AUTOMATISATION DE LA DÉTECTION DE L'INTOXICATION À L'ALCOOL DANS LA PAROLE[*]

Alice Breton, Ghyslain Cantin-Savoie
Université du Québec à Montréal

1. Introduction et cadre théorique

L'intoxication à l'alcool a un impact visible et audible sur le comportement humain, notamment sur la production langagière. Une personne a d'ailleurs la capacité de reconnaître si un autre individu est sobre ou intoxiqué correctement à 74% du temps, sans connaître préalablement la personne à son état sobre (Hollien et al. 2009). Ainsi, on peut tenter d'évaluer les différences acoustiques et articulatoires de la parole dans le but de différencier un signal sonore produit par une personne sobre d'un signal produit par une personne intoxiquée. Pour ce faire, certains spécialistes et certaines spécialistes se sont penchés sur l'analyse des effets de l'alcool au niveau de la production et aux signaux sonores.

Dans un corpus contenant trois personnes, il a été montré que les phonèmes /l/, /r/, /s/, /ʃ/ et /ts/ sont les plus affectés par la consommation à l'alcool (Trojan et Kryspin-Exner 1968, cité dans Pisoni et Martin 1989)Par la suite, un autre petit corpus a été analysé et les auteurs ont observé un allongement consonantique lors de syllabe non-accentuée, un dévoisement des occlusives en dernier mot d'une phrase, un changement du lieu d'articulation du /s/ et une désaffrication des t͡ʃ et d͡ʒ (Lester et Skousen 1974, cité dans Pisoni et Martin 1989). De plus, dans un corpus de quatre étudiants, en utilisant l'avis professionnel de deux phonéticiens, il a été jugé que la consommation de l'alcool a un effet sur la durée des voyelles ainsi que l'articulation partielle ou la suppression des consonnes /l/ et /r/, des consonnes nasales /n/ et /ŋ/ et de de la consonne dentale /d/. Ensuite, la désafrication du /ts/ a été observé par ces deux phonéticiens (Pisoni et Martin 1989). Aussi, dans un corpus composé d'enregistrements d'hommes intoxiqués à plusieurs niveaux d'alcoolémie, le nombre d'erreurs de prononciation augmentait significativement en corrélation avec le taux d'alcool (Kunzel et al. 1992, cité dans Hollien *et al.* 2001). On peut constater que ces analyses sont basées sur des petits corpus composés intégralement d'hommes, ce qui peut impliquer la présence de biais méthodologiques. L'observation d'un corpus de 162 personnes allemandes, le Alcohol Language Corpus (ALC), composé d'une quantité équilibrée de femmes et d'hommes, montre que le nombre et la durée des hésitations, le nombre d'erreurs de prononciation ainsi que la quantité de répétitions augmentent aussi avec la consommation d'alcool (Barfüßer et Schiel 2010).

Les effets de l'alcool sur la prononciation sont aussi perceptibles dans le signal acoustique, ce qui a été le sujet d'analyse de plusieurs chercheuses et chercheurs. Tout d'abord, la fréquence fondamentale (F0) est une des variables phonétiques affectées par

[*] Nous voulons remercier Denis Foucambert, Lucie Ménard, Thomas Leu, ainsi que tous nos pairs du cours Projet de recherche.

Actes du congrès annuel de l'Association canadienne de linguistique 2020.
Proceedings of the 2020 annual conference of the Canadian Linguistic Association.

l'effet de l'intoxication à l'alcool dans la parole. Il est important de mentionner que la F0 est la vibration sinusoïdale la plus lente qui est produite par le nombre d'ouvertures des cordes vocales, c'est par celle-ci que nous pouvons identifier un son aigu ou un son grave. Pour en revenir à l'effet de l'intoxication à l'alcool sur la F0, les études se contredisent ; certaines constatent une augmentation de celle-ci (Hollien *et al.* 2001, Klingholz *et al.*, 1988), d'autres montrent une baisse (Alderman *et al.* 1995, Watanabe *et al.* 1994). Enfin certaines études montrent qu'il n'y a aucun changement de la F0 (Sobell *et al.* 1982). Les conclusions divergentes que mettent en lumière ces études sur l'effet de l'intoxication à l'alcool sur la F0 peut être expliquée en observant différents niveaux d'intoxications (Braun et Kunzel 2003). En effet, l'impact de l'intoxication à l'alcool est plus grand sur la F0 des individus ayant un taux d'alcool dans le sang de 0,05% à 0,1%. Lorsque les individus ont un taux d'alcoolémie supérieure à 0.1%, les effets de l'intoxication ne paraissent plus autant dans l'observation de la F0. Par ailleurs, ces expériences ont été conduites sur des petits corpus, contenant entre 4 et 35 participants, et qui étaient souvent constitués exclusivement d'hommes. Un corpus avec une plus grande quantité de personnes, et constitué d'un nombre égal d'hommes et de femmes a justement été produit entre 2007 et 2010 pour, entre autres, étudier les effets de l'alcool sur la fréquence fondamentale.

Plus récemment, de multiples études ont été exécutées sur le Alcohol Language Corpus. Comme c'est celui que nous utilisons, il sera décrit en détail dans la section "Corpus" ci-dessous. L'une d'entre elles montre que 79,1% du temps la F0 des locuteurs augmentait (Baumeister *et al.* 2012). Par ailleurs, certaines études se sont penchées sur l'effet de l'intoxication à l'alcool sur les valeurs formantiques. Ces dernières sont représentées en Hertz et, selon leurs valeurs, elles montrent le dégré d'ouverture de la mâchoire (F1), la position de la langue (F2), la position des lèvres (F3) et la position du voile du palais (F4). Dans une étude de ces différences formantiques F1, F2, F3 et F4, il a été démontré qu'il y a une hausse de la médiane de la F1 seulement pour les voyelles de l'allemand [a:], [i:] et [u:] et que la valeur de F4 prise au quatrième quartile augmente pour l'ensemble des voyelles (Schiel *et al.* 2010). En se concentrant sur la facette prosodique de l'étude de la parole, une autre étude a montré que la durée des segments augmente lorsqu'ils sont produits par une personne intoxiquée à l'alcool (Schiel *et al.* 2010). Par ailleurs, une augmentation de la distance entre les contours rythmiques de la parole pour les personnes intoxiquées a été observée sur un échantillon d'enregistrements de 20 participants et participantes au ALC (Heinrich et Schiel 2014). En d'autres mots, ce sont des groupes rythmiques de la parole qui sont souvent représentés par les virgules dans une phrase écrite, donc la durée de ces segments est plus longue. Par ailleurs, une faible hausse de l'intensité (du volume) a aussi été corrélée avec l'intoxication à l'alcool (Fairbairn *et al.* 2015). En somme, on peut constater une hausse de la F0, une hausse de la F1 pour certaines voyelles, une hausse de la F4 prise au quatrième quartile pour l'ensemble des voyelles, et une augmentation du volume, de la durée des segments et de la distance des contours rythmiques. Nous avons décidé de prendre en compte une partie de ces variables, ainsi que des variables spectrales telles que l'écart type, le centre de gravité, l'aplatissement (*Kurtosis*) et l'asymétrie (*Skewness*) qui sont liées aux caractéristiques de la consonne. Une description de ces dernières est incluse ci-dessous dans la section Analyse.

Les différences acoustiques entre les productions langagières d'une personne sobre et d'une personne intoxiquée peuvent être utilisées pour automatiser la détection de l'intoxication à l'alcool dans la parole. Depuis la création du ALC, on peut constater que certaines études ont tenté d'établir un algorithme de classification dont le taux de bonnes réponses serait supérieur à 74% qui est le taux de bonne réponse observé chez des individus (Hollien *et al.* 2009). En extrayant les valeurs de huit paramètres acoustiques à chaque période de 10 millisecondes, sur une segmentation par tranches de 20 millisecondes du signal acoustique, le taux de bonnes réponses final est de 71.4% (Bone *et al.* 2014). Ce type de segmentation ne considère ni les différences entre une voyelle et une consonne ni les différences entre plusieurs types de pauses (vides, remplies, etc.). Bref, une telle segmentation par la durée ne tient pas compte des propriétés propres aux phonèmes et aux unités prosodiques de la parole. Par ailleurs, une étude a utilisé une autre unité de segmentation, nommée "unités phrastiques", dans le but de permettre une détection automatique de l'intoxication à l'alcool dans la parole (Levit *et al.* 2001). Celles-ci sont des segments séparés par les creux prosodiques, c'est-à-dire les endroits où, pour un bref instant, il y a une pause vide. Ces unités correspondent à 85% à des points, des virgules ou des espaces. Elles font directement référence aux contours rythmiques mentionnés précédemment. Après avoir été comparées à une segmentation à chaque 20 millisecondes, l'utilisation de cette "unité phrastique" comme type de segmentation a permis un taux de bons résultats à 69% pour la détection automatique de l'intoxication à l'alcool (Levit *et al.* 2001). Ce taux de bonnes classifications n'est pourtant pas plus élevé que celui de 71,4%, qui avait été obtenu avec l'utilisation d'un corpus segmenté à chaque 20 millisecondes (Bone *et al.* 2014). Il se peut que ce 69% soit dû à des biais méthodologiques dans la création du corpus, qui est constitué de 33 hommes qui répètent plusieurs fois la même tâche dans différents états d'intoxications. Cependant, à l'instar des phonèmes et contrairement aux séparations par millisecondes, les unités phrastiques sont justifiées par des critères linguistiques. Ces unités émergeant d'informations prosodiques et acoustiques, elles seront appelées Unités prosodiques pour le restant de cet article.

Il est crucial de déterminer les données d'entrées d'un système de classification automatique. Par exemple, si on bâtit un système d'automatisation de la détection de l'intoxication à l'alcool dans les mots fonctionnels de locuteurs de l'allemand, cela implique une séparation du signal en mots de l'allemand et une catégorisation des mots fonctionnels. Chacune de ces modifications du signal sonore restreint l'utilisation de l'algorithme aux mots fonctionnels de l'allemand, ce qui empêche l'utilisation ultérieure du même programme sur d'autres types de mots ou dans une autre langue. Une automatisation serait optimale dans la variabilité de son utilisation en serait donc une qui prédit l'intoxication en effectuant seulement les modifications préalables minimales au signal sonore d'entrée. On peut par exemple séparer le signal à chaque pause prosodique, par phonèmes, ou par syllabes. Par contre, si on décide de segmenter par mots, ou par de plus petites unités porteuses de sens; par morphèmes, cela implique les étapes additionnelles de choisir la langue de l'analyse, en catégoriser les mots et en segmentant les morphèmes. Ces segmentations restreignent donc l'éventail de langues sur lesquelles le programme de détection serait efficace.

Notre recherche vise à déterminer la segmentation préférable pour une automatisation de la détection de l'intoxication à l'alcool dans la parole qui n'implique pas de sélection préalable d'une langue en particulier. Pour ce faire, nous avons comparé deux segmentations du signal sonore une par phonèmes et une par unités prosodiques, dans le cadre d'une détection automatique de l'intoxication à l'alcool dans la parole en se posant la question de recherche suivante : « Quelle segmentation, entre une découpe par phonèmes et par unité prosodique, permet une meilleure performance en matière de détection automatique de l'intoxication à l'alcool dans la parole? »

Notre hypothèse est que la segmentation par phonèmes permettra une meilleure détection automatique de l'intoxication à l'alcool dans la parole. Puisque la plupart des variables phonétiques (la F1, la F2, le centre de gravité) sont plus appropriées pour les phonèmes, cette qualité de l'information permettra un meilleur classement. De plus, ce type de segmentation entraîne les plus petites unités distinctives, ainsi qu'une plus grande quantité de segments que la segmentation par unités prosodiques. Bref, nous croyons qu'en combinant une plus grande qualité et quantité de données, nous pourrons parvenir à une meilleure reconnaissance de l'intoxication à l'alcool, et qu'en conséquence la segmentation optimale en sera un par phonèmes.

2. Méthode

2.1 Corpus

Cette partie de l'article montre une brève description du corpus utilisé pour notre recherche. Réalisé en coopération entre l'*Institut of Legal Medicine of Munich* et le *Bund gegen Alkohol und Drogen im Strassenverkehr* (BADS), le ALC est le premier grand corpus accessible à la communauté scientifique constitué d'une collection d'enregistrements de 162 locuteurs et locutrices de l'allemand, dont 85 hommes et 77 femmes âgées de 21 à 64 ans, une fois sous l'effet de l'alcool, et une autre fois sobre. Ces personnes sont généralement des étudiantes et étudiants en droit, des juges, des policiers et policières ainsi que des travailleurs et travailleuses du domaine public. La procédure de l'expérience est divisée en deux parties. La première consistait à boire du vin ou de la bière jusqu'à ce que le locuteur ou la locutrice atteigne le niveau d'intoxication avec lequel il ou elle était confortable. Ensuite, une prise de sang ainsi qu'une mesure par le souffle étaient effectuées. Le taux d'alcool était généralement entre 0,05% et 0,25%. Suivant l'obtention du résultat par personne, l'enregistrement commençait. Les tâches étaient d'une durée totale de 15 minutes pour éviter l'évacuation de l'alcool de l'organisme. Celles-ci étaient enregistrées dans une voiture pour réduire la variabilité de l'arrière-plan acoustique et dans le but de favoriser l'utilisation d'outils dérivés de l'ALC dans les voitures. Chaque participant et participante intoxiquée devait lire sur un écran des phrases, des chiffres et des commandes vocales variées. De plus, ces tâches étaient aussi composées d'une production de la parole spontanée: le locuteur ou la locutrice devait répondre à cinq questions posées par le chercheur. En tout, il y a un total de 30 tâches. Les phrases élicitées étaient des commandes vocales souvent utilisées dans les voitures dotées d'un système de reconnaissance vocale et elles étaient inspirées par les recherches sur les différences de prononciations sous l'effet

de l'intoxication à l'alcool en allemand. Enfin, deux semaines suivant cette première partie d'enregistrement, les participants refaisaient les mêmes tâches, mais cette fois contenant deux fois plus d'items que la première (pour un total de 60 items) et dans l'état sobre. Pour contrôler et s'assurer que la différence de longueur de tâche n'affecte pas les résultats, 20 participants ont refait la séance de 30 items dans un état sobre. Toutes les tâches sont enregistrées simultanément sur deux micros ; un premier déjà intégré à la voiture et un autre provenant d'un casque d'écoute. Ce corpus contient 30 360 enregistrements, représentant plus de 100h de productions langagières (Schiel *et al.* 2012, Schiel et Heinrich 2009).

Ce corpus a été étiqueté avec plusieurs informations à propos des locuteurs. On peut retrouver des informations sur la ville d'origine du locuteur ou de la locutrice, la marque de la voiture utilisée pour l'enregistrement et la condition émotive du locuteur ou de la locutrice allant d'heureuse, fatiguée à dépressive. On peut aussi retrouver leur état suivant l'enregistrement. Les habitudes de consommation d'alcool et de tabac ont aussi été considérées dans l'étiquetage du corpus. En plus d'avoir ces informations, le tout est segmenté phonétiquement à l'aide du partiteur automatique du *Bavarian Archive for Speech Signals* BAS). Suivant cette segmentation automatique, un travail manuel a été effectué pour s'assurer de sa qualité, mais aussi pour identifier les hésitations, les erreurs d'articulation et les pauses vides (Schiel et Heinrich 2009). En tout, on compte environ 30 000 fichiers (audio) *Wave* segmentés en étant accompagnés de leur *Textgrids*, ceux-ci contiennent environ 2 600 000 phonèmes. Les fichiers *Textgrids* sont la découpe et la transcription phonétique de chaque son (phonème). C'est par ces fichiers qu'une association entre le fichier audio en *Wave* et son fichier *Textgrid* qu'est la segmentation phonémique. Au sein de cette segmentation, il se trouvait des segments nommés <p>, qui représentaient des pauses, nous les avons utilisés pour créer des unités prosodiques, nous avons ciblé ces segments "pauses" et nous les avons utilisés comme bornes pour notre deuxième segmentation, appelée "Unités prosodiques", ou UP, résultant en environ 180 000 fichiers *Wave* distincts contenant chacun une unité prosodique. En d'autres mots, tout ce qui se trouvait entre ces pauses a été considéré comme la segmentation par UP. Nous avions donc à notre disposition deux ensembles contenant les segmentations à l'étude ; un déjà effectué contenant un total de 2 600 000 phonèmes et un autre délimité par les pauses comptant 180 000 unités prosodiques. De ces deux ensembles, des variables ont été sélectionnées pour extraire les données du signal acoustique. La section qui suit définit les variables que nous avons utilisées pour extraire des données du signal acoustique.

2.2 Explication des variables phonétiques

Lors de la phonation, il y a transmission d'une onde sonore produite par les cordes vocales et par la suite modulée par les différents articulateurs comme le voile du palais, la langue et les lèvres. Ces ondes sont périodiques et complexes, c'est-à-dire elles ont un patron régulier et elles sont composées par la combinaison de plusieurs ondes simples pour enfin être complexe. Elles sont représentées par l'amplitude (en décibel), la période ou la durée d'un cycle (en secondes ou millisecondes) et la fréquence qui est le nombre de cycles par seconde (en Hertz). Suivant cette brève description de l'onde sonore produite lors de la

production de la parole, il se trouve plusieurs variables phonétiques qui peuvent être utilisées pour décrire les patrons d'une onde sonore produite par une personne. Pour atteindre l'objectif de notre recherche, nous avons utilisé les mêmes 13 variables phonétiques pour les deux segmentations. Nous avons décidé d'utiliser la fréquence fondamentale (F0) qui peut, par exemple, identifier si un son est aigu ou grave. Cette fréquence dépend de la masse et de la vibration des cordes vocales. Par exemple, plus les cordes vocales sont longues et épaisses, souvent chez les hommes, plus la vibration sera lente pour une fréquence plus basse (son grave). Par la suite, nous avons utilisé le premier et le deuxième formant (F1 & F2). Celles-ci sont les harmoniques. Le premier harmonique est appelé F1, le deuxième est appelé F2, et ainsi de suite. Elles sont produites par les articulateurs qui filtrent l'onde sonore émise par les cordes vocales. Le F1 est modulé par le degré d'ouverture de la mâchoire et le F2 par la position de la langue dans la cavité buccale. Ces variables phonétiques sont souvent utilisées pour décrire une voyelle. Par exemple, un /a/ aura un F1 haut puisque la mâchoire est très ouverte et un F2 bas puisque la langue sera positionnée à l'arrière dans la cavité buccale. Par ailleurs, il arrive cependant qu'une représentation spectrale d'une onde sonore ne contienne pas d'harmoniques ; c'est ce qu'on appelle des ondes apériodiques. Ces dernières représentent plus particulièrement les consonnes. Il n'est pas possible d'en extraire des valeurs par l'unité de mesure Hertz. Cependant, les variables appropriées aux consonnes sont la répartition du spectre sur le *partiel*. Le centre de gravité et son écart type du *partiel* correspondent au lieu d'articulation de la consonne, alors que le *Kurtosis* et le *Skewness* correspondent à la tension globale du conduit vocal (Thibeault 2011). De plus, il se trouve d'autres variables phonétiques comme l'intensité qui représente un son faible ou fort et enfin la durée d'un son.

2.3 Extraction des données linguistiques pour nos variables indépendantes (VI)

Notre recherche étant une comparaison, chaque variable choisie a été utilisée pour l'extraction des données linguistiques sur les deux ensembles de fichiers distincts. Puisque la littérature précédente indiquait des différences entre les individus sobres et intoxiqués dans la F0, la F1, le volume (Intensité) et la durée des segments, nous avons extrait les données de ces quatre variables à l'aide d'un script *Praat* (Boersma et Weenink 2009). Le volume a été extrait et est mesuré en décibel (dB), où le 0 dB représente le seuil d'audibilité. Pour une meilleure représentation des voyelles, nous avons extrait la F1 ainsi que la F2. La valeur moyenne des deux formants a été prise au premier quart, au centre, puis au troisième quart de chaque segment, résultant ainsi en 6 variables décrivant les valeurs des variables F1 et F2. De plus, nous avons extrait les valeurs de variables spectrales appropriées pour les consonnes comme le centre de gravité et son écart type, ainsi que le *Kurtosis* et le *Skewness*. Les variables spectrales des consonnes que nous avons prélevées le sont habituellement, en phonétique, au centre du segment. Puisqu'il y avait encore les <p> et que nous ne voulions pas que notre modèle les prenne en considération, par exemple, la F1 d'une pause, nous avons décidé de retirer les segments <p> de la segmentation par phonèmes. L'ensemble d'unités prosodiques, segments définis par ce qui se trouve entre les pauses, ne nécessitait pas cette modification.

3. Analyse

3.1 La variable dépendante (VD)

Notre classification est faite en fonction de la présence ou de l'absence d'alcool dans le sang du locuteur ou de la locutrice enregistrée. Le corpus contient 2 fois plus d'enregistrements de personnes sobres que d'enregistrement de personnes intoxiquées. L'idée d'utiliser 0.08%, qui est la limite légale du taux d'alcool dans le sang pour conduire un automobile au Canada (SAAQ 2015), comme limite à partir de laquelle nous rendons notre variable continue en variable binaire, car elle a été sélectionnée dans de précédentes études (Levit *et al.* 2001). Cependant, cela entraînait une répartition du corpus de 81% en bas de 0.08% pour 19% au-dessus de 0.08%. L'inégalité de cette répartition étant une cause probable de l'apparition d'*overfitting*, discutée plus bas dans la section Discussion, nous avons procédé à un nettoyage du corpus. Le corpus final nettoyé contenait un segment produit par un locuteur ou une locutrice sobre pour chaque segment produit par un locuteur ou une locutrice intoxiquée pour égaliser les proportions du corpus. Certains retraits de données erronées dues à des erreurs de microphone nous obligeant à retirer du corpus un nombre inégal de données, la quantité de personnes sobres présentes dans le corpus analysé était d'environ 49.2 sobres pour 50.8 intoxiquées. Les ensembles qui résultent de ces modifications n'ont plus la même taille. Les analyses statistiques ont donc eu lieu sur 1 707 510 phonèmes et sur 110 716 unités prosodiques.

3.2 Régressions logistiques binomiales

En considérant les relations entre la variable dépendante, taux de l'alcool rendu binaire (sobre ou intoxiqué), et les variables indépendantes phonétiques qui sont précédemment exposées, nous avons décidé de faire une régression pour prédire si un segment, avec toutes ses informations linguistiques extraites, sera produit par une personne sobre ou intoxiquée. Plus précisément, la variable dépendante de notre analyse n'étant pas continue, sa valeur ne se situe pas dans un intervalle et ainsi, une régression linéaire n'est pas possible. Par contre, les régressions logistiques permettent, pour chaque valeur d'une VI, d'évaluer la probabilité que le sujet analysé appartienne à chaque catégorie de la VD. Notre variable dépendante, qui est l'état du locuteur ou de la locutrice, a deux valeurs (intoxiqué ou sobre) et ainsi nous utilisons une régression logistique binomiale. Par exemple, dans le cas qui nous concerne, pour chaque valeur de la F0 d'un segment, un modèle de régression logistique associera un chiffre de 0 à 1, 1 signifiant que le locuteur ou de la locutrice ayant prononcé ce segment est intoxiqué, et 0 indiquant que la personne est sobre. Une régression fonctionnera avec une formule semblable à celle vulgarisée en (1).

(1) $Y = aX$

Se basant sur toutes les valeurs de la VD (Y) et de la VI (X), une régression produit un coefficient a. C'est ce qu'il utilisera pour faire des prédictions. Cependant, notre analyse contient plus d'une variable indépendante et ainsi, la formule ressemble plus à celle en (2).

(2) $Y = aX_1 + bX_2 + cX_3 + ...$

Cette formule est celle, vulgarisée, d'un modèle linéaire général, ou GLM. Le résultat de ce calcul étant une multitude de coefficients, nous avons besoin de quelques autres outils pour les interpréter efficacement.

Pour nous aider à interpréter les résultats de notre modèle linéaire général, nous utilisons le *odds ratio*, ou OR (Barnier *et al.* 2018). Par exemple, un OR de 1 signifie que la variable indépendante (VI) n'a aucun effet sur la variable dépendante (VD). Un OR supérieur à 1 signifie que la présence de la VI augmente la probabilité que la VD a une certaine valeur (intoxiqué ou sobre), alors qu'un OR inférieur à 1 indique que la présence de la VI diminue les probabilités d'appartenance de la VD à une certaine catégorie (Glen 2014). Si, dans le cas qui nous concerne, un OR supérieur à 1 est associé à la variable *Durée*, cela signifie que plus la durée des segments augmente, plus la chance que le locuteur ou la locutrice qui les a prononcés soit intoxiqué à l'alcool est grande. Notons que cette relation en est une de corrélation et non de causalité, et ainsi va dans les deux sens. Ainsi, si la variable *Durée* a un OR supérieur à 1, cela signifie aussi que plus la probabilité que la personne soit intoxiquée est grande, plus la valeur de *Durée* sera grande.

L'*Akaike's Informtion Criteria*, ou AIC, est un indice de la parcimonie d'un modèle (Glen 2014) ; plus il est bas, plus le modèle est parcimonieux. La parcimonie d'un modèle est inspirée du principe du rasoir d'Occam, qui indique qu'on ne devrait pas utiliser plus d'éléments que nécessaire. Dans le cas qui nous concerne, nous devrions donc utiliser le modèle qui ne contient que les variables nous permettant de répondre à notre question, sans en intégrer plus. Les modèles avec beaucoup de paramètres par données tendent à être plus parcimonieux, avec le défaut qu'ils sont moins utiles pour faire des prédictions sur un autre ensemble de données. Par exemple, si on fait un modèle qui prédit la présence d'alcool chez une personne en considérant une très grande quantité de paramètres (par exemple son âge, la taille de son chapeau, le nombre d'enfants, etc), ce modèle sera performant seulement pour prédire l'intoxication chez les individus qui répondent à tous ces critères (donc qui ont un chapeau, des enfants, etc.). L'AIC n'est donc pas une mesure véritable du "meilleur modèle", mais nous permet de sélectionner le modèle qui contient le moins de variables inutiles pour qu'il soit plus facilement applicable à différents ensembles de données.

3.3 Sélection du modèle

Après avoir normalisé toutes les variables et produit un GLM avec l'entièreté des variables, nous obtenons les résultats montrés ci-dessous pour la segmentation par phonème. Notons que les mêmes étapes ont été exécutés pour l'ensemble contenant des unités prosodiques.

Tableau 1. Premier Modèle pour la segmentation par phonèmes (Maechler et al. 2019)

Variable	Estimate	Std.Error	z value	Pr(> |z|)
(Intercept)	0.259792	0.012600	20.619	$< 2e-16$ ***
CF1_c	-0.012904	0.014335	-0.900	0.3680
CF1_1_4	-0.004901	0.011992	-0.409	0.6828
CF1_3_4	-0.026218	0.012169	-2.155	0.0312 *
CF2_c	0.009029	0.015240	0.592	0.5535
CF2_1_4	-0.067121	0.013045	-5.145	2.67e-07 ***
CF2_3_4	-0.065426	0.012507	-5.231	1.68e-07 ***
F0_U	0.758717	0.009500	79.861	$< 2e-16$ ***
Duree	2.031865	0.193183	10.518	$< 2e-16$ ***
Intensite	-0.271026	0.012608	-21.497	$< 2e-16$ ***
Centre_de_gravite	0.167168	0.022886	7.304	$2.78e-13$ ***
Ecart_type	-0.477634	0.021017	-22.726	$< 2e-16$ ***
Skewness	-2.637655	0.097663	-27.008	$< 2e-16$ ***
Kurtosis	4.072297	0.377251	10.795	$< 2e-16$ ***
AIC: 2 355 726				

Le nombre d'astérisques à la droite du tableau nous expose la significativité des variables dans le modèle. Plus il y a d'astérisques, plus la significativité est forte. Cela nous montre que ces variables sont significatives selon la variable dépendante qui est le taux d'alcool binarisé (sobre ou intoxiqué). Nous avons retiré les variables ne montrant aucun astérisque une à une pour trouver le modèle qui obtiendrait un AIC le plus bas. Le modèle ayant un AIC le plus bas est celui que nous avons utilisé pour notre analyse. Enfin, pour respecter notre objectif de trouver quelle est la meilleure segmentation, nous avons utilisé une même composition de variable pour les deux segmentations.

4. Résultats

4.1 La segmentation par phonème

Après comparaison des AICs, le modèle contenant le moins de variables non significatives pour l'ensemble segmenté par phonèmes était celui dont le F1 pris au premier quart, le F1 pris au centre et le F2 pris au centre ont été retirés du modèle, avec un AIC de 2 355 721. Notre variable dépendante est catégorielle avec deux valeurs possibles, 0 qui signifie sobre et 1 qui signifie intoxiqué. Après l'avoir transformé en facteur à deux niveaux, nous avons procédé à une régression logistique binomiale (Maechler *et al.* 2019). Nous avons ensuite considéré l'OR et le P, soit l'indice de significativité des différentes variables, à l'aide d'un modèle Forest (Kennedy 2018).

Toutes les corrélations observées dans cet ensemble sont hautement significatives ($P < 0.001$) à l'exception de la variable *F1 prise au troisième quart* (CF1_3_4 ; $P < 0,05$), qui est moins fortement significative. L'observation de l'OR de la variable *Skewness* nous

indique que plus le *Skewness* d'un segment est grand, plus la chance que le locuteur ou la locutrice qui a produit le segment soit intoxiqué est faible. L'OR de la variable *Kurtosis* montre au contraire une très forte corrélation positive avec la VD. Les coefficients estimés par notre modèle sont observables dans le tableau ci-dessous.

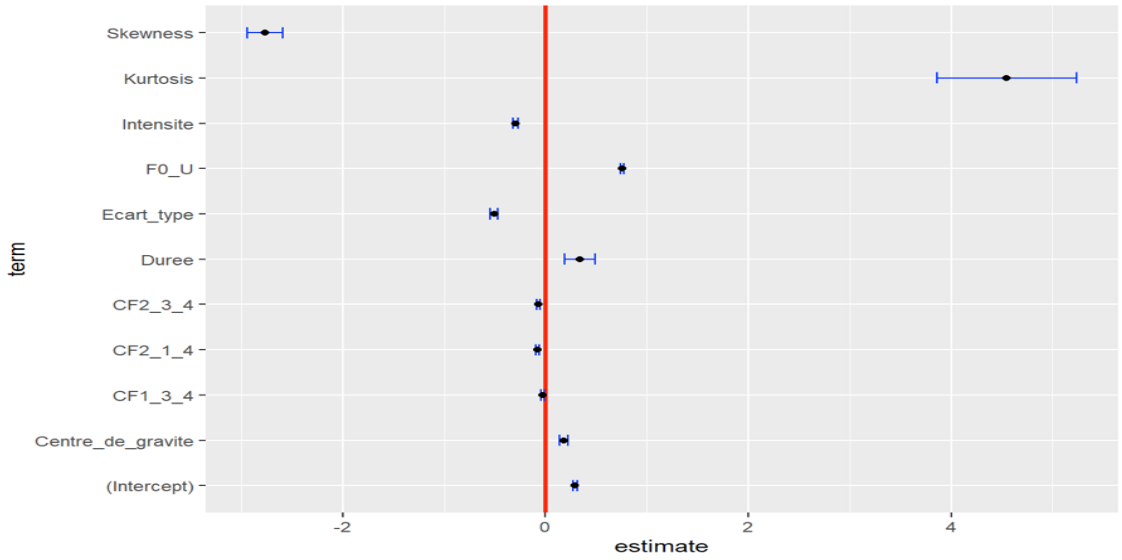

Figure 1. Coefficients estimés pour les phonèmes (Schloerke et al. 2018)

Le coefficient estimé de la variable *Kurtosis* étant proche de 5, on peut en comprendre que le kurtosis d'un segment provenant d'un locuteur intoxiqué augmentera de 5 par rapport au kurtosis d'un segment produit par un locuteur sobre.

Le tableau ci-dessous indique les nombres de prédictions du modèle utilisant une segmentation du signal par phonèmes.

Tableau 2. Prédictions pour les phonèmes

	Prédiction du modèle	
État de l'individu	Sobre	Intoxiqué
Sobre	410 296	373 777
Intoxiqué	428 776	494 661

Lorsque le modèle prédit que la personne est sobre mais qu'elle est en fait intoxiquée, ou vice versa, le résultat est erroné et affiché en rouge. Cela nous montre que le modèle de régression logistique binomiale prédira que le sujet est sobre sur 49.14% des segments, ayant raison 48.89% du temps. Il prédira que le sujet est intoxiqué sur 50.85% de l'échantillon, ayant raison 56.95% du temps. Cela résulte en un pourcentage de bonne réponse total de 52.99%.

4.2 La segmentation par unité prosodique

Après comparaison des AICs, le modèle ayant un AIC le plus bas est celui qui contient l'ensemble des variables avec un AIC de 152 538. Malgré le fait d'avoir enlevé les variables n'étant pas significatives, le AIC augmentait. Toutefois, dans le but de respecter la question de recherche qui vise à comparer la segmentation par phonème et la segmentation par unité prosodique, nous avons utilisé le même modèle pour les deux segmentations. Les variables F1 prise au premier quart, la F1 prise au centre, et la F2 prise au centre ont donc été retirées aussi. Notre variable dépendante est catégorielle avec deux valeurs possibles, 0 qui signifie sobre et 1 qui signifie intoxiqué. Après l'avoir transformé en facteur à deux niveaux, nous avons procédé à une régression logistique binomiale (Maechler *et al.* 2019). Nous avons ensuite considéré, comme dans le cas des phonèmes, l'OR et le P, soit l'indice de significativité des différentes variables, à l'aide d'un modèle Forest (Kennedy 2018).

Nous avons pu voir que la variable *F1 prise au troisième quart* (CF1_3_4) n'est pas significative (P = 0.664). Les variables *F1 prise au centre* (CF1_c ; P < 0.05) et *Kurtosis* (P < 0.05) sont significatives. Toutes les autres variables de cet ensemble sont hautement significatives (P < 0.001). En observant l'OR, on peut observer que, en excluant les variables formantiques, toutes les variables ont une corrélation, négative ou positive, plus ou moins prononcée avec la VD. Tout comme dans l'ensemble séparé en phonèmes, on peut voir une forte corrélation positive entre la VD et la VI *Kurtosis*, ainsi qu'une corrélation négative avec la VI *Skewness*. Chez les unités prosodiques, on peut voir que les autres variables spectrales entre elles aussi en corrélation avec la VD. On peut ainsi voir une corrélation positive la variable *Centre de gravité* et une corrélation négative avec la variable *Écart type*. Les coefficients estimés par notre modèle pour les unités prosodiques sont observables dans le tableau ci-dessous.

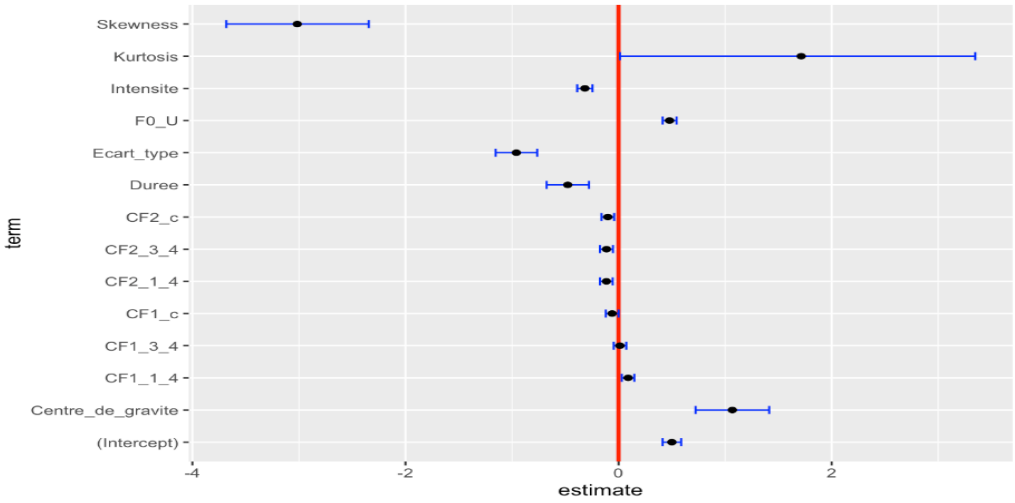

Figure 2. Coefficients estimés pour la segmentation par unités prosodiques (Schloerke et al. 2018)

Dans le tableau ci-dessus, on peut tout de suite voir que les coefficients estimés sont moins forts que pour l'ensemble séparé par phonème en observant les chiffres indiqués sur l'axe des X. Le coefficient entre la variable *Kurtosis* et la VD se rend presqu'à 4, signifiant que le *Kurtosis* d'un segment augmentera de 4 lorsque prélevé sur une unité prosodique produite par un locuteur intoxiqué. Rappelons-nous que cette corrélation montrait un coefficient estimé allant jusqu'à 5 pour l'ensemble séparé par phonèmes. On peut aussi voir que la variable *Skewness* montre un coefficient estimé allant presque jusqu'à -4, ce qui signifie que lorsque prélevé sur un locuteur intoxiqué, un segment aura une valeur de la variable *Skewness* diminuée de 4.

Le tableau ci-dessous indique les nombres de prédictions du modèle utilisant une segmentation du signal par unités prosodiques.

Tableau 3. Prédictions pour les unités prosodiques

	Prédiction du modèle	
État de l'individu	Sobre	Intoxiqué
Sobre	16905	15555
Intoxiqué	35999	42257

Lorsque le modèle prédit que la personne est sobre mais qu'elle est en fait intoxiquée, ou vice versa, le résultat est erroné et affiché en rouge. Cela nous montre que le modèle de régression logistique binomiale prédira que le sujet est sobre sur 47.78% des segments, ayant raison 31.95% du temps. Il prédira que le sujet est intoxiqué sur 52.22% de l'ensemble, ayant raison 73.09% du temps. Cela résulte en un pourcentage de bonne réponse total de 53.44%.

5. Discussion

En utilisant le *Alcohol Language Corpus*, nous avons présenté deux exécutions d'un modèle de régression logistique binomiale dans le but de comparer deux segmentations linguistiques: une par unité phonémique et une autre par unité prosodique. Il ne semble pas y avoir de différence entre le taux de prédictions correctes d'un modèle de régression logistique binomiale selon la segmentation du signal sonore. C'est-à-dire que les prédictions pour les unités prosodiques ainsi que pour les phonèmes sont tout aussi basses, avec un taux de bonnes réponses d'environ 53%. Il nous est donc impossible de confirmer notre hypothèse de départ qui prévoyait de meilleurs résultats sur la segmentation par phonèmes.

Pour améliorer le taux de bonnes réponses prédites par le modèle, plusieurs modifications pourraient être apportées. Par exemple, notre modèle étant binomial, il faisait ses prédictions sur un facteur à deux niveaux. Ainsi, ce que nous avons fait en revient à classifier la parole comme étant prononcé soit par l'archétype du "saoul", soit par l'archétype du "sobre". Cependant, la signification de ce que sont les états "saouls" et "sobres" qui sont différents pour chaque individu. Une future analyse par régression logistique multiniveaux permettrait de prendre en considération la différence

interindividuelle des réactions à l'intoxication, en considérant chaque individu comme variable indépendante. La variable dépendante de notre analyse étant catégorielle, elle ne permet pas au modèle de prédire les différences entre beaucoup et moindrement intoxiqué. Aussi, la variable continue du taux d'alcool dans le sang pourrait directement être considérée comme la variable dépendante de l'analyse statistique. En revanche, notre analyse a impliqué un processus d'équilibration des valeurs de la variable dépendante, en s'assurant qu'à chaque enregistrement d'un locuteur sobre, il y ait un enregistrement d'un locuteur intoxiqué. Préalablement à cette modification, le taux de bonnes réponses du modèle étant exactement égal à la proportion d'enregistrements sobres dans l'échantillon. Si à première vue un pourcentage de 81% de bons résultats est une bonne nouvelle, nous avons constaté que ce taux correspondait à la proportion de fichiers étiquetés "sobres". C'est-à-dire que le modèle prédisait tout simplement "sobre" à chaque segment, peu importe les variables. Changer la proportion d'individus sobres présents dans l'échantillon changeait donc directement le nombre de bonnes prédictions du modèle, du moins jusqu'à ce que la proportion s'approche de 50%. Ce phénomène est peut-être dû à un *overfitting*, qui pourrait être contrebalancé par les réajustements précédemment proposés, ou avec une reconsidération du choix de modèle statistique dans le but de contrer l'*overfitting*.

Une autre possibilité de poursuite de la question se trouve dans le choix d'une différente segmentation à comparer. Les segmentations par unités prosodiques et par phonèmes ne sont pas les seuls types de segmentation qui précèdent la sélection préalable de la langue étudiée. Une comparaison plus systématique avec des segmentations non linguistiquement justifiées, comme un découpage par dix millisecondes pourrait être accompli. Sinon, la syllabe est elle aussi une segmentation préalable à la sélection de la langue, mais avec l'avantage d'être bâtie autour de considérations linguistiques.

Les choix des variables indépendantes sont aussi ouverts à l'amélioration. La majorité de nos variables étant phonétique, elle s'applique donc à des phonèmes. Ainsi, il serait intéressant de faire une analyse avec le prélèvement de variables prosodiques telles que les formants dits "hauts", comme la F4, F5, ou la F6 sur des segments prosodiquement justifiés. Les valeurs des variables spectrales sont prélevées sur la répartition des valeurs dans le partiel d'un spectre. Ceci est fait en utilisant des informations sur la répartition globale des valeurs aussi souvent utilisées en statistiques, comme l'aplatissement, l'asymétrie, l'écart type et le centre de gravité. Le même type d'information sur la distribution pourrait aussi être prélevé sur la représentation de l'intensité d'un segment par rapport au temps, ou sur celle de la fréquence d'un segment par rapport au temps. De telles variables auraient l'avantage, contrairement aux variables phonétiques sélectionnées dans cette analyse, d'être tout aussi pertinentes sur les deux types de segmentations. En plus d'ajouter d'autres variables phonétiques plus pertinentes, l'ajout de variable comme le nombre d'erreurs de prononciation, d'hésitation et d'omission, aurait pu participer à l'amélioration de l'algorithme de classification. La sélection des variables indépendantes aurait donc probablement avantage à être considérée avec plus d'attention. Cependant, cette attention ne doit pas nécessairement être produite par nous-mêmes. En effet, un système d'apprentissage automatique par réseau de neurones pourrait, si ce n'est répondre à la question par celui-ci, au moins aider dans une sélection plus performante des variables intégrées dans l'analyse.

6. Conclusion

Nous avons tenté d'aborder la problématique de la détection automatique de l'intoxication à l'alcool par la parole. Cet article contient une brève revue de la littérature sur le sujet, ainsi qu'une explicitation des contradictions qu'on y retrouve. Nous introduisons ensuite le *Alcohol Language Corpus*, ou ALC, que nous utilisons pour faire notre propre analyse. Nous décrivons ensuite notre processus d'extraction de données linguistiques selon 13 variables phonétiques du Corpus, suivi de définitions des concepts que nous utilisons dans notre analyse : modèles de régressions logistiques binomiales, *Odds Ratio* (ou OR), et *Akaike's Information Criteria* (ou AIC). Nous procédons ensuite à l'étalement des résultats de notre analyse, montrant notamment des corrélations entre l'intoxication à l'alcool et les variables spectrales *Kurtosis* et *Skewness*. Des matrices de confusions nous indiquent ensuite que les pourcentages de bonnes prédictions de nos modèles, soit 52.99% pour les phonèmes et 53.60% pour les unités prosodiques, est insuffisante. Ce taux de bonne réponse est inférieur au 74% de bonnes réponses obtenues par une détection faite par l'humain. De plus, notre hypothèse prévoyant qu'une segmentation du signal sonore par phonèmes permettrait une meilleure détection automatique n'a pu être confirmée ni infirmée. Nous procédons ensuite à l'énumération des raisons potentielles pour l'obtention de telles réponses, ainsi que plusieurs pistes de solutions pour remédier aux résultats de notre analyse. Nous finissons donc en mentionnant qu'une analyse multiniveau qui prend en compte les locuteurs, l'utilisation d'un autre système d'apprentissage automatique, une sélection plus attentive des variables et l'ajout d'un autre type de segmentation préalable au choix de la langue observée, comme les syllabes, pourraient représenter des pistes intéressantes d'analyses ultérieures et provoquer de meilleurs résultats.

Références

Alderman, G. Allan, Harry Hollien, Camilo Martin, et Gea DeJong. 1995. *Shifts in fundamental frequency and articulation resulting from intoxication.* The Journal of the Acoustical Society of America. *97*(5): 3363–3364. doi: 10.1121/1.412694

Barfüßer, Sabine et Florian Schiel. 2010. *Disfluencies in alcoholized speech*.

Barnier, Julien, François Briatte, et Joseph Larmarange. 2018. *questionr: Functions to Make Surveys Processing Easier*. Récupéré de https://CRAN.R-project.org/package=questionr

Baumeister, Barbara, Christian Heinrich, et Florian Schiel. 2012. *The influence of alcoholic intoxication on the fundamental frequency of female and male speakers.* The Journal of the Acoustical Society of America, 132(1): 442-451. doi: 10.1121/1.4726017

Boersma, Paul et David Weenink. 2009. *Praat: doing phonetics by computer (Version 5.1.13)*. Récupéré de http://www.praat.org

Bone, Daniel, Ming Li, Matthew P. Black, et Shrikanth S. Narayanan. 2014. *Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functionals and GMM supervectors.* Computer Speech & Language, 28(2): 375-391. doi: 10.1016/j.csl.2012.09.004

Braun, Angelika et Hermann J. Künzel. 2003. *The effect of alcohol on speech prosody.* Dans Proceedings of the International Congress of Phonetic Sciences, Barcelona (vol. 2645, p. 2648).

Cooney, Orla M., Kevin G. McGuigan, Peter J. P. Murphy, et Ronan M. Conroy. 1998. *Acoustic analysis of the effects of alcohol on the human voice.* The Journal of the Acoustical Society of America, 103(5): 2895-2895. doi: 10.1121/1.421829

Fairbairn, Catharine E., Michael A. Sayette, Marlissa C. Amole, John D. Dimoff, Jeffrey F. Cohn, et Jeffrey M. Girard. 2015. *Speech volume indexes sex differences in the social-emotional effects of alcohol.* Experimental and Clinical Psychopharmacology, 23(4): 255-264. doi: 10.1037/pha0000021

Glen, Stephanie. 2014. *Odds Ratio Calculation and Interpretation*. Récupéré de https://www.statisticshowto.com/odds-ratio/

Heinrich, Christian et Florian Schiel. 2014. *The influence of alcoholic intoxication on the short-time energy function of speech.* The Journal of the Acoustical Society of America, 135(5): 2942-2951. doi: 10.1121/1.4870705

Hollien, Harry, Gea DeJong, Camilo A. Martin, Reva Schwartz, et Kristen Liljegren. 2001. *Effects of ethanol intoxication on speech suprasegmentals.* The Journal of the Acoustical Society of America, 110(6): 3198-3206. doi: 10.1121/1.1413751

Hollien, Harry, James D. Harnsberger, Camilo A. Martin, Rebecca Hill, et G. Allan Alderman. 2009. *Perceiving the Effects of Ethanol Intoxication on Voice.* Journal of Voice, 23(5): 552-559. doi: 10.1016/j.jvoice.2007.11.005

Kennedy, Nick (2018). *forestmodel: Forest Plots from Regression Models*. Récupéré de https://CRAN.R-project.org/package=forestmodel

Klingholz, Fritz, Randolph Penning, et E. Liebhardt. 1988. *Recognition of low-level alcohol intoxication from speech signal*. The Journal of the Acoustical Society of America, 84(3): 929-935. doi: 10.1121/1.396661

Levit, Michael, Richard Huber, Anton Batliner, et Elmar Noeth. 2001. *Use of prosodic speech characteristics for automated detection of alcohol intoxication.* Dans ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, et Kurt Hornik. 2019. *Cluster Analysis Basics and Extensions*. (s. l. : n. é.).

Pisoni, David B. et Christopher S. Martin. 1989. *Effects of Alcohol on the Acoustic-Phonetic Properties of Speech: Perceptual and Acoustic Analyses.* Alcoholism: Clinical and Experimental Research, 13(4): 577-587. doi: 10.1111/j.1530-0277.1989.tb00381.x

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing. Récupéré de http://www.R-project.org/

SAAQ. 2015. *Profil détaillé des faits et des statistiques touchant les véhicules lourds.* Récupéré de http://collections.banq.qc.ca/ark:/52327/2755492

SAAQ. 2018. *Alcool au volant : ce que dit la loi.*

Schiel, Florian et Christian Heinrich. 2009. *Laying the foundation for in-car alcohol detection by speech.* Dans Tenth Annual Conference of the International Speech Communication Association.

Schiel, Florian, Christian Heinrich, et Sabine Barfüsser. 2012. *Alcohol language corpus: the first public corpus of alcoholized German speech.* Language Resources and Evaluation, 46(3): 503-521. doi: 10.1007/s10579-011-9139-y

Schiel, Florian, Christian Heinrich, et Veronika Neumeyer. 2010. *Rhythm and formant features for automatic alcohol detection.* Dans Eleventh Annual Conference of the International Speech Communication Association.

Schloerke, Barret, Jason Crowley, Di Cook, François Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg et Joseph Larmarange. 2018. *GGally: Extension to « ggplot2 »*. Récupéré de https://CRAN.R-project.org/package=GGally

Sobell, Linda C., Mark B. Sobell, et Robert F. Coleman. 1982. *Alcohol-Induced Dysfluency in Nonalcoholics.* Folia Phoniatrica et Logopaedica, 34(6): 316-323. doi: 10.1159/000265672

Thibeault, Mélanie. 2011. *Les émotions : Une étude Articulatoire, Acoustique et Perceptive.*

Watanabe, Hiroshi, Takemoto Shin, Hiromichi Matsuo, Fumio Okuno, Tsutomu Tsuji, Midori Matsuoka, Junichi Fukaura, et Hisashi Matsunaga. 1994. *Studies on vocal fold injection and changes in pitch associated with alcohol intake.* Journal of Voice, 8(4): 340-346. doi: 10.1016/S0892-1997(05)80282-6