XXXIIe Journées d'Études sur la Parole
4-8 June 2018, Aix-en-Provence, France

Perception des consonnes et voyelles nasales en parole vocodée : Analyse de la contribution des niveaux de résolution spectrale et temporelle.

Olivier Crouzet[1,2]
(1) LLING – Laboratoire de Linguistique de Nantes – UMR6310 Université de Nantes / CNRS, chemin de la Censive du Tertre, 44312 Nantes Cedex, France
(2) ENT Department – University Medical Center Groningen, Rijksuniversiteit Groningen, Pays-Bas
olivier.crouzet@univ-nantes.fr

Résumé

Nous présentons une série d'expériences dans lesquelles nous étudions l'impact des niveaux de résolution spectrale et temporelle de signaux de parole vocodée par canaux de bruit sur la classification des consonnes et voyelles du français en nous intéressant plus particulièrement à la comparaison orales / nasales. Les réponses perceptives ont été recueillies séparément pour l'identification des consonnes et des voyelles dans des tâches de classification à choix forcé à N alternatives. Nous étudions dans un premier temps la relation entre les performances de classification et les niveaux de résolution spectrale et temporelle des signaux. Nous présentons ensuite des analyses d'entropie mutuelle qui permettent d'évaluer le degré de préservation des différents « traits » associés aux catégories sonores. Ces résultats font ressortir des difficultés particulières posées par le trait de nasalité pour la classification des voyelles. Un certain nombre de questions méthodologiques et théoriques émergent de ces données et sont discutées.

Abstract

Perceptual classification of nasal consonants and vowels in vocoded speech: Contribution of spectral and temporal resolution levels.

A series of experiments is described in which we investigated the impact of spectral and temporal resolution of channel-vocoded speech on French consonants and vowels, focusing our main interests on a comparison of oral and nasal segments. Participants provided perceptual responses in N-Alternative Forced-Choice tasks focusing on either consonant or vowel identification. We first discuss overall classification performance levels in relation to spectral and temporal resolution of signals. We then turn to mutual entropy analyses which let us estimate information transmission preservation for individual featural categories. From these analyses, it is observed that specific difficulties seem to be associated with nasal vowels. Further methodological and theoretical issues are then discussed.

Mots-clés : parole vocodée ; implants cochléaires ; résolution spectrale ; résolution temporelle ; perception ; théorie de l'information ; entropie mutuelle ; matrices de confusion.

Keywords: vocoded speech; cochlear implants; spectral resolution; temporal resolution; perception; information theory; mutual entropy; confusion matrices.

10.21437/JEP.2018-46

1 Introduction

Les types de stimulation produits par les implants cochléaires sont décrits comme transmettant une information de type « variations d'énergie intra-bande » (Shannon *et al.*, 1995) et induisent donc une perte d'information spectrale par rapport aux capacités naturelles du système auditif puisque la résolution spectrale de l'implant dépend essentiellement du nombre d'électrodes implantées (ce nombre étant actuellement lui-même limité par des contraintes techniques liées notamment à la taille de ces électrodes et aux risques de diffusion de potentiel –en anglais « current spread »– liés à cette taille). Néanmoins, on observe depuis les travaux princeps de Van Tasell *et al.* (1987) que des formes acoustiques de ce type (signaux à canaux vocodés) sont en mesure de fournir des informations relativement efficaces pour la classification phonétique. Les données plus récentes portant sur la modélisation du traitement des signaux par les implants cochléaires conduisent à considérer que la « qualité de la résolution spectrale » n'est pas primordiale pour la reconnaissance de la parole (Shannon *et al.*, 1995; Kanedera *et al.*, 1999) et que l'information phonétique serait principalement portée par les « fréquences de modulation d'amplitude » (Kanedera *et al.*, 1999; Christiansen & Greenberg, 2010) qui sont associées aux canaux spectraux, lesquelles correspondent aux modulations d'énergie de basse fréquence qui caractérisent les variations d'amplitude de l'onde dans un canal spectral.

Chez les patients porteurs d'implants cochléaires, outre des variations individuelles importantes en termes de « récupération » rendue possible par l'appareillage, l'une des propriétés qui semble résister à la prise en charge orthophonique est la nasalité. Ces problèmes correspondent aussi bien à des phénomènes de « qualité de la voix » (hyper– / hypo–nasalisation ; Fletcher *et al.*, 1999; Baudonck *et al.*, 2015) que de discrimination phonologique en perception (Borel, 2015, pour les voyelles du français).

Même si les aspects articulatoires de la nasalité peuvent parfois paraître relativement simples, aussi bien la question des propriétés articulatoires des distinctions orales / nasales (Delvaux, 2012; Demolin *et al.*, 2003; Carignan *et al.*, 2015) que celle des effets acoustiques de la nasalité (House & Stevens, 1956; Maeda, 1982; Stevens, 1998; Feng & Castelli, 1996; Rossato, 2000) continuent de poser des problèmes cruciaux en termes de modélisation théorique des relations articulatoire–acoustique et des mécanismes de perception. Or il se trouve que les analyses phonétiques qui sont proposées dans le cadre du modèle source-filtre suggèrent une complexité du spectre de sortie très marquée, laquelle s'explique par le phénomène de couplage entre les cavités orale et nasale (Maeda, 1982; Stevens, 1998, cf. Fig. 1). Si les propriétés des consonnes et des voyelles nasales diffèrent (en raison des contributions relatives différentielles des cavités orale et nasale et de leur organisation temporelle), les mécanismes fondamentaux impliqués peuvent générer certains phénomènes acoustiques similaires comme l'élargissement de la largeur de bande des formants ou l'atténuation de l'énergie du signal. Certains de ces mécanismes pourraient constituer un frein à la transmission des informations acoustiques pertinentes dans un implant cochléaire.

Afin d'évaluer les contributions respectives des paramètres de résolution spectrale et temporelle associés à la perception des nasales, nous avons mis en place une série d'expériences dans lesquelles nous étudions la catégorisation d'un sous ensemble des consonnes et voyelles du français dans deux tâches distinctes (une tâche de classification des consonnes, une de classification des voyelles). Notre objectif est d'évaluer le rôle de ces deux domaines de résolution sensorielle pour la classification des catégories sonores et d'étudier plus particuliè-

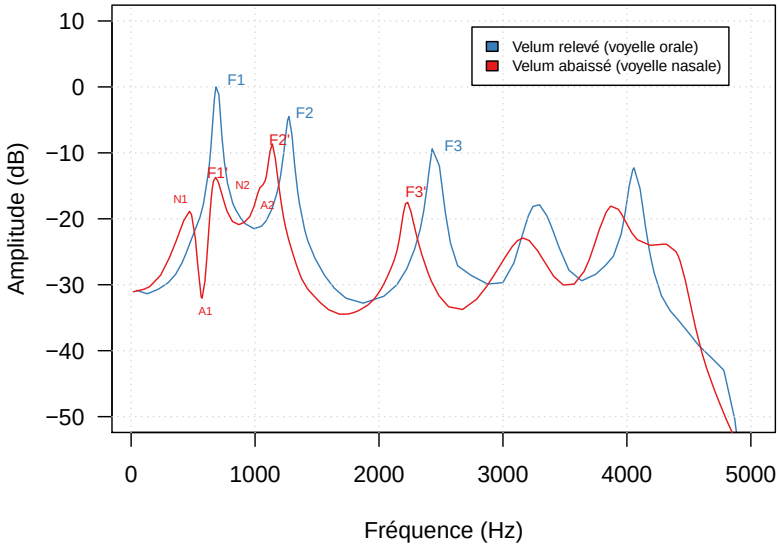


FIGURE 1: Modélisation articulatoire–acoustique de l'abaissement du velum pour la production des voyelles proposée par Maeda (1982). Cette simulation met en évidence l'impact du couplage oral–nasal et des anti-résonances (A_1, A_2) qui en découlent sur le spectre de sortie : diminution de l'amplitude de certains formants « oraux »($F_1 \rightarrow F_1'$, $F_2 \rightarrow F_2'$, $F_3 \rightarrow F_3'$), décalages en fréquence, combinaison de résonances « orales » et « nasales » (N_1, N_2) et des anti-résonances provoquant une augmentation des largeurs de bandes. Graphique adapté de Maeda (1982).

rement le comportement des catégories nasales dans les performances de classification. Ces données pourraient contribuer à améliorer le traitement des informations acoustiques pour les personnes déficientes auditives portant un implant cochléaire mais fourniraient aussi des informations essentielles pour la modélisation acoustique des nasales et la compréhension des mécanismes perceptifs qui leur sont associés.

2 Méthode

Deux expériences parallèles ont été conçues afin d'évaluer les impacts perceptifs des propriétés de résolution spectrale et temporelle de la parole sur l'identification des consonnes et voyelles nasales. Dans la première expérience on étudie la classification perceptive de consonnes dans un contexte VCV. Dans la seconde expérience, l'identification de voyelles est abordée à travers la classification de segments vocaliques isolés ne présentant pas de transitions formantiques. Dans les deux cas, les participants réalisent une tâche de classification à choix forcé à N alternatives.

2.1 Expérience 1 : Consonnes

2.1.1 Participants

Dix-neuf participants volontaires adultes n'ayant pas de troubles auditifs connus ont pris part à l'expérience. Les résultats de 3 d'entre eux ont été retirés des données en raison de

problèmes techniques n'ayant pas permis de réaliser l'intégralité de l'expérience. Les résultats des 16 participants restants sont présentés.

2.1.2 Matériel

Les stimuli sont des séquences VCV [1] sans signification. La consonne (choisie parmi 19 consonnes du français dans l'ensemble {b, d, g, p, t, k, v, z, ʒ, f, s, ʃ, l, ʁ, w, j, m, n, ɲ}) est produite dans 3 contextes vocaliques différents ({i, a, u}). Les stimuli ont été produits par une locutrice adulte et numérisés sans compression à une fréquence d'échantillonnage de 16 kHz avec un taux de quantification de 16 bits. Ils ont ensuite été traités par un vocodeur à canaux de bruits développé dans l'environnement Octave (Eaton *et al.*, 2015) en faisant varier 2 paramètres : le nombre de canaux spectraux (2, 4, 6, 8 canaux de bruit) et la fréquence de coupure des modulations d'amplitude (filtre passe-bas ; 4, 16, 128 Hz). Dans la condition 4 Hz, seules les fréquences de modulation lentes sont préservées. Pour la fréquence de coupure 16 Hz, les fréquences de modulation lentes et moyennement rapides sont préservées. À 128 Hz, toutes les fréquences de modulation sont préservées. Les fréquences de coupure des modulations d'amplitude ont été choisies sur la base d'évaluations subjectives informelles. Les fréquences du banc de filtres passe-bande utilisé pour créer les canaux de bruit suivent une progression régulière en ERB (Moore & Glasberg, 1983) selon l'implémentation de Slaney (1993).

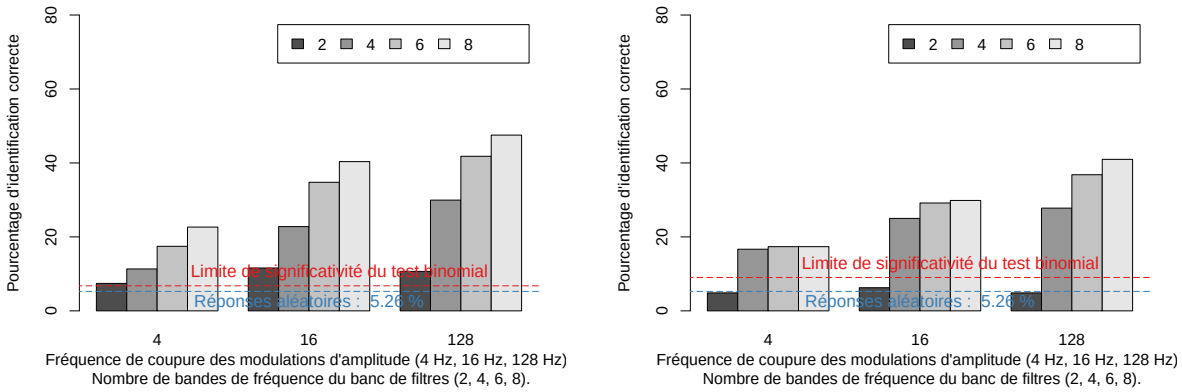
2.1.3 Procédure

Le recueil des données se faisait dans une pièce calme. Les stimuli étaient présentés à travers un casque à un niveau sonore jugé confortable par les participants. L'expérience débutait par une familiarisation avec la tâche. Les stimuli étaient ensuite présentés dans un ordre aléatoire. Un tableau représentant les consonnes sous forme orthographique était affiché sur l'écran de l'ordinateur sur lequel l'expérience était réalisée. Les participants avaient pour tâche de cliquer sur la case qui contenait la consonne identifiée. Il n'était pas possible de réécouter le stimulus. Dès que la réponse était donnée, le stimulus suivant était diffusé. Les participants pouvaient faire autant de pauses qu'ils le souhaitaient pendant la session. Chaque stimulus était présenté 1 fois. Au total, chaque participant donnait 684 réponses (19 consonnes $\times$ 3 voyelles $\times$ 4 conditions de nombre de canaux spectraux $\times$ 3 conditions de résolution temporelle).

2.1.4 Résultats

Les taux de reconnaissance correcte ont été calculés sur l'ensemble de l'expérience mais les résultats sont présentés en séparant les consonnes orales et les consonnes nasales de manière à évaluer les différences potentielles entre ces deux catégories. Ces résultats sont représentés dans la figure 2. Nous étudions l'impact de la résolution spectrale et temporelle sur les performances de classification. Un test binomial a été appliqué pour chaque condition afin de déterminer si le taux de réponses correctes dépasse significativement la probabilité aléatoire de réponses correctes ($1/19 \times 100 = 5.26\%$). Les nasales étant moins nombreuses que les orales, le seuil de significativité est plus haut pour les nasales.

1. Voyelle-Consonne-Voyelle



(a) (b)

FIGURE 2: Performances de classification des consonnes orales (Fig. 2a) et nasales (Fig. 2b). Le trait horizontal bleu indique le pourcentage de réponses correctes correspondant au taux théorique de réponses au hasard dans la tâche de classification à choix forcé. Le trait rouge caractérise le niveau de pourcentage de réponses correctes au-delà duquel les performances mesurées dépassent significativement (pour un test binomial et un seuil $p < 0.05$) le taux aléatoire de réponses correctes.

Globalement, on observe que la performance progresse conjointement à l'amélioration de la résolution spectrale et temporelle. La plupart des conditions permettent de dépasser le taux de réponses correctes aléatoire. C'est le cas dans toutes les conditions pour les consonnes orales. Pour les consonnes nasales, seule la condition la plus dégradée (2 canaux quelles que soient les fréquences de modulation disponibles) empêche les participants de dépasser le seuil de réponses au hasard. Dans toutes les autres conditions et de manière similaire aux consonnes orales, la performance est significativement supérieure au hasard et s'améliore avec l'accroissement des niveaux de résolution spectrale et temporelle. On voit donc que la difficulté de classification des consonnes nasales est légèrement plus marquée que pour les orales mais que globalement la forme de la progression est similaire. Cette légère différence pourrait notamment s'expliquer par la plus faible proportion de nasales dans la langue et / ou dans le matériel de l'expérience ainsi que par de plus grandes difficultés à traiter l'information acoustique propre aux nasales.

2.2 Expérience 2 : Voyelles

2.2.1 Participants

Les mêmes dix-neuf participants ont participé à l'expérience au cours de la même session. Les résultats de 2 d'entre eux ont été retirés des données en raison de problèmes techniques et seules les données des 17 participants restants sont présentées.

2.2.2 Matériel

Les stimuli sont des voyelles isolées qui ont été produites par la même locutrice au cours de la même session. Lors de l'enregistrement, des mots contenant ces voyelles dans leur

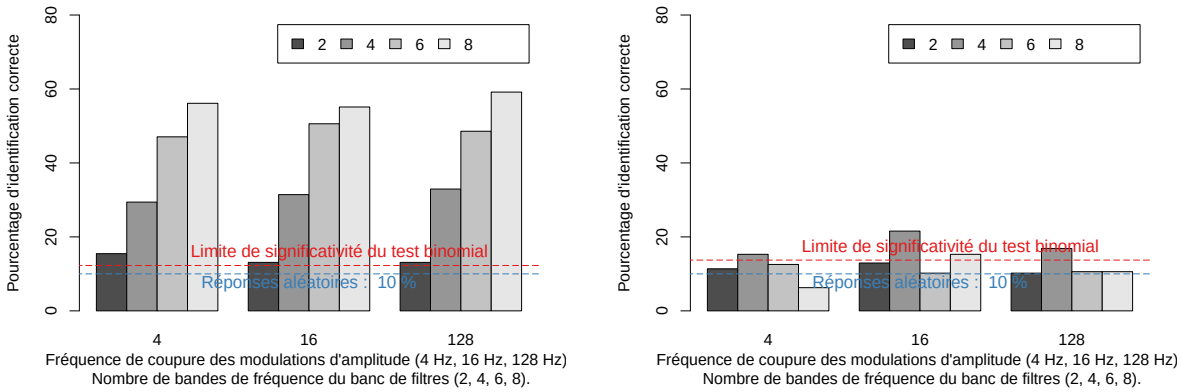


FIGURE 3: Performances de classification des voyelles orales (Fig. 3a) et nasales (Fig. 3b).

première syllabe (ouverte) étaient présentés à l'écran et la locutrice devait produire le mot en maintenant la réalisation de la première voyelle pendant plusieurs secondes. Les enregistrements ont ensuite été édités pour extraire une portion spectralement stable de 250 ms de chaque voyelle en appliquant une enveloppe d'accroissement puis d'atténuation progressive et rapide (10ms de temps de montée / descente) de l'amplitude au début et à la fin du signal. Les 10 voyelles utilisées sont : {i, e, a, y, ø, u, o, ɛ̃, ɑ̃, ɔ̃}.

Les stimuli ont été traités selon les mêmes principes que dans l'expérience 1. Il est important de noter que pour cette expérience, il n'est attendu aucun impact des fréquences de coupure des modulations d'amplitude puisqu'il n'y a normalement aucune information temporelle distinctive dans ces voyelles : elles sont stables du point de vue du spectre et de l'énergie. Nous avons choisi de conserver les mêmes conditions dans la perspective d'une étude ultérieure des propriétés dynamiques des voyelles.

2.2.3 Procédure

La procédure était la même que dans l'expérience 1. L'interface graphique utilisée affichait des équivalents orthographiques des voyelles pour recueillir les réponses des participants. Afin d'obtenir approximativement la même quantité de données dans les deux expériences, chaque stimulus était répété 5 fois au cours de l'expérience. Au total, chaque participant donnait 600 réponses (10 voyelles $\times$ 4 conditions de nombre de canaux spectraux $\times$ 3 conditions de résolution temporelle $\times$ 5 répétitions).

2.2.4 Résultats

Nous avons procédé aux mêmes analyses que dans l'expérience 1. Les résultats observés respectivement pour les voyelles orales et nasales sont présentés dans la figure 3. Le taux théorique de réponses au hasard est de $1/10 \times 100 = 10\%$. De même qu'avec les consonnes, les voyelles nasales étant moins nombreuses que les voyelles orales, le seuil de significativité est plus haut pour les nasales

Pour les voyelles orales, on observe un comportement des réponses en fonction des niveaux de résolution spectrale et temporelle tout à fait similaire aux consonnes orales. À l'inverse, les performances observées pour les voyelles nasales sont fortement dégradées. D'une part, très peu de conditions donnent lieu à des taux de reconnaissance correcte qui dépassent significativement le seuil de réponses au hasard. Seules la condition à 4 canaux (pour toutes les fréquences de modulation d'amplitude) et la condition à 8 canaux pour la fréquence de coupure 16 Hz donnent lieu à des performances dépassant le seuil de significativité du test binomial. D'un point de vue global, l'amélioration progressive des performances qui est observée sur les consonnes (orales et nasales) et sur les voyelles orales avec l'accroissement de la résolution spectrale est totalement absente des résultats observés sur les voyelles nasales.

2.3 Analyses d'entropie mutuelle

Afin d'affiner l'interprétation des données de performance, nous avons procédé à l'analyse quantitative des matrices de confusion à travers des calculs d'entropie mutuelle (qu'on appelle aussi « Taux de transfert d'information » à travers un canal / un medium ; Shannon, 1948; Miller & Nicely, 1955; Christiansen & Greenberg, 2012) en nous focalisant sur des traits articulatoires permettant de regrouper les segments en catégories. Afin de compenser les problèmes d'interprétation liés aux déséquilibres des effectifs des différentes catégories et au nombres de catégories distinctes, nous avons calculé l'entropie mutuelle *relative* (ou « taux de transfert d'information normalisé »).

Par manque de place, il n'est pas possible de détailler ces résultats ici. Les mesures effectuées confirment la difficulté à différencier les voyelles orales des nasales dans toutes les conditions de résolution étudiées. La différenciation entre consonnes orales et nasales, correcte mais néanmoins assez limitée, semble comparable à la classification de la place d'articulation ou du mode, le voisement étant l'information la mieux perçue. Pour les voyelles, les propriétés d'aperture, de position et d'arrondissement semblent aussi bien transmises que le voisement pour les consonnes.

Globalement, il semble donc que l'information de nasalité des voyelles soit très nettement dégradée en parole vocodée et ce, quelles que soient les conditions de résolution spectrale étudiées dans l'expérience.

3 Discussion

Les résultats de performance semblent mettre en évidence une forte difficulté des participants à identifier correctement les voyelles nasales. Cette observation est notamment soulignée par la forme globale des résultats : il ne semble y avoir aucune amélioration progressive de la performance de classification avec le niveau de résolution spectrale. Le fait que certaines conditions donnent lieu à des taux de réponse qui dépassent significativement le seuil de réponses au hasard pourrait être la conséquence de résultats statistiques aléatoires liés au risque d'erreur de Type I. Néanmoins, il est intéressant de constater que cette situation s'observe de manière cohérente pour les 3 fréquences de coupure des modulations d'amplitude affectées aux stimuli composés de 4 canaux. Or ce chiffre pourrait correspondre au nombre de canaux idéalement efficaces pour représenter les régularités statistiques de la parole (Ming &

Holt, 2009). Les fréquences limites des filtres utilisés dans le vocodage des signaux à 4 canaux (autour de 600, 1400, 4000 Hz pour les 3 frontières intermédiaires) semblent assez proches des limites identifiées par Ueda & Nakajima (2017). À l'issue d'une analyse factorielle réalisée sur des signaux de parole issus de 8 langues différentes, Ueda & Nakajima (2017) déduisent 4 « bandes spectrales idéales » qui seraient *optimales* pour porter l'information acoustique et les 3 frontières identifiées par les auteurs sont respectivement 540, 1720 et 3300 Hz. Cette hypothèse devra être approfondie car elle pourrait fournir une piste essentielle pour l'étude de la transmission des informations nasales dans un implant.

Les résultats des analyses d'entropie mutuelle qui n'ont été que partiellement évoqués ici semblent plutôt suggérer que l'information de distinction orale / nasale associée aux voyelles est très fortement dégradée pour l'ensemble des conditions de résolution spectrale et temporelle. Il reste cependant que le *design* mis en œuvre ne permet pas de comparer à la fois l'analyse perceptive de la distinction orale / nasale et celle de la différenciation des nasales entre elles. En effet, les mesures d'entropie mutuelle associées à des paramètres comme l'arrondissement, l'aperture ou la position pour les voyelles intègrent nécessairement les résultats des voyelles orales. Nous prévoyons d'étudier spécifiquement le comportement des voyelles orales et nasales dans un cadre qui permettra d'obtenir des données pertinentes pour étudier cette question en mettant en place un design spécifiquement adapté. Par ailleurs, il semble essentiel de généraliser les résultats de cette étude à d'autres locuteurs et à des conditions de résolution spectrale plus fines.

Remerciements

Ce travail a reçu le soutien financier du Conseil Scientifique de l'Université de Nantes (Programme « Interdisciplinarités »).

Références

Baudonck N., Van Lierde K., D'Haeseleer E. & Dhooge I. (2015). Nasalance and nasality in children with cochlear implants and children with hearing aids. *International Journal of Pediatric Otorhinolaryngology*, **79**(4), 541–545.

Borel S. (2015). *Perception auditive, visuelle et audiovisuelle des voyelles nasales par les adultes devenus sourds. Lecture labiale, implant cochléaire, implant du tronc cérébral.* PhD thesis, Université de la Sorbonne Nouvelle – Paris 3.

Carignan C., Shosted R. K., Fu M., Liang Z.-P. & Sutton B. P. (2015). A real-time MRI investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel system of French. *Journal of Phonetics*, **50**, 34–51.

Christiansen T. U. & Greenberg S. (2010). Frequency selective filtering of the modulation spectrum and its impact on consonant identification. In *Linguistic Theory and Raw Sound*, volume 40 of *Copenhagen Studies in Language*, p. 119. Copenhagen, DK : Samfundslitteratur.

Christiansen T. U. & Greenberg S. (2012). Perceptual Confusions Among Consonants, Revisited : Cross-Spectral Integration of Phonetic-Feature Information and Consonant

Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, **20**(1), 147–161.

Delvaux V. (2012). *Les voyelles nasales du français : Aérodynamique, articulation, acoustique et perception.* Peter Lang, Éditions Scientifiques Internationales.

Demolin D., Delvaux V., Metens T. & Soquet A. (2003). Determination of velum opening for french nasal vowels by magnetic resonance imaging. *Journal of Voice*, **17**(4), 454–467.

Eaton J. W., Bateman D., Hauberg S. & Wehbring R. (2015). *GNU Octave version 4.0.0 manual : a high-level interactive language for numerical computations.* UK : Network Theory Limited. 2002.

Feng G. & Castelli E. (1996). Some acoustic features of nasal and nasalized vowels : A target for vowel nasalization. *The Journal of the Acoustical Society of America*, **99**(6), 3694–3706.

Fletcher S., Mahfuzh F. & Hendarmin H. (1999). Nasalence in the speech of children with normal hearing and children with hearing loss. *American Journal of Speech Language Pathology*, **8**, 241–248.

House A. S. & Stevens K. N. (1956). Analog studies of the nasalization of vowels. *Journal of Speech and Hearing Disorders*, **21**(2), 218–232.

Kanedera N., Arai T., Hermansky H. & Pavel M. (1999). On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Communication*, **28**, 43–55.

Maeda S. (1982). The role of the sinus cavities in the production of nasal vowels. In *ICASSP – IEEE International Conference on Acoustics Speech and Signal Processing*, volume 7, p. 911–914.

Miller G. A. & Nicely P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, **27**(2), 338–352.

Ming V. L. & Holt L. L. (2009). Efficient coding in human auditory perception. *The Journal of the Acoustical Society of America*, **126**(3), 1312–1320.

Moore B. C. J. & Glasberg B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, **74**, 750–753.

Rossato S. (2000). *Du son au geste, inversion de la parole : le cas des voyelles nasales.* Thèse de doctorat, Université Sthendal, Grenoble, France.

Shannon C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 399–423, 623–656.

Shannon R., Zeng F., Kamath V., Wygonski J. & Ekelid M. (1995). Speech recognition with primarily temporal cues. *Science*, **270**, 303–304.

Slaney M. (1993). *An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank.* Rapport interne 35, Apple Computer.

Stevens K. N. (1998). *Acoustic Phonetics.* Cambridge, Mass., USA : The MIT Press.

Ueda K. & Nakajima Y. (2017). An acoustic key to eight languages/dialects : Factor analyses of critical-band-filtered speech. *Scientific Reports*, **7**, 42468.

Van Tasell D., Soli S., Kirby V. & Widin G. (1987). Speech waveform envelope cues for consonant recognition. *The Journal of the Acoustical Society of America*, **77**, 1069–1077.