Caractérisation psycho-acoustique de l'identité du locuteur

Etienne Gaudrain, Roy D. Patterson

10ème **C**ongrès **F**rançais d'**A**coustique

Lyon, 12-16 Avril 2010

Caractérisation psycho-acoustique de l'identité du locuteur

Etienne Gaudrain[1,2], Roy D. Patterson[2]

[1] MRC Cognition and Brain Sciences Unit, Cambridge, Royaume-Uni, etienne.gaudrain@mrc-cbu.cam.ac.uk

[2] Centre for the Neural Basis of Hearing, University of Cambridge, Royaume-Uni, rdp1@cam.ac.uk

L'identité vocale d'un locuteur est largement définie par l'anatomie de son appareil phonatoire. La hauteur fondamentale de sa voix est définie par la taille et la masse de ses cordes vocales (ainsi que par la tension qu'il y applique). La longueur de son tractus vocal définie une grandeur d'*échelle acoustique* que l'auditeur associera à la taille du locuteur. En combinant ces deux informations, il est possible d'estimer la taille, l'âge et le genre du locuteur à partir se sa voix. Cependant, la fréquence fondamentale est rarement fixe et varie pour participer à la prosodie. De même, la longueur du tractus vocal varierait faiblement pour transmettre certaines émotions. L'étude présentée ici vise à caractériser les variations de hauteur fondamentale et de longueur de tractus vocal qu'un auditeur peut tolérer avant de décider que le locuteur a changé. Les résultats montrent que les auditeurs sont bien plus tolérants pour un changement de hauteur fondamental que pour un changement de longueur de tractus vocal, alors même qu'ils sont moins sensibles à ce dernier. Ces résultats concordent avec les variations observées dans des enregistrements de parole et suggèrent que les auditeurs ont développé un modèle de locuteur par apprentissage. Les implications de ces résultats pour l'étude de l'analyse des scènes auditives et pour l'étude de l'organisation du système auditif sont discutées.

1 Introduction

La fréquence d'oscillation des cordes vocales (FCV [1]) et la longueur du tractus vocal (LTV [1]) sont deux paramètres physiques fondamentaux de l'appareil vocal. La FCV est influencée par la masse et la longueur des cordes vocales et détermine la fréquence fondamentale (F_0) de la voix. La LTV détermine la durée des résonances formantiques dues aux cavités bucales. Cette dimension peut donc être qualifiée d'*échelle acoustique* (S_f). De récentes études ont montré qu'un auditeur peut déterminer la taille d'un locuteur [1] ainsi que son âge et son genre [1, 2] uniquement à partir de ces deux variables. Ces caractéristiques vocales jouent donc un rôle important pour la perception de l'identité d'un locuteur, et leur perception est ainsi essentielle pour la séparation perceptive de voix concurrentes [3, 4].

Cependant, alors que la longueur du tractus vocal est une caractéristique anatomique fixe, la fréquence d'oscillation des cordes vocales peut aussi être influencée par la tension des cordes vocales que le locuteur peut modifier à loisir. La hauteur fondamentale ($\hat{F}_0$) de la voix, liée à la FCV, ne semble donc pas un indice très fiable pour l'identification d'un locuteur. Il peut ainsi paraître surprenant que le système auditif soit en réalité plus sensible à une différence de F_0 (une difference de 0,3 demitons est détectée de façon fiable [5]) qu'à une différence d'échelle acoustique (un différence de 0,5 demitons [2] est détectée de façon fiable [6]). Ceci signifie que même s'il dispose d'une représentation très claire de la différence de FCV qui sépare deux voix, un auditeur ne pourra déterminer si cette différence est causée par la présence de deux locuteurs ayant des FCV moyennes différentes, ou s'il s'agit du même locuteur qui aurait modifié la hauteur de sa voix [7].

L'objectif de la présente étude était de quantifier les variations de FCV et LTV qu'un auditeur tolère avant de juger que l'identité du locuteur a changé. Étant donné que la FCV d'un locuteur change avec la prosodie, il était attendu que les auditeurs soient plus tolérants pour des changements suivant cette dimension que pour des changements de LTV qui supposent un changement de taille du locuteur. Par ailleurs, cette tolérance a été évaluée pour des locuteurs présentant des combinaisons typiques et atypiques de FCV et LTV. Si le motif de tolérance observé pour des locuteurs typiques se réplique pour des locuteurs atypiques, celà signifie que le critère de jugement est transférable à des voix auxquelles nous sommes peu ou pas exposés.

2 Méthode

Les participants devaient juger la similarité entre deux locuteurs en indiquant s'il leur semblait possible que deux stimuli aient été prononcés par le même locuteur. Deux expériences légèrement différentes ont été réalisées. Les éléments méthodologiques communs sont d'abord décrit ci-dessous.

De nombreux indices perceptifs peuvent être utilisés pour l'identification d'un locuteur (comme par exemple la composante rythmique de la prosodie [8]). Afin de contrôler les aspects liés à la prosodie et au champ lexical spécifiques à un locuteur, seules des triplés de syllabes

1. Ces abbréviations correspondent dans la littérature à, respectivement, GPR pour *glottal pulse rate* et à VTL pour *vocal tract length*.

2. Bien que l'échelle acoustique soit homogène à une longueur, il est possible de représenter une différence de LTV en termes fréquentiels et donc de la quantifier en demitons.

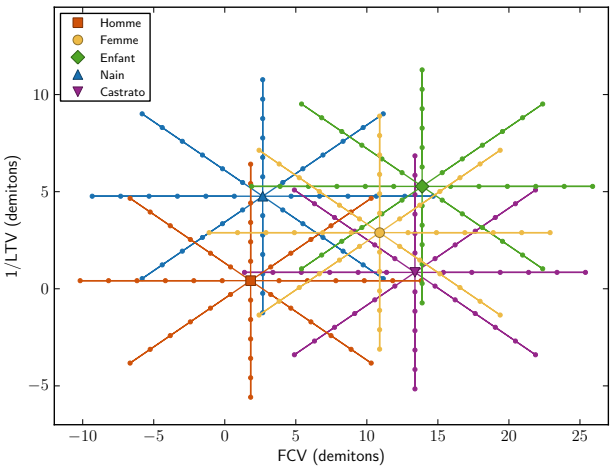

FIGURE 1 – Fréquence d'oscillation des cordes vocales (FCV) et longueur du tractus vocal (LTV) des locuteurs de référence (gros symboles) et de comparaison (points). Les axes sont en demitons relatifs à la voix originale enregistrée.

ont été utilisés comme stimuli dans cette étude. Par ailleurs, pour contrôler tous les aspects liés aux caractéristiques physiques du locuteur ainsi qu'à son accent, toutes les syllabes ont été enregistrées par un même locuteur (l'auteur RP). Les changements de FCV et LTV ont ensuite été simulés grace au vocodeur STRAIGHT [9].

Les syllabes étaient les 65 combinaisons des 13 consonnes /b, d, f, g, h, k, l, m, n, p, r, s, t/ avec les 5 voyelles /a, e, i, o, u/. Les voyelles originales avaient une durée d'environ 500 ms, mais ont été réduites à 200 ms en réduisant la partie stationnaire de ces syllabes. Le niveau de toutes ces syllabes a été ajusté pour être à environ 70 dB-SPL. Dans chaque triplé, les syllabes étaient séparées de 50 ms, et leurs FCV et LTV suivaient chacun un contour aléatoirement choisi parmi les options suivantes : montant, descendant, montant-descendant, descendant-montant. Pour chaque contour, le pas entre deux syllabes consécutives etait de 0,5 demitons, à la fois pour FCV et pour LTV.

Cinq locuteurs de référence, représentés Figure 1, ont été choisis : trois locuteurs typiques (un homme, une femme et un enfant) et deux locuteurs atypiques (un "nain" et un "castrato"). Les FCV et LTV sont exprimées relativement au locuteur original (RP) dont la FCV moyenne est d'environ 120 Hz, et la LTV est estimée à environ 16 cm. Les valeurs choisis pour les locuteurs simulés correspondent à des valeurs typiques de FCV et de LTV par rapport aux données canoniques collectées par Peterson et Barney [10]. Autour de chacun de ces 5 locuteurs de référence, 48 locuteurs de comparaison ont été créés. Ces locuteurs sont disposés sur 8 rayons d'une ellipse centrée sur le locuteur de référence et ayant un rayon de 12 demitons selon l'axe FCV et 6 demitons selon l'axe LTV. La disposition en ellipse suppose que c'est une distance euclidienne dans le plan FCV–LTV qui va piloter les résultats comme observé par Vestergaard *et al.* [4] dans une tâche de ségrégation de syllabes concurrentes. Les dimensions de l'ellipse ont été établies de façon a optimiser l'échantillonnage de l'espace de test grâce à une pré-expérience où les rayons de l'ellipse étaient de 24 demitons selon FCV et 8 demitons selon LTV.

À chaque essai, un triplé prononcé par un locuteur de comparaison est présenté avant ou après un triplé prononcé par le locuteur de référence correspondant. Cette opération est répétée 10 fois pour chacune des $5 = 240$ voix de comparaison plus 10 autres fois où le locuteur de référence est utilisé pour les deux triplés, résultants ainsi en un total de 2450 comparaisons par sujet.

2.1 Expérience 1 : locuteurs anonymes

Dans cette expérience, l'objectif était de minimiser les effets d'apprentissage liés aux locuteurs de référence. Comme ceux-ci sont présentés dans 51% des essais alors que toutes les autres voix ne représentent que 1% des essais, il y a un risque que les participants s'habituent rapidement à ces voix de référence. Il devient alors difficile de savoir si les participants comparent réellement les deux voix qu'ils entendent sur la base de leurs caractéristiques acoustiques, ou s'ils jugent à quel point la voix de comparaison leur est familière. Bien que ces deux jugements soient conceptuellement reliés, des mesures ont été prises dans cette expérience pour minimiser l'effet d'apprentissage des voix.

Dans tous les essais de cette expérience, la paire de voix sélectionnée étaient translatée aléatoirement dans le plan FCV–LTV. Le vecteur défini par les deux voix sélectionnées dans le plan FCV–LTV était donc déplacé sans être modifié. L'amplitude de la translation suivant FCV était comprise entre -0,5 et +0,5 demitons, et entre -1 et +1 demiton pour LTV. Par ailleurs, l'ordre de présentation des deux triplés était tiré aléatoirement de façon à ce que le sujet ne puisse savoir lequel de ces triplés représente la voix de référence. La question posée au sujet était : "Est-il possible que les deux sons présentés aient été prononcés par le même locuteur ?"

Dix étudiants (5 hommes et 5 femmes) de l'Université de Cambridge ont participé à cette expérience. Tous étaient normo-entendants (seuils audiométriques inférieurs à 15 dB-HL) et leur participation à cette expérience était rémunérée.

2.2 Expérience 2 : locuteurs connus

Comme la familiarité est un corrélat important de l'indentité d'un locuteur, une seconde expérience visant cette fois à maximiser les effets de familiarisation a été réalisée. Dans cette expérience, des noms ont été attribués aux 5 locuteurs de référence : James pour l'homme, Mary pour la femme, Ethan pour l'enfant, Tony pour le nain et Alessandro pour le castrato. À chaque essai, la voix de référence était présentée en premier et son nom était indiqué au sujet, de plus la question qui lui était posé était : "Est-il possible que le second son ait été prononcé par *nom* ?", où *nom* était remplacé par le nom du locuteur. Enfin l'expérience commençait cette fois par une phase d'habituation aux voix de référence où chacune de ces voix était présentée 2 fois, accompagnée une fois d'une voix de comparaison très similaire, et l'autre fois d'une voix de comparaison très différente. Contrairement à l'expérience 1, aucune translation aléatoire des

voix n'était appliquée.

Dix autres étudiants (5 hommes et 5 femmes) de l'Université de Cambridge ont participé à cette expérience. Tous étaient aussi normo-entendants (seuils audiométriques inférieurs à 15 dB-HL) et leur participation a également été rémunérée.

3 Résultats

Les résultats sont présentés Figure 2. Les données des deux expériences ont été analysés en appliquant un même modèle linéaire mixte sur lequel une analyse de variance a été effectuée. Le modèle spécifié comprenait 4 effets fixes : expérience, locuteur de référence, distance radiale au locuteur de référence (décrite plus loin) et angle dans le plan FCV–LTV. La variable sujet a été spécifiée comme effet aléatoire, de façon à autoriser des ordonnées à l'origine indépendantes pour chaque sujet et chaque locuteur de référence. Les résultats de cette analyse sont reproduits Table 1. Cette analyse a révélé que le numéro d'expérience n'avait pas d'effet significatif (et seule une des intéractions s'est révélée significative). Les résultats présentés Figure 2 correspondent donc au données moyennées sur les deux expériences.

La distance radiale est définie comme la distance euclidienne pondérée dans le plan FCV–LTV [4] :

$$d_\xi = \sqrt{\xi^2\ \Delta\mathrm{FCV}^2 + \Delta\mathrm{LTV}^2} \qquad (1)$$

ΔFCV et ΔLTV sont les différences en demitons, et ξ est un facteur d'équivalence entre FCV et LTV. Les données de la Figure 2 sont tracées avec $\xi = 1$. Cependant en supposant une symétrie radiale des données, une valeur optimale de ce paramètre peut-être évaluée en ajustant un autre modèle : la courbe de réponse est définie comme une fonction Γ-cumulée qui ne dépend que la distance radiale d_ξ, comme montré dans la Figure 3. Alors que Vestergaard *et al.* [4] avaient ainsi trouvé une valeur de 1,6 pour ξ, les données collectées dans la présente expérience conduisent à une valeur proche de son inverse : $\xi = 0{,}6$. Ceci signifie qu'en moyenne, dans l'étude de Vestergaard *et al.* , une différence de 1 demiton en

Locuteur	$F(4, 72) = 19{,}02$	$p < 0{,}0001$
Expérience	$F(1, 18) = 0{,}225$	$p = 0{,}64$
Angle	$F(1, 4670) = 34{,}16$	$p < 0{,}0001$
Distance	$F(1, 4670) = 4662$	$p < 0{,}0001$
Loc. × Expé.	$F(4, 72) = 2{,}175$	$p = 0{,}08$
Loc. × Angle	$F(4, 4670) = 14{,}27$	$p < 0{,}0001$
Expé. × Angle	$F(1, 4670) = 0{,}542$	$p = 0{,}46$
Loc. × Dist.	$F(4, 4670) = 1{,}510$	$p = 0{,}20$
Expé. × Dist.	$F(1, 4670) = 0{,}459$	$p = 0{,}50$
Angle × Dist.	$F(1, 4670) = 12{,}02$	$p < 0{,}001$
L. × E. × A.	$F(4, 4670) = 3{,}002$	$p < 0{,}05$
L. × E. × D.	$F(4, 4670) = 1{,}463$	$p = 0{,}21$
L. × A. × D.	$F(4, 4670) = 6{,}504$	$p < 0{,}0001$
E. × A. × D.	$F(1, 4670) = 2{,}019$	$p = 0{,}16$
L. × E. × A. × D.	$F(4, 4670) = 0{,}721$	$p = 0{,}58$

TABLE 1 – Analyse de la variance du modèle linéaire mixte. Les noms des facteurs sont abbréviés par leur première lettre quand nécessaire.

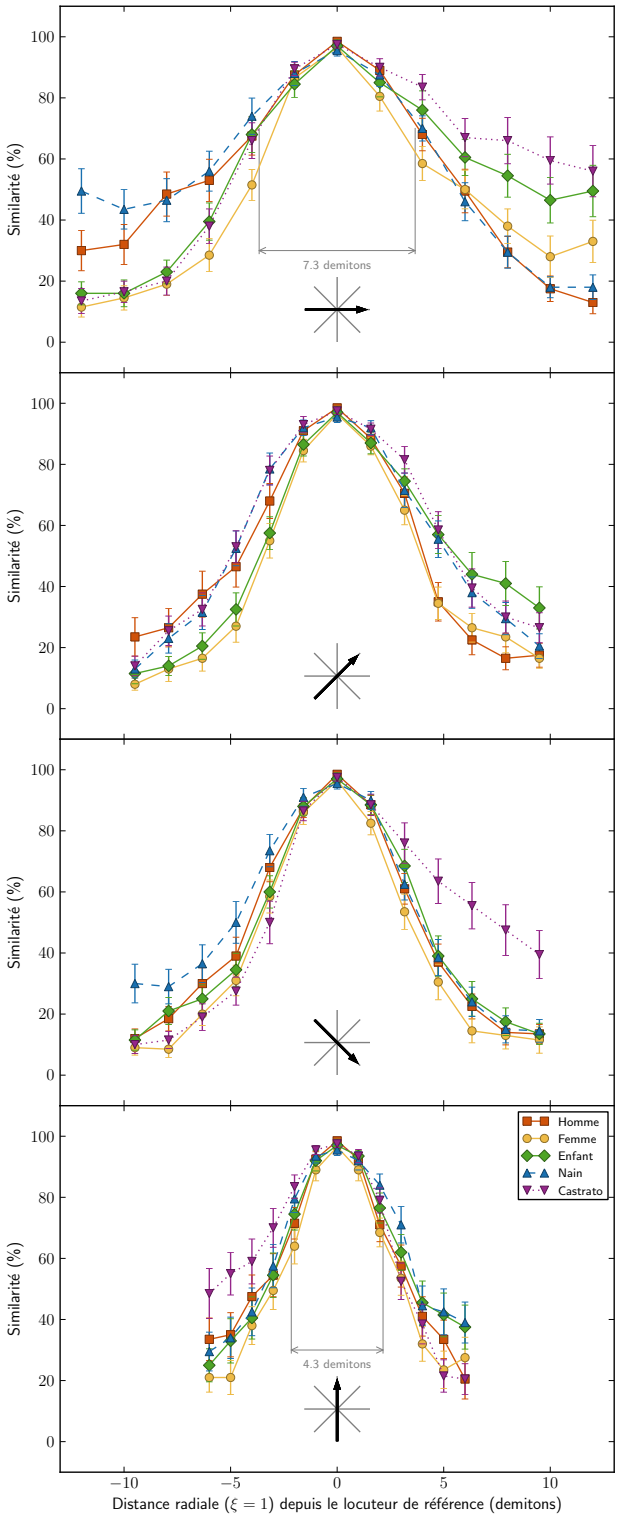

FIGURE 2 – Jugement de similarité (%) en fonction de la distance radiale ($\xi = 1$) depuis le locuteur (voir texte). Chaque couleur et symbole représente un locuteur de référence. Chaque cadre correspond à un diamètre de l'ellipse circonscrite aux locuteurs de comparaison, c'est-à-dire à une direction dans le plan FCV–LTV comme indiqué par les flèches noires. Les indications de largeur de distribution en gris correspondent à deux écarts-types.

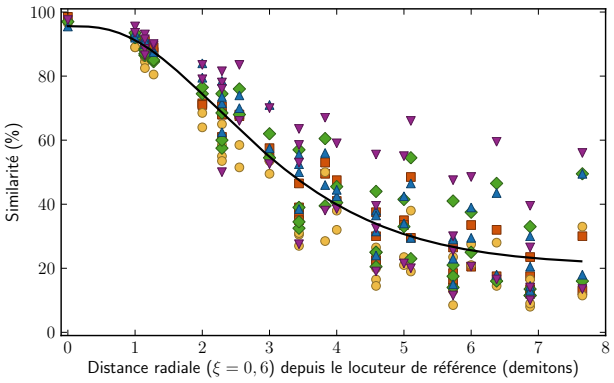

FIGURE 3 – Jugement de similarité (%) en fonction de la distance radiale depuis le locuteur pour la valeur optimale de ξ. Comme dans la Figure 2, chaque couleur et symbole représente un locuteur de référence. La ligne continue noire représente le modèle ajusté aux données.

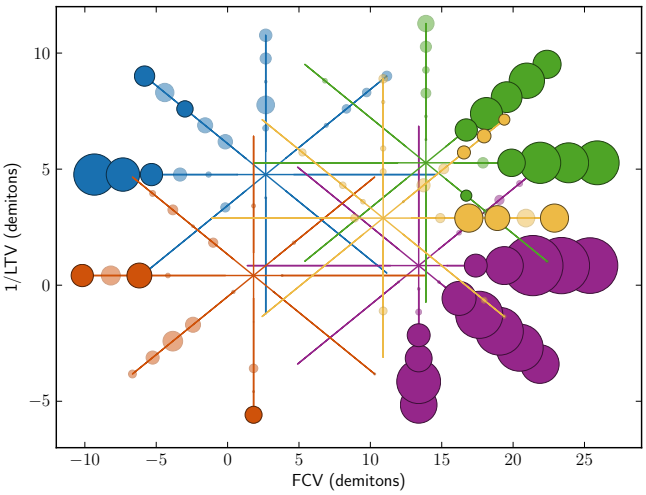

FIGURE 4 – Assymétrie du jugement de similarité par rapport au rayon opposé pour chaque locuteur de référence. Les couleurs sont les mêmes que dans la Figure 1. Seules les assymétries positives (plus similaire) sont reproduites. L'aire des points est proportionnelle à la différence de similarité. Les points opaques représentent les différences significatives (t-test), tandis que les points transparents représentent les différences non significatives.

FCV était équivalente à une différence de 1,6 demiton en LTV. Ici, les sujets toléraient de façon identique un changement de 1 demiton en FCV qu'un changement de 0,6 demitons en LTV.

Le résultat principal de cette expérience est que la largeur de la distribution des jugements de similarité suivant l'axe FCV (7,3 demitons) est plus importante que celle mesurée suivant l'axe LTV (4,3 demitons). Cette différence coïncide avec un facteur d'équivalence ξ de 0,6. Ce résultat indique que les sujets étaient plus tolérants envers un changement de hauteur de la voix, qu'envers un changement de taille du locuteur, comme prédit par nos hypothèses. Par ailleurs, il peut être montré que les résultats le long des diagonales (quand il y a une différence selon FCV et LTV) peuvent être simplement prédits en ajoutant les contributions dues à chaque axe.

L'usage d'un facteur d'équivalence optimal entre FCV et LTV est destiné à contrôler l'effet de l'angle. Néanmoins cet effet reste significatif dans l'analyse rapportée en Table 1. Cet effet interagit significativement avec le locuteur et la distance radiale. Il apparaît en effet que toutes les courbes de similarité sont légèrement assymétriques. Il apparaît en fait que cette assymétrie est localisée à la périphérie du plan FCV–LTV exploré, et en particulier pour les voix ayant une hauteur fondamentale plus élevée, où le jugement de similarité se rapproche de 50% (Figure 4).

4 Discussion

Cette étude permet de quantifier la représentation interne qu'un auditeur se fait d'un locuteur. Des variations de hauteur fondamentale de ±3,6 demitons, et des variations d'échelle acoustique de ±2,2 demitons sont tolérées sans que l'identité du locuteur ne soit altérée. La représentation interne de l'identité d'un locuteur peut donc se résumer, comme illustré Figure 5, à une ellipse de 7,3 demitons suivant FCV et de 4,3 demitons suivant LTV. Ces paramètres pourront être utilisés, dans un contexte expérimental, pour générer des stimuli pour lesquels le variations de FCV et LTV sont réalistes pour un auditeur. Les résultats de cette étude pourront aussi être utilisés pour faciliter l'identification automatique du locuteur dans le contexte du traitement automatique de la parole.

Cette représentation du locuteur semble relativement robuste dans tout le plan FCV–LTV : la largeur des distributions varie peu avec le locuteur. En particulier, ces variations sont concentrées à la périphérie du plan et quatre explications peuvent être considérées. Une première explication tentante est que ces voix sont moins communes que les voix situées à l'intérieur du plan. Les limites du plan utilisé dans cette étude sont en effet définies à partir d'estimations de la répartition des valeurs de FCV et LTV pour la population [10]. Cependant, les deux locuteurs atypiques simulés dans l'étude ne donnent pas lieu à des distributions plus larges, du moins sur les rayons qui pointent vers l'intérieur du plan. La seconde explication serait que les participants ne se contentent pas simplement de juger si les deux voix présentées dans un essai peuvent provenir du même locuteur, mais s'aident dans leur jugement en catégorisant les voix : si une voix ressemble à celle de l'homme et l'autre à celle de la femme, alors il est peu probable que les deux stimuli proviennent du même locuteur ; en revanche si une voix ressemble à la voix d'homme et que la seconde est aussi une voix d'homme, mais beaucoup plus grave, alors le sujet pourra être plus tolérant pour ce type de changement puisqu'il n'y à pas ici de saut de catégorie. Cependant, alors que l'objectif de l'expérience 1 était de limiter cet effet, celui de l'expérience 2 était de le favoriser. Si l'interférence avec une stratégie générale de catégorisation était véritablement responsable de ces assymétries, le numéro d'expérience devrait avoir un effet significatif, ce qui n'est pas

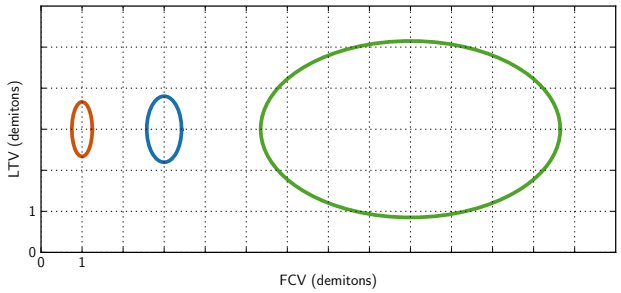

FIGURE 5 – Représentation, dans le plan FCV–LTV, de la plus petite différence détectable[3] (à gauche, en orange, d'après [5, 6]), de la différence nécessaire pour séparer 2 syllabes simultanées (au mileu, en bleu, d'après [4]) et de la différence tolérée sans altérer l'identité du locuteur (à droite, en vert).

le cas. La troisième explication concerne les limites de la technique utilisée pour manipuler les voix des locuteurs. Le mécanisme de synthèse de STRAIGHT est extrêmement robuste et à été utilisé avec succès pour explorer des domaines du plan FCV–LTV bien plus larges [5]. Néanmoins il existe une limite physique à cette manipulation : si la fréquence fondamentale est plus élevée que le premier formant, alors celui-ci ne recevra aucune énergie et disparaîtra du signal. Ce problème intervient donc dans le coin inférieur droit du plan FCV–LTV tel que présenté Figure 1, et concerne donc le locuteur Castrato. L'effet de sous-échantillonnage est clairement visible sur le troisième cadre de la Figure 2 où le jugement de similarité pour le Castrato (triangles violets) se démarque clairement de celui des autres locuteurs. Ce sous-échantillonnage rend la compréhension de certaines voyelles plus difficiles, mais peut aussi rendre la perception de l'échelle acoustique moins fiable. Cet effet peut donc expliquer la forte assymétrie des réstultats obtenus pour le Castrato, mais ne peut expliquer les assymétries observées tout autour du plan. Une dernière explication s'appuie sur la constatation que cette assymétrie est essentiellement due aux sujets féminins. Les participantes pourraient en effet utiliser une stratégie différente de celles des sujets masculins. Il est en effet optimal pour une femme de catégoriser les voix perçues par rapport à une voix de femme, ou leur propre voix : de part sa position centrale dans le plan FCV–LTV, la voix du locuteur Femme, comme probablement celle de la participante, est idéalement située au centre du plan. Toutes les voix adjacentes appartiennent donc à une autre catégorie. Au contraire, pour les participants masculins, leur voix se situant dans un coin du plan, catégoriser les voix adjacentes n'est pas suffisant pour distinguer Femme et Enfant. Il n'est cependant pas possible, ici, de conclure avec certitude sur les origines de cet effet, et des études complémentaires sont nécessaire pour vérifier cette hypothèse.

Comme mentionné en introduction, la largeur des distributions observées ne reflète pas la sensibilité du système auditif pour chacune de ces manipulations (FCV et LTV). Afin de mieux comparer ces différences, le seuil de détection a été reproduit sous forme d'ellipse dans le plan FCV–LTV aux côtés de la largeur des distributions obtenues par Vestergaard *et al.* [4] pour la ségrégation de syllabes simultanées et de la représentation interne d'un locuteur telle que mesurée dans la présente étude (Figure 5). Il apparaît ainsi clairement que les seuils de détection ne peuvent expliquer la forme de l'ellipse représentant l'identité d'un locuteur. En revanche, la largeur de cette dernière distribution selon l'axe FCV correspond à la variation naturelle de de fréquence fondamentale dans la voix humaine. Kania *et al.* [11] ont mesuré les variations de FCV pour 15 locuteurs, masculins et féminins, lisant un texte pendant plusieurs minutes. Ils ont ainsi obtenu une distribution d'une largeur de 7,8 demitons comparable à la présente mesure de 7,3 demitons. Ceci suggère que les distributions mesurées ici ont été établies par apprentissage : en étant exposé à des locuteurs tout au long de leur développement, nous construirions ces distributions qui nous permette d'établir la probabilité que deux sons proviennent du même locuteur. Cette hypothèse pourrait être vérifiée en répétant cette étude chez l'enfant.

Enfin, alors qu'une différence de FCV (en demitons) était plus efficace pour aider à séparer deux syllabes simultanées que la même différence en LTV, la relation entre FCV et LVT est inversée pour l'identification d'un locuteur (Figure 5). Cette rotation de l'ellipse entre la présente étude et les précédentes pourrait refléter une différence de niveau d'intégration, et en particulier, des intégrations à des échelles temporelles différentes. Dans le paradigme de ségrégation de syllabes concurrentes, n'importe quelle différence instantanée entre les deux locuteurs peut être utilisée pour effectuer la tâche. Au contraire, pour juger de la similarité de deux locuteurs, les échantillons de voix collectés par l'auditeur sont comparés à des distributions de variation *à long terme* de ces variables. Cette étude suggère donc que la sensibilité relative suivant chacune de ces dimensions pourrait être un indicateur du niveau de traitement. Des études d'imagerie permettront de distinguer et localiser les structures impliquées dans chacune de ces tâches.

3. Alors que les valeurs données en Introduction correspondent à un d' de 1, les valeurs utilisées ici correspondent à 1 écart-type de la distribution de détection de façon à permettre la comparaison avec la présente étude.

Remerciements

Les auteurs souhaitent remercier Su Li et Vin Shen Ban pour leur aide dans la réalisation de ces expériences. Cette étude a été financée par deux bourses du UK Medical Research Council (G9900369 et G0500221).

Références

[1] Smith D.R.R., Patterson R.D., "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age", *J. Acoust. Soc. Am.* 118, 3177–3186 (2005).

[2] Smith D.R.R., Walters T.C., Patterson R.D., "Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled", *J. Acoust. Soc. Am.* 122, 3628–3639 (2007).

[3] Darwin C.J., Brungart D.S., Simpson B.D., "Effects of fundamental frequency and vocal-tract length

changes on attention to one of two simultaneous talkers", *J. Acoust. Soc. Am.* 114, 2913–2922 (2003).

[4] Vestergaard M.D., Fyson N.R.C., Patterson R.D., "The interaction of vocal tract length and glottal pulse rate in the recognition of concurrent syllables", *J. Acoust. Soc. Am.* 125, 1114–1124 (2009).

[5] Smith D.R.R., Patterson R.D., Turner R.E., Kawahara H., Irino T., "The processing and perception of size information in speech sounds", *J. Acoust. Soc. Am.* 117, 305–318 (2005).

[6] Ives D.T., Smith D.R.R., Patterson R.D., "Discrimination of speaker size from syllable phrases", *J. Acoust. Soc. Am.* 118, 3186–3822 (2005).

[7] Honorof D.N., Whalen D.H., "Perception of pitch location within a speaker's f0 range.", *J. Acoust. Soc. Am.* 117, 2193–2200 (2005).

[8] Remez R.E., Fellowes J.M., Nagel D.S., "On the perception of similarity among talkers.", *J. Acoust. Soc. Am.* 122, 3688–3696 (2007).

[9] Kawahara H., Irino T., "Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation", in *Speech separation by humans and machines*, édité par Divenyi P.L., 167–180 (Kluwer Academic, Massachusetts) (2004).

[10] Peterson G.E., Barney H.L., "Control methods used in a study of the vowels", *J. Acoust. Soc. Am.* 24, 175–184 (1952).

[11] Kania R.E., Hartl D.M., Hans S., Maeda S., Vaissiere J., Brasnu D.F., "Fundamental frequency histograms measured by electroglottography during speech : a pilot study for standardization", *J. Voice* 20, 18–24 (2006).