, G. (2011). What the study of voice recognition in normal subjects and brain-damaged patients tells us about models of familiar people recognition. *Neuropsychologia, 49*, 2273-2282. doi: 10.1016/j.neuropsychologia.2011.04.027

Gainotti, G. (2013). Laterality effects in normal subjects' recognition of familiar faces, voices and names. Perceptual and representational components. *Neuropsychologia, 51*, 1151-1160. doi: 10.1016/j.neuropsychologia.2013.03.009

Gainotti, G. (2014a). Cognitive models of familiar people recognition and hemispheric asymmetries. *Frontiers in Bioscience (Elite Edition), 6*, 148-158. doi: 10.5334/pb.at

Gainotti, G. (2014b). The neuropsychology of familiar person recognition from face and voice. *Psychologica Belgica, 54*. doi: 10.5334/pb.at

Gainotti, G. (2015). Implications of recent findings for current cognitive models of familiar people recognition. *Neuropsychologia, 77*, 279-287. doi: 10.1016/j.neuropsychologia.2015.09.002

Gainotti, G. (2018). How can familiar voice recognition be intact if unfamiliar voice discrimination is impaired? An introduction to this special section on familiar voice recognition. *Neuropsychologia, 116*, 151-153. doi: 10.1016/j.neuropsychologia.2018.04.003

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J. R., . . . Duchaine, B. (2009). Developmental phonagnosia: A selective deficit of vocal identity recognition. *Neuropsychologia, 47*, 123-131. doi: 10.1016/j.neuropsychologia.2008.08.003

Gilbert, A. C., Boucher, V. J. et Jemel, B. (2014). Perceptual chunking and its effect on memory in speech processing: ERP and behavioral evidence. *Frontiers in psychology, 5*, 220. doi: 10.3389/fpsyg.2014.00220

Glenn, J. W. et Kleiner, N. (1968). Speaker identification based on nasal phonation. *Journal of the Acoustical Society of America, 43*, 368-372. doi: 10.1121/1.1910788

Gobbini, M. I. et Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia, 45*, 32-41. doi: 10.1016/j.neuropsychologia.2006.04.015

Goggin, J. P., Thompson, C. P., Strube, G. et Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition, 19*, 448-458. doi: 10.3758/BF03199567

Gonzalez, I. Q., Bobes Leon, M. A., Belin, P., Martinez-Quintana, Y., Galan Garcia, L. et Sanchez Castillo, M. (2011). Person identification through face and voices: An ERP study. *Brain research, 1407*, 13-26. doi: 10.1016/j.brainres.2011.03.029

Graux, J., Gomot, M., Roux, S., Bonnet-Brilhault, F. et Bruneau, N. (2015). Is my voice just a familiar voice? An electrophysiological study. *Social cognitive and affective neuroscience, 10*, 101-105. doi: 10.1093/scan/nsu031

Graux, J., Gomot, M., Roux, S., Bonnet-Brilhault, F., Camus, V. et Bruneau, N. (2013). My voice or yours? An electrophysiological study. *Brain Topography, 26*, 72-82. doi: 10.1007/s10548-012-0233-2

Gunji, A., Koyama, S., Ishii, R., Levy, D., Okamoto, H., Kakigi, R. et Pantev, C. (2003). Magnetoencephalographic study of the cortical activity elicited by human voice. *Neuroscience Letters, 348*, 13-16. doi: 10.1016/S0304-3940(03)00640-2

Hailstone, J. C., Crutch, S. J., Vestergaard, M. D., Patterson, R. D. et Warren, J. D. (2010). Progressive associative phonagnosia: A neuropsychological analysis. *Neuropsychologia, 48*, 1104-1114. doi: 10.1016/j.neuropsychologia.2009.12.011

Hammersley, R. et Read, J. D. (1985). The effect of participation in a conversation on recognition and identification of the speakers' voices. *Law and Human Behavior, 9*, 71. doi: 10.1007/BF01044290

Haxby, J. V., Hoffman, E. A. et Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*, 223-233. doi: 10.1016/S1364-6613(00)01482-0

Heath, A. J. et Moore, K. (2011). Earwitness memory: Effects of facial concealment on the face overshadowing effect. *International Journal of Advanced Science and Technology, 33*, 131-140.

Helminen, T. M., Pasanen, T. P. et Hietanen, J. K. (2016). Learning under your gaze: The mediating role of affective arousal between perceived direct gaze and memory performance. *Psychological Research, 80*, 159-171. doi: 10.1007/s00426-015-0649-x

Hepper, P. G., Scott, D. et Shahidullah, S. (1993). Newborn and fetal response to maternal voice. *Journal of Reproductive and Infant Psychology, 11*, 147-153. doi: 10.1080/02646839308403210

Hirotani, M., Stets, M., Striano, T. et Friederici, A. D. (2009). Joint attention helps infants learn new words: Event-related potential evidence. *Neuroreport, 20*, 600-605. doi: 10.1097/WNR.0b013e32832a0a7c

Holeckova, I., Fischer, C., Giard, M.-H., Delpuech, C. et Morlet, D. (2006). Brain responses to a subject's own name uttered by a familiar voice. *Brain research, 1082*, 142-152. doi: 10.1016/j.brainres.2006.01.089

Hollien, H. (1990). *The acoustics of crime*. New York: Springer. doi:10.1007/978-1-4899-0673-1

Hollien, H., Huntley Bahr, R. et Harnsberger, J. D. (2014). Issues in forensic voice. *Journal of Voice, 28*, 170-184. doi: 10.1016/j.jvoice.2013.06.011

Hollien, H., Huntley Bahr, R., Künzel, H. J. et Hollien, P. (1995). Criteria for earwitness lineups. *Forensic Linguistics, 2*, 143-153. doi: 10.1558/ijsll.v2i2.143

Hollien, H., Huntley, R., Kunzel, H. et Hollien, P. A. (2013). Criteria for earwitness lineups. *International Journal of Speech Language and the Law, 2*, 143-153. doi: 10.1558/ijsll.v2i2.143

Hollien, H., Majewski, W. et Doherty, E. T. (1982). Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics, 10*, 139-148. doi: 10.1016/S0095-4470(19)30953-2

Hood, B. M., Macrae, C. N., Cole-Davies, V. et Dias, M. (2003). Eye remember you: The effects of gaze direction on face recognition in children and adults. *Developmental Science, 6*, 67-71. doi: 10.1111/1467-7687.00256

Humble, D., Schweinberger, S. R., Dobel, C. et Zäske, R. (2019). Voices to remember: Comparing neural signatures of intentional and non-intentional voice learning and recognition. *Brain research, 1711*, 214-225. doi: 10.1016/j.brainres.2019.01.028

Imaizumi, S., Mori, K., Kiritani, S., Kawashima, R., Sugiura, M., Fukuda, H., . . . Hatano, K. (1997). Vocal identification of speaker and emotion activates differerent brain regions. *Neuroreport, 8*, 2809-2812. doi: 10.1097/00001756-199708180-00031

Jessen, M. (2008). Forensic phonetics. *Language and linguistics compass, 2*, 671-711. doi: 10.1111/j.1749-818X.2008.00066.x

Jiang, J., Borowiak, K., Tudge, L., Otto, C. et von Kriegstein, K. (2017). Neural mechanisms of eye contact when listening to another person talking. *Social cognitive and affective neuroscience, 12*, 319-328. doi: 10.1093/scan/nsw127

Johnson, M. K., De Leonardis, D. M., Hashtroudi, S. et Ferguson, S. A. (1995). Aging and single versus multiple cues in source monitoring. *Psychology and Aging, 10*, 507. doi: 10.1037/0882-7974.10.4.507

Kaufmann, J. M., Schweinberger, S. R. et Burton, A. M. (2009). N250 ERP correlates of the acquisition of face representations across different images. *Journal of Cognitive Neuroscience, 21*, 625-641.

Kerstholt, J. H., Jansen, N. J. M., van Amelsvoort, A. G. et Broeders, A. P. A. (2006). Earwitnesses: Effects of accent, retention and telephone. *Applied Cognitive Psychology, 20*, 187-197. doi: 10.1002/acp.1175

Kisilevsky, B. S., Hains, S. M., Brown, C. A., Lee, C. T., Cowperthwaite, B., Stutzman, S. S., . . . Huang, H. (2009). Fetal sensitivity to properties of maternal speech and language. *Infant Behavior and Development, 32*, 59-71. doi: 10.1016/j.infbeh.2008.10.002

Kisilevsky, B. S., Hains, S. M., Lee, K., Xie, X., Huang, H., Ye, H. H., . . . Wang, Z. (2003). Effects of experience on fetal voice recognition. *Psychological Science, 14*, 220-224. doi: 10.1111/1467-9280.02435

Köster, O., Hess, M. M., Schiller, N. O. et Künzel, H. J. (1998). The correlation between auditory speech sensitivity and speaker recognition ability. *International Journal of Speech, Language and the Law, 5*, 22-32. doi: 10.1558/sll.1998.5.1.22

Köster, O. et Schiller, N. O. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics, 4*, 18-28. doi: 10.1558/ijsll.v4i1.18

Kreiman, J. et Papcun, G. (1991). Comparing discrimination and recognition of unfamiliar voices. *Speech Communication, 10*, 265-275. doi: 10.1016/0167-6393(91)90016-M

Kreiman, J. et Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons. doi:10.1002/9781444395068

Kroese, D. P., Taimre, T. et Botev, Z. I. (2013). *Handbook of monte carlo methods*. John Wiley & Sons. doi:10.1002/9781118014967

Lafleur, A. et Boucher, V. J. (2015). The ecology of self-monitoring effects on memory of verbal productions: Does speaking to someone make a difference? *Consciousness and Cognition, 36*, 139-146. doi: 10.1016/j.concog.2015.06.015

Laub, C. E., Wylie, L. E. et Bornstein, B. H. (2013). Can the courts tell an ear from an eye: Legal approaches to voice identification evidence. *Law and Psychology Review, 37*, 119-158.

Lavner, Y., Rosenhouse, J. et Gath, I. (2001). The prototype model in speaker identification by human listeners. *International Journal of Speech Technology, 4*, 63-74. doi: 10.1023/A:1009656816383

Lee, G. Y. et Kisilevsky, B. S. (2014). Fetuses respond to father's voice but prefer mother's voice after birth. *Developmental psychobiology, 56*, 1-11. doi: 10.1002/dev.21084

Legge, G. E., Grosmann, C. et Pieper, C. M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 298. doi: 10.1037/0278-7393.10.2.298

Lehmann, D. et Skrandies, W. (1980). Reference-free identification of components of checkerboard-evoked multichannel potential fields. *Electroencephalography and clinical Neurophysiology, 48*, 609-621. doi: 10.1016/0013-4694(80)90419-8

Levi, S. V. (2018). Another bilingual advantage? Perception of talker-voice information. *Bilingualism: Language and Cognition, 21*, 523-536. doi: 10.1017/S1366728917000153

Levy, D., Granot, R. et Bentin, S. (2003). Neural sensitivity to human voices: ERP evidence of task and attentional influences. *Psychophysiology, 40*, 291-305. doi: 10.1111/1469-8986.00031

Levy, D. A., Granot, R. et Bentin, S. (2001). Processing specificity for human voice stimuli: Electrophysiological evidence. *Neuroreport, 12*, 2653-2657. doi: 10.1097/00001756-200108280-00013

Locke, J. L. et Bogin, B. (2006). Language and life history: A new perspective on the development and evolution of human language. *Behavioral and Brain Sciences, 29*, 259. doi: 10.1017/S0140525X0600906X

Lucchelli, F. et Spinnler, H. (2008). A reappraisal of person recognition and identification. *Cortex, 44*, 230-237. doi: 10.1016/j.cortex.2006.11.001

Luck, S. J. et Kappenman, E. S. (2011). *The Oxford handbook of event-related potential components*. New York: Oxford university press.

Luzzi, S., Coccia, M., Polonara, G., Reverberi, C., Ceravolo, G., Silvestrini, M., . . . Gainotti, G. (2017). Selective associative phonagnosia after right anterior temporal stroke. *Neuropsychologia*. doi: 10.1016/j.neuropsychologia.2017.05.016

Macrae, C. N., Hood, B. M., Milne, A. B., Rowe, A. C. et Mason, M. F. (2002). Are you looking at me? Eye gaze and person perception. *Psychological Science, 13*, 460-464. doi: 10.1111/1467-9280.00481

Maguinness, C., Roswandowitz, C. et von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia*. doi: 10.1016/j.neuropsychologia.2018.03.039

Mai, X., Xu, L., Li, M., Shao, J., Zhao, Z., deRegnier, R.-A., . . . Lozoff, B. (2012). Auditory recognition memory in 2-month-old infants as assessed by event-related potentials. *Developmental neuropsychology, 37*, 400-414. doi: 10.1080/87565641.2011.650807

Manesi, Z., Van Lange, P. A. et Pollet, T. V. (2016). Eyes wide open: Only eyes that pay attention promote prosocial behavior. *Evolutionary Psychology, 14*, 1474704916640780. doi: 10.1177/1474704916640780

Mann, V. A., Diamond, R. et Carey, S. (1979). Development of voice recognition: Parallels with face recognition. *Journal of Experimental Child Psychology, 27*, 153-165. doi: 10.1016/0022-0965(79)90067-5

Marcus, S. M. (1981). Acoustic determinants of perceptual center (P-center) location. *Perception & Psychophysics, 30*, 247-256. doi: 10.3758/BF03214280

Maris, E. et Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods, 164*, 177-190. doi: 10.1016/j.jneumeth.2007.03.024

Marzi, T. et Viggiano, M. P. (2007). Interplay between familiarity and orientation in face processing: An ERP study. *International Journal of Psychophysiology, 65*, 182-192. doi: 10.1016/j.ijpsycho.2007.04.003

Matheson, H. E. et Barsalou, L. W. (2018). Embodiment and grounding in cognitive neuroscience. *Stevens' handbook of experimental psychology and cognitive neuroscience, 3*, 1-27. doi: 10.1002/9781119170174.epcn310

McAllister, H. A., Dale, R. H., Bregman, N. J., McCabe, A. et Cotton, C. R. (1993). When eyewitnesses are also earwitnesses: Effects on visual and voice identifications. *Basic and Applied Social Psychology, 14*, 161-170. doi: 10.1207/s15324834basp1402_3

McGehee, F. (1937). The reliability of the identification of the human voice. *Journal of General Psychology, 17*, 249-271. doi: 10.1080/00221309.1937.9917999

McGehee, F. (1944). An experimental study of voice recognition. *Journal of General Psychology, 31*, 53-65. doi: 10.1080/00221309.1944.10545219

Mehler, J., Bertoncini, J., Barriere, M. et Jassik-Gerschenfeld, D. (1978). Infant recognition of mother's voice. *Perception, 7*, 491-497. doi: 10.1068/p070491

Moon, C. et Fifer, W. P. (1990). Syllables as signals for 2-day-old infants. *Infant Behavior and Development, 13*, 377-390. doi: 10.1016/0163-6383(90)90041-6

Morton, J., Marcus, S. et Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review, 83*, 405. doi: 10.1037/0033-295X.83.5.405

Mullennix, J. W., Pisoni, D. B. et Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85*, 365-378. doi: 10.1121/1.397688

Näätänen, R., Gaillard, A. W. et Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta psychologica, 42*, 313-329. doi: 10.1016/0001-6918(78)90006-9

Näätänen, R. et Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology, 24*, 375-425. doi: 10.1111/j.1469-8986.1987.tb00311.x

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Natano, K., . . . Kojima, S. (2001). Neural substrates for recognition of familiar voices: A PET study. *Neuropsychologia, 39*, 1047-1054. doi: 10.1016/S0028-3932(01)00037-9

Neuner, F. et Schweinberger, S. R. (2000). Neuropsychological impairments in the recognition of faces, voices, and personal names. *Brain and Cognition, 44*, 342-366. doi: 10.1006/brcg.1999.1196

Nolan, F. (1997). Speaker recognition and forensic phonetics. *Handbook of phonetic sciences*, 744-767. doi: 10.1002/9781444317251

Nolan, F. (2003). A recent voice parade. *Forensic Linguistics, 10*, 277-291. doi: 10.1558/sll.2003.10.2.277

Nolan, F. et Grabe, E. (1996). Preparing a voice lineup. *International Journal of Speech, Language and the Law, 3*, 74-94. doi: 10.1558/ijsll.v3i1.74

Nolan, F., McDougall, K. et Hudson, T. (2013). Effects of the telephone on perceived voice similarity: Implications for voice line-ups. *International Journal of Speech, Language and the Law, 20*, 229-246. doi: 10.1558/ijsll.v20i2.229

O'Mahony, C. et Newell, F. N. (2012). Integration of faces and voices, but not faces and names, in person recognition. *British Journal of Psychology, 103*, 73-82. doi: 10.1111/j.2044-8295.2011.02044.x

Öhman, L., Eriksson, A. et Granhag, P. A. (2010). Mobile phone quality vs. direct quality: How the presentation format affects earwitness identification accuracy. *European Journal of Psychology Applied to Legal Context, 2*.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*, 97-113. doi: 10.1016/0028-3932(71)90067-4

Oostenveld, R., Fries, P., Maris, E. et Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience, 2011*, 1. doi: 10.1155/2011/156869

Orchard, T. L. et Yarmey, A. D. (1995). The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology, 9*, 249-260. doi: 10.1002/acp.2350090306

Papcun, G., Kreiman, J. et Davis, A. (1989). Long-term memory for unfamiliar voices. *Journal of the Acoustical Society of America, 85*, 913-925. doi: 10.1121/1.397564

Peretz, I., Kolinsky, R., Tramo, M., Labrecque, R., Hublet, C., Demeurisse, G. et Belleville, S. (1994). Functional dissociations following bilateral lesions of auditory cortex. *Brain, 117*, 1283-1301. doi: 10.1093/brain/117.6.1283

Perfect, T. J., Hunt, L. J. et Harris, C. M. (2002). Verbal overshadowing in voice recognition. *Applied Cognitive Psychology, 16*, 973-980. doi: 10.1002/acp.920

Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E., . . . Valdes-Sosa, M. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage, 119*, 164-174. doi: 10.1016/j.neuroimage.2015.06.050

Perrachione, T. K. (2017). Speaker recognition across languages. Dans S. Frühholz & P. Belin (dir.), *Oxford Handbook of Voice Perception*. Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780198743187.001.0001

Perrachione, T. K. et Wong, P. C. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia, 45*, 1899-1910. doi: 10.1016/j.neuropsychologia.2006.11.015

Perrodin, C., Kayser, C., Abel, T. J., Logothetis, N. K. et Petkov, C. I. (2015). Who is that? Brain networks and mechanisms for identifying individuals. *Trends in Cognitive Sciences, 19*, 783-796. doi: 10.1016/j.tics.2015.09.002

Pinheiro, A. P., Rezaii, N., Nestor, P. G., Rauber, A., Spencer, K. M. et Niznikiewicz, M. (2016). Did you or I say pretty, rude or brief? An ERP study of the effects of speaker's identity on emotional word processing. *Brain and language, 153*, 38-49. doi: 10.1016/j.bandl.2015.12.003

Plante-Hébert, J. et Boucher, V. J. (2014). *L'identification vocale: Pour une quantification des effets de la familiarité.* Communication présentée Journée d'Études sur la Parole, Le Mans. Repéré à http://www.afcp-parole.org/doc/Archives_JEP/2014_XXXe_JEP_LeMans/2014_XXXe_JEP_LeMans.pdf

Plante-Hébert, J. et Boucher, V. J. (2015a). *Effects of nasality and utterance length on the recognition of familiar speakers.* Communication présentée 18th International Congress of Phonetic Sciences, Glasgow. Repéré à

https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0772.pdf

Plante-Hébert, J. et Boucher, V. J. (2015b). *L'effet de la familiarité sur l'identification des locuteurs : Pour un perfectionnement de la parade vocale.* (Université de Montréal, Montréal). Repéré à http://hdl.handle.net/1866/11890

Plante-Hébert, J., Boucher, V. J. et Jemel, B. (2021). The processing of intimately familiar and unfamiliar voices: Specific neural responses of speaker recognition and identification. *PLoS ONE, 16*, e0250214. doi: 10.1371/journal.pone.0250214

Pollack, I., Pickett, J. M. et Sumby, W. H. (1954). On identification of speakers by voice. *Journal of the Acoustical Society of America, 26*, 403-406. doi: 10.1121/1.1907349

Rachman, L. (2018). *The" other-voice" effect: How speaker identity and language familiarity influence the way we process emotional speech.* (Sorbonne Université, Paris). Repéré à https://hal.archives-ouvertes.fr/tel-01983748

Robertson, D. M. C. et Schweinberger, S. R. (2010). The role of audiovisual asynchrony in person recognition. *Quarterly Journal of Experimental Psychology, 63*, 23-30. doi: 10.1080/17470210903144376

Robinson, J. (2014). Likert Scale. Dans A. C. Michalos (dir.), *Encyclopedia of Quality of Life and Well-Being Research* (p. 3620-3621). Dordrecht: Springer Netherlands. doi:10.1007/978-94-007-0753-5_1654

Roebuck, R. et Wilding, J. (1993). Effects of vowel variety and sample lenght on identification of a speaker in a line-up. *Applied Cognitive Psychology, 7*, 475-481. doi: 10.1002/acp.2350070603

Rogier, O., Roux, S., Belin, P., Bonnet-Brilhaut, F. et Bruneau, N. (2010). An electrophysiological correlate of voice processing in 4- to 5-year-old children. *International Journal of Psychophysiology, 75*, 44-47. doi: 10.1016/j.ijpsycho.2009.10.013

Roswandowitz, C., Kappes, C., Obrig, H. et von Kriegstein, K. (2017). Obligatory and facultative brain regions for voice-identity recognition. *Brain, 141*, 234-247. doi: 10.1093/brain/awx313

Roswandowitz, C., Maguinness, C. et von Kriegstein, K. (2019). Deficits in voice-identity processing: Acquired and developmental phonagnosia. Dans S. Frühholz & P. Belin (dir.), *Oxford handbook of Voice Perception*. Oxford: Oxford University Press. doi:10.20944/preprints201806.0280.v2

Roswandowitz, C., Mathias, S. R., Hintz, F., Kreitewolf, J., Schelinski, S. et von Kriegstein, K. (2014). Two cases of selective developmental voice-recognition impairments. *Current Biology, 24*, 2348-2353. doi: 10.1016/j.cub.2014.08.048

Roswandowitz, C., Schelinski, S. et von Kriegstein, K. (2017). Developmental phonagnosia: Linking neural mechanisms with the behavioural phenotype. *NeuroImage, 155*, 97-112. doi: 10.1016/j.neuroimage.2017.02.064

Rugg, M. D. et Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences, 11*, 251-257. doi: 10.1016/j.tics.2007.04.004

Schall, S., Kiebel, S. J., Maess, B. et von Kriegstein, K. (2013). Early auditory sensory processing of voices is facilitated by visual mechanisms. *NeuroImage, 77*, 237-245. doi: 10.1016/j.neuroimage.2013.03.043

Schall, S., Kiebel, S. J., Maess, B. et von Kriegstein, K. (2014). Voice identity recognition: Functional division of the right STS and its behavioral relevance. *Journal of Cognitive Neuroscience, 27*, 280-291. doi: 10.1162/jocn_a_00707

Schelinski, S., Riedel, P. et von Kriegstein, K. (2014). Visual abilities are important for auditory-only speech recognition: Evidence from autism spectrum disorder. *Neuropsychologia, 65*, 1-11. doi: 10.1016/j.neuropsychologia.2014.09.031

Schiller, N. O. et Köster, O. (1996). Evaluation of a foreign speaker in forensic phonetics: A report. *International Journal of Speech, Language and the Law, 3*, 176-185. doi: 10.1558/ijsll.v3i1.176

Schiller, N. O. et Köster, O. (1998). The ability of expert witnesses to identify voices: a comparison between trained and untrained listeners. *Forensic Linguistics, 5*, 1-9. doi: 10.1558/ijsll.v5i1.1

Schmidt-Nielsen, A. et Stern, K. R. (1985). Identification of known voices as a function of familiarity and narrow-band coding. *Journal of the Acoustical Society of America, 77*, 658-663. doi: 10.1121/1.391884

Schweinberger, S. R. (2001). Human brain potential correlates of voice priming and voice recognition. *Neuropsychologia, 39*, 921-936. doi: 10.1016/S0028-3932(01)00023-9

Schweinberger, S. R., Herholz, A. et Sommer, W. (1997). Recognizing famous voices influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language, and Hearing Research, 40*, 453-463. doi: 10.1044/jslhr.4002.453

Schweinberger, S. R., Kawahara, H., Simpson, A. P., Skuk, V. G. et Zäske, R. (2014). Speaker perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*, 15-25. doi: 10.1002/wcs.1261

Schweinberger, S. R., Kloth, N. et Robertson, D. (2011). Hearing facial identities: Brain correlates of face-voice integration in person identification. *Cortex, 47*, 1026-1037. doi: 10.1016/j.cortex.2010.11.011

Schweinberger, S. R., Robertson, D. et Kaufman, J. M. (2007). Hearing facial identities. *The Quarterly Journal of Experimental Psychology, 60*, 1446-1456. doi: 10.1080/17470210601063589

Schweinberger, S. R., Walther, C., Zäske, R. et Kovács, G. (2011). Neural correlates of adaptation to voice identity. *British Journal of Psychology, 102*, 748-764. doi: 10.1111/j.2044-8295.2011.02048.x

Sheffert, S. M. et Olson, E. (2004). Audiovisual speech facilitates voice learning. *Perception & Psychophysics, 66*, 352-362. doi: 10.3758/BF03194884

Sherrin, C. (2014). Earwitness evidence: the reliability of voice identifications. *Osgoode Hall Law Journal, 52*, 819-865. doi: 10.2139/ssrn.2628313

Sidtis, D. et Kreiman, J. (2012). In the beginning was the familiar voice: Personally familiar voices in the evolutionary and contemporary biology of communication. *Integrative Psychological and Behavioral Science, 46*, 146-159. doi: 10.1007/s12124-011-9177-4

Skuk, V. G. et Schweinberger, S. R. (2013). Gender differences in familiar voice identification. *Hearing research, 296*, 131-140. doi: 10.1016/j.heares.2012.11.004

Smith, I., Foulkes, P. et Sóskuthy, M. (2017). Speaker identification in whisper. *Letras de Hoje, 52*, 5-14. doi: 10.15448/1984-7726.2017.1.26659

Solan, L. M. et Tiersma, P. M. (2002). Hearing voices: Speaker identification in court. *Hastings Law Journal, 54*, 373.

Sørensen, M. H. (2012). Voice line-ups: Speakers' F0 values influence the reliability of voice recognitions. *International Journal of Speech, Language and the Law, 19*. doi: 10.1558/ijsll.v19i2.145

Speaks, C. E. (2017). *Introduction to sound: acoustics for the hearing and speech sciences*. Plural Publishing. doi:10.1007/978-1-4899-7196-8

Spreckelmeyer, K. N., Kutas, M., Urbach, T., Altenmüller, E. et Münte, T. F. (2009). Neural processing of vocal emotion and identity. *Brain and Cognition, 69*, 121-126. doi: 10.1016/j.bandc.2008.06.003

Stevenage, S. V. (2018). Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings. *Neuropsychologia, 116*, 162-178. doi: 10.1016/j.neuropsychologia.2017.07.005

Stevenage, S. V., Clarke, G. et McNeill, A. (2012). The “other-accent” effect in voice recognition. *Journal of Cognitive Psychology, 24*, 647-653. doi: 10.1080/20445911.2012.675321

Stevenage, S. V., Hale, S., Morgan, Y. et Neil, G. J. (2014). Recognition by association: Within-and cross-modality associative priming with faces and voices. *British Journal of Psychology, 105*, 1-16. doi: 10.1111/bjop.12011

Stevenage, S. V., Howland, A. et Tippelt, A. (2011). Interference in eyewitness and earwitness recognition. *Applied Cognitive Psychology, 25*, 112-118. doi: 10.1002/acp.1649

Stevenage, S. V. et Neil, G. J. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica, 54*, 266-281. doi: 10.5334/pb.ar

Stevenage, S. V., Neil, G. J., Parsons, B. et Humphreys, A. (2018). A sound effect: Exploration of the distinctiveness advantage in voice recognition. *Applied Cognitive Psychology, 32*, 526-536. doi: 10.1002/acp.3424

Su, L. S., Li, K. P. et Fu, K. S. (1974). Identification of speakers by use of nasal coarticulation. *Journal of the Acoustical Society of America, 56*, 1876-1883. doi: 10.1121/1.1903526

Sullivan, K. P. et Schlichting, F. (2007). Speaker discrimination in a foreign language: First language environment, second language learners. *International Journal of Speech, Language and the Law, 7*, 95-112. doi: 10.1558/sll.2000.7.1.95

Tanaka, J. W., Curran, T., Porterfield, A. L. et Collins, D. (2006). Activation of preexisting and acquired face representations: The N250 event-related potential as an index of face familiarity. *Journal of Cognitive Neuroscience, 18*, 1488-1497. doi: 10.1162/jocn.2006.18.9.1488

Thompson, C. P. (1985). Voice identification: Speaker identifiability and a correction of the record regarding sex effects. *Human Learning: Journal of Practical Research & Applications, 4*, 19-27.

Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology, 1*, 121-131.

Tomlin, R. J., Stevenage, S. V. et Hammond, S. (2016). Putting the pieces together: Revealing face–voice integration through the facial overshadowing effect. *Visual Cognition, 25*, 629-643. doi: 10.1080/13506285.2016.1245230

Tulving, E. (1972). Episodic and semantic memory. *Organization of memory, 1*, 381-403.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie canadienne, 26*, 1. doi: 10.1037/h0080017

Tulving, E., Kapur, S., Craik, F., Moscovitch, M. et Houle, S. (1994). *Hemispheric encoding/retrieval asymmetry in episodic memory: Positron emission tomography findings*. Communication présentée Proceedings of the National Academy of Sciences, U.S.A. doi: 10.1073/pnas.91.6.2016

Tulving, E. et Murray, D. (1985). Elements of episodic memory. *Canadian Psychology, 26*, 235-238.

Van Lancker, D. et Canter, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain and Cognition, 1*, 185-195. doi: 10.1016/0278-2626(82)90016-1

Van Lancker, D. et Kreiman, J. (1985). Unfamiliar voice discrimination and familiar voice recognition are independent and unordered abilities. *UCLA Working Papers in Phonetics*, 50-60.

Van Lancker, D., Kreiman, J. et Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices. *Journal of Phonetics, 13*, 19-38. doi: 10.1016/S0095-4470(19)30723-5

Van Lancker, D. R., Cummings, J. L., Kreiman, J. et Dobkin, B. H. (1988). Phonagnosia: A dissociation between familiar and unfamiliar voices. *Cortex, 24*, 195-209. doi: 10.1016/S0010-9452(88)80029-7

Van Lancker, D. R. et Kreiman, J. (1987). Voice discimination and recognition are separate abilities. *Neuropsychologia, 25*, 829-834. doi: 10.1016/0028-3932(87)90120-5

Von Kriegstein, K., Dogan, Ö., Grüter, Martina, Giraud, A.-L., Kleinschmidt, A. et Kiebel, S. J. (2008). *Simulation of talking faces in the human brain improves auditory speech recognition.* Communication présentée PNAS. doi: 10.1073/pnas.0710826105

Von Kriegstein, K. et Giraud, A.-L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage, 22*, 948-955. doi: 10.1016/j.neuroimage.2004.02.020

Von Kriegstein, K. et Giraud, A.-L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology, 4*, e326. doi: 10.1371/journal.pbio.0040326

Von Kriegstein, K., Kleinschmidt, A. et Giraud, A.-L. (2006). Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. *Cerebral Cortex, 16*, 1314-1322. doi: 10.1093/cercor/bhj073

Von Kriegstein, K., Kleinschmidt, A., Sterzer, P. et Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience, 17*, 367-376. doi: 10.1162/0898929053279577

Vuilleumier, P., George, N., Lister, V., Armony, J. et Driver, J. (2005). Effects of perceived mutual gaze and gender on face processing and recognition memory. *Visual Cognition, 12*, 85-101.

Watkins, K. E., Strafella, A. P. et Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia, 41*, 989-994. doi: 10.1016/S0028-3932(02)00316-0

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A. et Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior, 44*, 3. doi: 10.1037/lhb0000359

Wheeler, M. A., Stuss, D. T. et Tulving, E. (1997). Toward a theory of episodic memory: The frontal lobes and autonoetic consciousness. *Psychological bulletin, 121*, 331. doi: 10.1037/0033-2909.121.3.331

Wilding, E. L. et Ranganath, C. (2011). Electrophysiological correlates of episodic memory processes. Dans S. J. Luck & E. S. Kappenman (dir.), *Oxford handbook of event-related potential components*. New York: Oxford University Press. doi:10.1093/oxfordhb/9780195374148.013.0187

Wilding, J. et Cook, S. (2000). Sex differences and individual consistency in voice identification. *Perceptual and Motor Skills, 91*, 535-538. doi: 10.2466/pms.2000.91.2.535

Wilson, S. M., Saygin, A. P., Sereno, M. I. et Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7*, 701-702. doi: 10.1038/nn1263

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception, & Psychophysics, 72*, 2031-2046. doi: 10.3758/BF03196680

Xie, X. et Myers, E. (2015). The impact of musical training and tone language experience on talker identification. *Journal of the Acoustical Society of America, 137*, 419-432. doi: 10.1121/1.4904699

Xu, X., Biederman, I., Shilowich, B. E., Herald, S. B., Amir, O. et Allen, N. E. (2015). Developmental phonagnosia: Neural correlates and a behavioral marker. *Brain and language, 149*, 106-117. doi: 10.1016/j.bandl.2015.06.007

Yarmey, A. D. (1991). Voice identification over the telephone 1. *Journal of Applied Social Psychology, 21*, 1868-1876. doi: 10.1111/j.1559-1816.1991.tb00510.x

Yarmey, D. A. (2001). Earwitness descriptions and speaker identification. *Forensic Linguistics, 8*, 113-122. doi: 10.1558/sll.2001.8.1.113

Yarmey, D. A. (2014). The psychology of speaker identification and earwitness memory (*Handbook Of Eyewitness Psychology* (Vol. 2, p. 115-150). New Jersey, États-Unis: Routledge. doi:10.4324/9780203936368

Yarmey, D. A., Yarmey, L. A., Yarmey, M. J. et Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology, 15*, 183-299. doi: 10.1002/acp.702

Yonelinas, A. P. (2001). Components of episodic memory: The contribution of recollection and familiarity. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 356*, 1363-1374. doi: 10.1098/rstb.2001.0939

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441-517. doi: 10.1006/jmla.2002.2864

Young, A. W., Frühholz, S. et Schweinberger, S. R. (2020). Face and voice perception: Understanding commonalities and differences. *Trends in Cognitive Sciences*. doi: 10.1016/j.tics.2020.02.001

Zäske, R., Limbach, K., Schneider, D., Skuk, V. G., Dobel, C., Guntinas-Lichius, O. et Schweinberger, S. R. (2018). Electrophysiological correlates of voice memory for young and old speakers in young and old listeners. *Neuropsychologia*. doi: 10.1016/j.neuropsychologia.2017.08.011

Zäske, R., Mühl, C. et Schweinberger, S. R. (2015). Benefits for voice learning caused by concurrent faces develop over time. *PLoS ONE, 10*, e0143151. doi: 10.1371/journal.pone.0143151

Zäske, R., Volberg, G., Kovács, G. et Schweinberger, S. R. (2014a). Electrophysiological correlates of voice learning and recognition. *Journal of Neuroscience, 34*, 10821-10831. doi: 10.1523/JNEUROSCI.0581-14.2014

Zäske, R., Volberg, G., Kovács, G. et Schweinberger, S. R. (2014b). Electrophysiological Correlates of Voice Learning and Recognition. *The Journal of Neuroscience, 34*, 10821-10831. doi: 10.1523/jneurosci.0581-14.2014

Zhang, C., Pugh, K. R., Mencl, W. E., Molfese, P. J., Frost, S. J., Magnuson, J. S., . . . Wang, W. S. Y. (2016). Functionally integrated neural processing of linguistic and talker information: An event-related fMRI and ERP study. *NeuroImage, 124*, 536-549. doi: 10.1016/j.neuroimage.2015.08.064

Zinke, K., Thöne, L., Bolinger, E. M. et Born, J. (2018). Dissociating long and short-term memory in three-month-old infants using the mismatch response to voice stimuli. *Frontiers in psychology, 9*, 31. doi: 10.3389/fpsyg.2018.00031

Annexes

Caractéristiques des locuteurs					
Âge	F_0mp	Région natale	Âge	F_0mp	Région natale
34	129,5	Chaudière-Appalaches	29	114,6	Montréal
30	111,2	Centre-du-Québec	44	91,3	Montréal
39	125,2	Montréal	38	127,0	Ontario
22	120,8	Montréal	43	125,9	Outaouais
31	122,5	Montréal	29	107,5	Montréal
20	118,7	Outaouais	22	152,1	Laurentides
30	105,8	Lanaudière	27	127,3	Montérégie
29	116,9	Québec	29	122,5	Montréal
31	127,2	Montréal	34	120,0	Montérégie
36	112,9	Outaouais	24	99,8	Lanaudière
34	140,9	Montréal	19	119,4	Montréal
25	149,5	Montréal	21	121,8	Laurentides
26	115,6	Montréal	20	140,4	Montréal
30	126,8	Montréal	23	94,2	Montréal
26	110,1	Québec	22	103,8	Montréal
23	97,5	Montréal	20	122,4	Montréal
24	107,5	Montréal	28	126,2	Lanaudière
60	92,2	Chaudière-Appalaches	24	132,6	Montréal
25	107,1	Québec	24	132,7	Lanaudière

Tableau 6. – Statistiques descriptives des locuteurs enregistrés dans l'élaboration des stimuli de l'expérience 1. Les cases ombragées représentes les locuteurs retenus.

Stimuli d'apprentissage				
agrume	boulot	chameau	cheveu	divan
écaille	entête	festin	guidon	homard
kayak	lézard	ourson	ovale	poumon
raisin	rouleau	salon	théière	wagon

Tableau 7. – Stimuli d'entrainement l'expérience 2 en orthographe standard et en ordre alphabétique

Stimuli expérimentaux			
L1	L2	L3	L4
accès	acteur	agence	alarme
acheteur	adulte	album	arcade
adresse	antenne	asile	assise
annexe	arome	auberge	autruche
annonce	assiette	avenue	balai
argent	baguette	balade	banane
automne	ballon	banquet	bassin
bagage	bandeau	baron	berceau
bambou	billet	bonnet	bouchée
bétail	bonhomme	boucherie	bouteille
bijou	bouton	bovin	cadet
bouillon	bureau	cadeau	canon
buffet	canard	carotte	casier
canal	chalet	chauffage	chapeau
carré	chasseur	chrono	chariot
cerveau	cheminée	cité	chauffeur
cheval	chorale	coffret	clocher
cheville	ciseau	coquille	cohorte
colonne	comète	culotte	colis
commis	coussin	dauphin	corail
courrier	couture	diner	débris
déesse	démo	dossier	dessin
denrée	détenu	drapeau	douzaine
donneur	dosage	éclat	dureté
écran	échelle	équipe	école
façade	engrais	érable	enseigne
forage	épine	fauteuil	essence
fusée	éponge	fourrure	étage
horloge	farine	gâteau	faisceau
iris	foulard	kiosque	feuillage
joyau	gala	labo	galette
jumeau	jury	lotus	laurier
local	loterie	maman	logo
marée	loyer	manteau	lunette
mécène	maillon	menu	marché
morceau	mari	métal	marin
motard	médaille	muraille	métro
neurone	moteur	olive	monnaie
objet	nickel	orbite	musique
orange	noyau	ordi	orage

panier	oiseau	paquebot	palais
pavé	papier	parade	paroi
pilier	parrain	pépin	piano
poignet	péage	photo	pignon
police	pillage	pilote	pilule
purée	poubelle	racine	pivot
recette	rallye	reçu	radar
rival	recueil	reflet	régie
robot	resto	roman	remède
saumon	rocher	roulette	routier
sifflet	roquette	ruban	ruisseau
stylo	sirène	rugby	sabot
tailleur	tango	soleil	sonnette
tapis	taureau	taxi	terreau
troupeau	tennis	télé	théâtre
vaisselle	tuyau	ténor	tissu
vallée	vapeur	velours	vernis
voleur	vélo	verger	vignette
voyou	village	verrou	visa

Tableau 8. – Stimuli de l'expérience 2 en orthographe standard et en ordre alphabétique pour chaque locuteur. Seuls les stimuli de la première partie présentée dans l'article 2 sont présents.