Mémoire de Master 2 Traitement Automatique des Langues

Parcours Recherche et Développement

Présenté par Lila KIM

Classification automatique de voyelles nasales pour une caractérisation de la qualité de voix des locuteurs : réseaux de neurones convolutifs, analyse acoustique et perception

Sous la direction de Cédric Gendrot et de Nicolas Audibert

Soutenu le 17 juin 2021 devant :

Angélique Amelot, Ingénieur de Recherche CNRS

Nicolas Audibert, Maître de Conférences, Université Sorbonne-Nouvelle

Didier Demolin, Professeur des Universités, Université Sorbonne-Nouvelle

Cécile Fougeron, Directeur de Recherche CNRS

Cédric Gendrot, Maître de Conférences, Université Sorbonne-Nouvelle

Martine Adda-Decker, Directeur de Recherche CNRS

Remerciements

Mes remerciements s'adressent particulièrement à mes directeurs de mémoire, Cédric Gendrot et Nicolas Audibert, pour leur accompagnement, leurs précieux conseils, leur patience, leur gentillesse et leurs aides inoubliables pendant l'élaboration de ce mémoire.

Je souhaite remercier Serge Fleury pour son accompagnement de tous les instants, ses encouragements et son soutien depuis la troisième année de Licence Sciences du Langages.

Je suis très reconnaissante à toutes les personnes qui ont participé au test de perception, en particulier les personnes sur la liste pluriTal et la liste du Laboratoire de Phonétique et Phonologie.

Je tiens par avance à remercier les membres du jury, Martine Adda-Decker, Cecile Fougeron, Cedric Gendrot, Nicolas Audibert, Angelique Amelot, Serge Fleury et Didier Demolin pour la lecture de mon mémoire et leur participation à la soutenance.

Table des matières

Table des matières **5**

Liste des figures **8**

Liste des tableaux **10**

Résumé **13**

Introduction **14**

État de l'art **15**

Acoustique des voyelles nasales 17

Physiologie des voyelles nasales 20

Qualité de voix nasale et caractérisation du locuteur 23

Apprentissage automatique et profond 26

Terminologie 27

Réseaux de neurones convolutifs 27

Couche de convolution 28

Couche de pooling 29

Couche dense 30

Corpus **32**

Méthodes **33**

Extraction des données 33

Sélection des échantillons 33

Création des spectrogrammes pour les réseaux de neurones convolutifs 35

Découpage en trois tronçons 37

Réseau de neurones convolutif 39

Présentation des systèmes CNN 39

Analyse en items 40

Analyse sur l'influence sur la classification 41

Contextes et durée 41
Fréquence fondamentale et intensité 43
Analyse sur la généralisation des CNNs 44
Analyse en mesures acoustiques 45
Analyse perceptive 46
Résultats et discussion **49**
Analyse en items 49
Segment 49
Tronçon 51
Locuteur 53
Influence sur la classification 57
Contexte 57
Sexe 59
Durée 62
Fréquence fondamentale 65
F0 au milieu de la voyelle 65
F0 en moyenne 67
Intensité 68
Intensité au milieu de la voyelle 69
Intensité en moyenne 70
Généralisation du modèle 72
Modèles entraînés et testés sur deux voyelles 72
Modèles entraînés avec deux voyelles et testés sur quatres voyelles 75
Modèles entraînés et testés sur quatre voyelles 77
Mesures acoustiques 79
H1c 79
H1A1c 81
Analyse perceptive 88
Test de perception 88
Analyse qualitative 93
Discussion 97

Interprétation des résultats pour la caractérisation des locuteurs 97
Validation acoustique et perceptive 103
Tentatives de généralisation et améliorations futures 106
Conclusion **108**
Annexes **110**
Bibliographie **111**

Liste des figures

Figure 1: Deux spectrogrammes illustrant plusieurs caractéristiques nasals dans les mots “gars” et “gant” notant le chevauchement de F1 dans la région attendue pour P1 dans "gars"' d'après Styler (2017) 14

Figure 2: Directions schématiques d'action des muscles impliqués dans le mouvement du velum et le port vélopharyngé d'après Serrurier (2008); 1: Tensor veli palatini; 2: Levator veli palatini; 3: Palatoglossus; 4: Palatopharyngeus; 5: Pharyngeal superior constrictor 16

Figure 3: Exemple d'un réseau de neurones comportant une couche d'entrée, une couche cachée et une couche de sortie d'après O’ Shea et Nash (2015) 22

Figure 4: Exemple de l'opération de convolution sur une image avec un filtre de 3x3 et un stride de 1 d'après Yamashita et al. (2018) 25

Figure 5: Exemple de l'opération de max-pooling avec un filtre de 2x2 et un stride de 2 d'après Yamashita et al. (2018) 26

Figure 6: Exemple de résultats obtenus après le passage dans une couche de pooling d'après Yamashita et al. (2018) 26

Figure 7: Fonction outer view port pour un son à 100 ms: 0, 2, 0, 3. 32

Figure 8: Fonction Paint sans axes 33

Figure 9: Zéro-padding sur le spectrogramme 33

Figure 10: Représentation visuelle de l'architecture de nos réseaux de neurones convolutifs 35

Figure 11: Résultat pour l'item "locuteur" 36

Figure 12: Résultat pour l'item "voyelle" (/a/, /ɑ̃/, /ɛ/ et /ɛ̃/ notés "a", "A", "E" et "I" respectivement dans le tableau) 37

Figure 13: Résultat pour l'item "tronçon" (1 pour première tiers, 2 pour deuxième tiers, 3 pour dernière tiers d'une image) 37

Figure 14: Analyse de probabilité d’appartenance de chaque occurrence 38

Figure 15: Fichier txt utilisé dans la phase d’extraction de f0 et intensité de chaque son 39

Figure 16: Analyse selon f0 et intensité de chaque occurrence 40

Figure 17: Paramétrage dans le logiciel VoiceSauce pour l’analyse en mesures acoustiques 41

Figure 18: Exemple de question du test de perception 43

Figure 19: TextGrid créé à l'aide d'une fonction "concatenate recoverably" 43

Figure 20: H1c selon la voyelle 76

Figure 21: H1c selon la classe 76

Figure 22: H1c selon la voyelle et la classe prédite 76

Figure 23: H1A1c selon la voyelle 77

Figure 24: H1A1c selon la classe prédite 77

Figure 25: H1A1c selon la voyelle et la classe prédite 78

Figure 26: H1c selon la voyelle et la classe prédite pour le locuteur 05_12_07_nb1_1_16 81

Figure 27: H1c selon la voyelle et la classe prédite pour le locuteur 28_11_07_nb2_1_16 81

Figure 28: H1c selon la voyelle et la classe prédite pour le locuteur 03_12_07_nb1_1_16 81

Figure 29: H1A1c selon la voyelle et la classe prédite pour le locuteur 05_12_07_nb1_1_16 82

Figure 30: H1A1c selon la voyelle et la classe prédite pour le locuteur 28_11_07_nb2_1_16 82

Figure 31: H1A1c selon la voyelle et la classe prédite pour le locuteur 03_12_07_nb1_1_16 82

Figure 32: H1c selon la voyelle et la classe prédite pour le locuteur 28_11_07_nb1_1_16 84

Figure 33: H1c selon la voyelle et la classe prédite pour le locuteur 29_11_07_nb1_2_16 84

Figure 34: H1c selon la voyelle et la classe prédite pour le locuteur 14_11_07_nb1_1_16 84

Figure 35: H1A1c selon la voyelle et la classe prédite pour le locuteur 28_11_07_nb1_1_16 85

Figure 36: H1A1c selon la voyelle et la classe prédite pour le locuteur 29_11_07_nb1_2_16 85

Figure 37: H1A1c selon la voyelle et la classe prédite pour le locuteur 14_11_07_nb1_1_16 85

Figure 25: Exemple du zéro-padding sur une partie du script (l'ensemble est disponible sur le GitHub) 108

Liste des tableaux

Table 1: Consonnes du français 15

Table 2: Voyelles du français (la colonne de droite pour chaque lieu d'articulation correspond aux contreparties arrondies 16

Table 3: Présentation des jeux de données 30

Table 4: Modèle entraîné avec /a/-/ɑ̃/ et testé sur /a/-/ɑ̃/ 46

Table 5: Modèle entraîné avec /a/-/ɑ̃/ et testé sur /a/-/ɑ̃/ et /ɛ/-/ɛ̃/ 46

Table 6: Modèle entraîné avec /a/-/ɑ̃/ et /ɛ/-/ɛ̃/, et testé sur /a/-/ɑ̃/ et /ɛ/-/ɛ̃/ 47

Table 7: Scores de classification obtenus par les trois modèles selon le tronçon 48

Table 8: Faux positifs (FP), vrais négatifs (VN) et faux négatifs (FN) pour les occurrences de voyelles /a/-/ɑ̃/ selon locuteur 51

Table 9: Nombre de faux négatifs selon le contexte 53

Table 10: Nombre de faux positifs selon le contexte 54

Table 11: Nombre de faux négatifs selon le contexte et le sexe pour les voyelles orales 55

Table 12: Nombre de faux positifs selon le contexte et le sexe pour les voyelles nasales 56

Table 13: Statistiques selon le sexe du locuteur 57

Table 14: Nombre de locuteurs selon la tendance d'erreurs calculée avec la durée 59

Table 15: Nombre de locuteurs selon la tendance d'erreurs calculée avec la durée (images découpées) 60

Table 16: Nombre de locuteurs selon la tendance d'erreurs calculée avec la f0 au milieu de la voyelle 61

Table 17: Nombre de locuteurs selon la tendance d'erreurs calculée avec la f0 au milieu de la voyelle (images découpées) 62

Table 18: Nombre de locuteurs selon la tendance d'erreurs calculée avec la f0 moyenne 63

Table 19: Nombre de locuteurs selon la tendance d'erreurs calculée avec la f0 moyenne (images découpées) 64

Table 20: Nombre de locuteurs selon la tendance d'erreurs calculée avec l'intensité au milieu de la voyelle 65

Table 21: Nombre de locuteurs selon la tendance d'erreurs calculée avec l'intensité au milieu de la voyelle (images découpées) 66

Table 22: Nombre de locuteurs selon la tendance d'erreurs calculée avec l'intensité moyenne 66

Table 23: Nombre de locuteurs selon la tendance d'erreurs calculée avec l'intensité moyenne (images découpées) 67

Table 24: Cinq locuteurs selon les modèles de test entraînés et testés sur deux voyelles 69

Table 25: Scores obtenus à partir du modèle normal et des modèles de test entraînés et testés sur deux voyelle 69

Table 26: Cinq locuteurs selon les modèles de test entraînés et testés sur deux voyelles (images découpées) 70

Table 27: Scores obtenus à partir du modèle normal et des modèles de test entraînés et testés sur deux voyelles (images découpées) 70

Table 28: Cinq locuteurs selon les modèles de test entraînés avec deux voyelles et testés sur quatre voyelles 71

Table 29: Scores obtenus à partir du modèle normal et des modèles de test entraînés avec deux voyelles et testés sur quatre voyelles 71

Table 30: Cinq locuteurs selon les modèles de test entraînés avec deux voyelles et testés sur quatre voyelles (images découpées) 72

Table 31: Scores obtenus à partir du modèle normal et des modèles de test entraînés avec deux voyelles découpées et testés sur quatre voyelles (images découpées) 72

Table 32: Cinq locuteurs selon les modèles de test entraînés et testés sur quatre voyelles 73

Table 33: Scores obtenus à partir du modèle normal et des modèles de test entraînés et testés sur quatre voyelles 73

Table 34: Cinq locuteurs selon les modèles de test entraînés et testés sur quatre voyelles (images découpées) 74

Table 35: Scores obtenus à partir du modèle normal et des modèles de test entraînés et testés sur quatre voyelles (images découpées) 74

Table 36: Récapitulatif des locuteurs qui ont produit les voyelles utilisées dans le test de perception 79

Table 37: Résultats pour deux voyelles orales correctes dans le test de perception 79

Table 38: Résultats pour deux voyelles orales incorrectes dans le test de perception 80

Table 39: Résultats pour deux voyelles nasales correctes dans le test de perception 81

Table 40: Résultats pour deux voyelles nasales incorrectes dans le test de perception 82

Table 41: Résultat pour deux locuteurs (V1 pour 04_12_07_nb2_2_16 et V2 pour 28_11_07_nb2_1_16) 83

Table 42: Résultat pour deux locuteurs (V1 pour 22_11_07_nb2_2_16 et V2 pour 29_11_07_nb1_2_16) 85

Table 43: Nombre de faux négatifs et de vrais positifs pour les occurrences de voyelles /a/-/ɑ̃/ selon neuf locuteurs 87

Table 44: Valeurs de probabilités pour les huit voyelles correctes du test de perception et toutes les voyelles 93

Table 45: Récapitulatif des résultats obtenus pour l'expérience de généralisation 94

Résumé

L'objectif de ce mémoire est d'évaluer, au moyen de réseaux de neurones convolutifs (CNN pour Convolutional Neural Networks), la capacité à discriminer acoustiquement une voyelle orale d'une voyelle nasale (voyelles /a/ et /ɑ̃/, mais aussi /ɛ/ et /ɛ̃/) sur un corpus de 45 locuteurs francophones, ainsi que la possibilité de généraliser cette discrimination sur des locuteurs et/ou des voyelles non entraînées au préalable. Pour une classification voyelle nasale vs. voyelle orale, les résultats obtenus pouvaient atteindre jusqu'à 97% d'identification correcte. Ces résultats ont été validés acoustiquement et perceptivement afin de montrer que certains locuteurs pouvaient être identifiés comme ayant une voix plus nasale que d'autres. Ces scores s'avèrent particulièrement élevés pour une analyse acoustique de la nasalité en comparaison d'autres mesures acoustiques obtenues dans la littérature (Chen, 1997 ; Lonchamp, 1988), ces résultats étant imputables à l'utilisation de réseaux de neurones profonds qui sont réputés pour leurs capacités d'extraction de paramètres (Hinton et al., 2012; Lozano-Diez et al., 2018), et à l'utilisation de grands corpus de parole segmentés au niveau phonémique avec une grande précision. Les limites de ce travail, ainsi que ses suites pour mettre en application sur l'ensemble de la parole sont abordées dans la discussion.

1. Introduction

Il existe très peu de langues où la nasalité n'est pas présente phonologiquement : 96% des langues du monde possèdent au minimum une consonne nasale. Elle se réalise par le couplage de deux cavités, nasale et buccale, qui a pour effet acoustique d'apporter des modifications sur les résonances des sons telles que l'introduction des résonances nasales sur des résonances orales existant ou la réduction de l'énergie dans le spectre, etc. Les sons émettent non seulement le contenu linguistique que le locuteur souhaite transmettre, mais également des informations sur le locuteur, tels que son humeur, ses intentions ou ses caractéristiques physiologiques.

Puisque le velum est un organe difficile à contrôler consciemment, et la morphologie de la cavité nasale qui est rigide et propre au locuteur, les nasales sont considérées comme contenant des informations pertinentes sur le locuteur.

Le domaine du traitement automatique de la parole est exploré depuis le XXième siècle. L'évolution de la machine et l'augmentation des quantité de données ont permis l'usage des réseaux de neurones profonds depuis les années 2010. Dans le cadre de la reconnaissance du locuteur, les réseaux de neurones convolutifs sont de plus en plus employés étant donné leur réputation pour leurs capacités d'extraction dans les images.

Les travaux de ce mémoire consistent à montrer et à évaluer la capacité à discriminer la nasalité dans les voyelles orales /a/ et /ɛ/ et les voyelles nasales /ɑ̃/ et /ɛ̃/, et à vérifier la généralisation des modèles sur cinq locuteurs ou des voyelles non vus par le système. Nous nous intéresserons également à l'influence des contextes phonémiques, du sexe, des caractéristiques acoustiques sur le taux de la classification du système, et à la validation des résultats obtenus.

Dans la première partie, nous nous pencherons sur les travaux existant non seulement dans le domaine phonétique sur l'acoustique et la physiologie des voyelles nasales, la qualité de voix et la caractérisation du locuteur, mais également dans le domaine du traitement automatique des langues sur la présentation des algorithmes employés dans notre étude. Dans la deuxième partie, nous décrirons le corpus utilisé, les méthodes d'extraction et les différentes analyses abordées. La dernière partie est dévolue par la présentation des résultats obtenus et l'interprétation de ces résultats.

2. État de l'art

La production de la parole est le résultat des processus distincts, qui peuvent être décrits selon trois niveaux : sous-glottique, glottique et supra-glottique. Ces systèmes peuvent s'expliquer par la position des organes comparée à la glotte. L'air est expulsé à travers des organes respiratoires différents, par exemple, des poumons, des bronches, et la trachée, faisant partie du système sous-glottique, puis se transforme en sons de la parole lorsqu'il atteint le niveau supra-glottique. Le système glottique, ou système phonatoire, correspond au larynx au sein duquel se trouvent les plis vocaux. C'est à ce niveau qu'interviennent les notions de fréquence fondamentale et de type de phonation. La fréquence fondamentale correspond au nombre de vibration des plis vocaux par seconde et est variable selon les locuteurs en fonction de facteurs physiologiques comme le sexe ou l'âge du locuteur, mais varie également en fonction de facteurs linguistiques tels que la position de la syllabe dans le mot, ou dans le groupe prosodique, etc. Les types de phonation varient en fonction du temps pendant lequel les plis vocaux sont ouverts, par exemple pour une voix soufflée, les plis vocaux seront ouverts entre 60% et 80% du cycle de vibration, alors que pour une voix craquée, les plis vocaux seront ouverts entre 30% et 45% du cycle de vibration. Enfin, au niveau supra-glottal, le flux d'air passe par le conduit vocal et se trouve altéré par les organes dits résonateurs constitués d'une cavité pharyngale, d'une cavité buccale et d'une cavité nasale. Le système d'articulation donne lieu à un contraste entre les consonnes et les voyelles. La production des consonnes peut s'expliquer par le mouvement de fermeture du conduit vocal avec quatre critères de classification des consonnes (Vaissière, 2015) : le mode de voisement, le lieu d'articulation, le degré de constriction (ou le mode d'articulation) et la nasalité.

	Bilabiales	Labiodentales	Dentales	Alvéolaires	Post-alvéolaires	Palatales	Vélaires	Uvulaires
Occlusives	p b		t d				k g	
Nasales	m		n			ɲ	ŋ	
Fricatives		f v	s z		ʃ ʒ			ʁ
Approximants						j		
Approximants latérales			l					

Table 1: Consonnes du français

La réalisation d'une consonne est dûe aux organes mobiles comme les lèvres, la langue ou la glotte qui empêchent complètement ou légèrement le passage de l'air dans le conduit vocal. Les voyelles, quant à elles, correspondent au mouvement d'ouverture du conduit vocal et se distinguent par l'aperture (ou l'ouverture) du conduit vocal, le degré d'antériorité de la langue et l'étirement des lèvres. Aucun obstacle n'est observé dans le passage de l'air et les plis vocaux vibrent presque toujours lors de la réalisation d'une voyelle. Dans la phase de l'articulation, le comportement du voile du palais permet de distinguer les voyelles orales des voyelles nasales

	Antérieure		Centrale		Postérieure	
Fermée	i	y				u
Mi-fermée	e	ø				o
Mi-ouverte	ɛ ɛ̃	œ œ̃				ɔ ɔ̃
Ouverte			a		ɑ ɑ̃	

Table 2: Voyelles du français (la colonne de droite pour chaque lieu d'articulation correspond aux contreparties arrondies)

Un son oral est réalisé dans la cavité buccale, l'air ne peut passer par la cavité nasale car le voile du palais est relevé. Contrairement au son oral, le voile du palais est abaissé lors de la production d'un son nasal et l'air sort donc par la cavité nasale s'il s'agit d'une consonne nasale, et à la fois par la cavité orale et nasale s'il s'agit d'une voyelle nasale.

En français, on compte /ɑ̃/, /ɛ̃/, et /ɔ̃/ qui sont les contre-parties nasales des voyelles orales /a/, /ɛ/ et /o/. Notons que les trois voyelles nasales ont des différences d'abaissement du voile du palais. Pour la voyelle /ɑ̃/, la langue se rapproche de la paroi pharyngée et le voile du palais est relativement moins élevé que la voyelle /ɛ̃/ pour laquelle le palais mou est le plus baissé parmi d'autre voyelles nasales avec la langue abaissée en son centre et éloignée de la paroi. La voyelle /ɔ̃/, quant à elle, le dos de la langue se positionne très proche du voile du palais à quel point qu'ils peuvent se superposer (Amelot, 2004).

Chaque contrepartie nasale a également des différences articulatoires vis-à-vis de la voyelle orale correspondante (Zerling). Ces différences vont amplifier les problèmes d'analyse acoustique que nous mentionnerons ci-dessous. Notons que la voyelle /œ̃/ a été laissée de côté dans cette étude car elle est le plus souvent par les systèmes d'alignement de la parole, qui la fusionnent avec /ɛ̃/.

2.1. Acoustique des voyelles nasales

La nasalité est utilisée pendant la production de parole, pour la distinction entre nasales et orales, que ce soit pour les voyelles (un /a/ et un /ɑ̃/ par exemple) ou bien les consonnes (un /b/ et un /m/ par exemple). Dans les langues comme le français où l'abaissement du voile du palais permet de distinguer les nasales et les orales et où les voyelles nasales sont considérées comme des phonèmes distincts, l'acte de communication peut être introduit par la production et la perception de la nasalité (Styler, 2017). Les voyelles nasales peuvent être également présentes dans les langues pour lesquelles elles ne sont pas des phonèmes, comme l'anglais par la notion de coarticulation nasale.

Le conduit nasal a une structure très complexe et il produit des effets acoustiques sur les sons nasals (Havel, 2016); certains auteurs disent qu'aucune relation n'est observée entre la voix et les sinus paranasaux qui ne peuvent pas fonctionner en tant que cavités de résonances (Amelot, 2004 ; Havel, 2016), d'autres estiment que les effets acoustiques constants et significatifs de la nasalité proviennent de la morphologie complexe des sinus paranasaux (Dang, 1996), des cavités asymétriques, qui fonctionnent comme absorbeurs acoustiques en baissant l'énergie dans de différentes régions de fréquences (Havel, 2016).

La nasalité s'effectue par un abaissement du voile du palais qui permet au flux d'air de traverser la cavité nasale, en plus de la cavité orale. Cette double cavité dans laquelle l'air vient résonner a pour effet acoustique d'apporter des modifications sur les voyelles nasales. La littérature distingue ces changements en quatre types : 1) introduction des résonances nasales; 2) interférence des résonances nasales avec des résonances orales; 3) changement au niveau de la structure globale des formants vocaliques; 4) changement de l'enveloppe spectrale des voyelles.

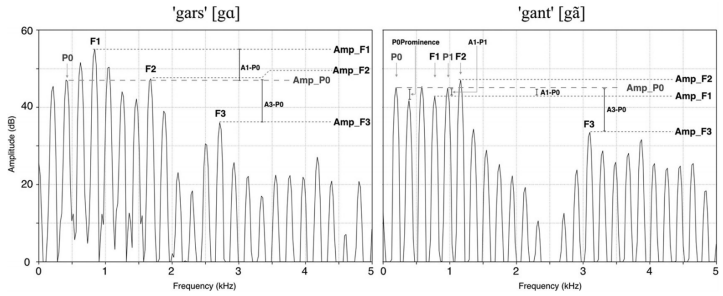


Figure 1: Deux spectrogrammes illustrant plusieurs caractéristiques nasals dans les mots “gars” et “gant” notant le chevauchement de F1 dans la région attendue pour P1 dans "gars"' d'après Styler (2017)

Le premier changement consiste à ajouter de nouvelles résonances au signal de parole en améliorant relativement les harmoniques de certaines régions, ce qui affecte les résonances orales existant et entraîne des formants avec le timbre de nasalité ou « nasal poles ». La détection de P0, un pôle nasal visible entre 250 Hz et 450 Hz avec une amplitude allant de 3 dB jusqu’à 5,5 dB, peut être difficile dans le cas des voyelles fermées pour lesquelles le premier formant se trouve dans la même région que P0 selon l’expérience de Chen. Ce pôle est généralement mesuré au moyen de la relation avec le premier harmonique du premier formant qui est également affecté par la nasalité, et dans le cas du chevauchement de P0 et A1 (comme les voyelles fermées), un deuxième pôle nasal, P1, situé entre 790 Hz et 1100 Hz peut être utilisé au lieu de P0 pour la comparaison avec A1. Pourtant, la détection de P1, qu’elle soit manuelle ou automatique, peut être plus difficile que P0 car P1 est sensible à l’interférence du premier formant et aussi du deuxième formant, et à la variation des positions de ces formants. P2, un autre pôle nasal situé environ 1250 Hz, est repérable dans le plus haut harmonique autour de 1250 Hz dans les voyelles nasales (Styler, 2017).

On parle de « zéros » ou « d’anti-formants » qui font baisser l’énergie dans le spectre au lieu de les renforcer comme le font les résonances de la cavité orale (Maeda, 1982). Cette caractéristique de la nasalité qui réduit l’amplitude de façon relative aux autres fréquences se trouve dans la région du premier, du deuxième et du troisième formant et peut intervenir en parallèle les pôles nasals (Styler, 2017).

Le changement des formants est observé dans les voyelles nasales car l’articulation orale est relativement modifiée lors de la réalisation des voyelles nasales en comparaison avec les

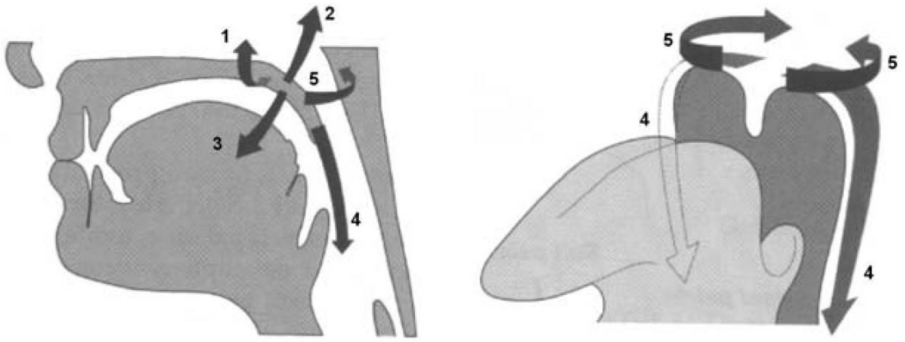
voyelles orales. Il peut être également dû aux pôles nasals et aux zéros modifiant l'amplitude près des formants. Enfin les formants nasals et les anti-résonances provoquent la baisse de l'amplitude des harmoniques les plus élevés de la voyelle (Styler, 2017).

Comme dit ci-dessus, les anti-formants vont rendre difficiles les analyses acoustiques de la nasalité. Les analyses acoustiques sont rendues d'autant plus difficiles que l'abaissement du voile du palais peut également avoir pour conséquence d'altérer la forme de la cavité orale et ainsi de modifier les formants (Carignan, 2017). A partir de la modélisation acoustique et de la compréhension de ces anti-formants, des analyses mesurant la différence d'amplitude entre les formants et les anti-formants ont été entreprises (Chen, 1997 ; Styler, 2017) pour catégoriser la nasalité des sons orals et nasals de manière automatique (Pruthi et Espy-Wilson, 2007). Néanmoins, des éléments segmentaux (le degré d'aperture des voyelles, le contexte phonétique) et non-segmentaux (le stress, la prosodie, les caractéristiques du locuteur et de la langue étudiée) peuvent donner lieu à des variations du degré de la nasalité, ces mesures acoustiques deviennent complexes et ne peuvent être effectuées que dans des conditions très spécifiques où tous les contextes sont contrôlés (Yuan et al, 1974 ; Styler W. 2017).

La nasalité est présente dans 96% des langues du monde. Elle se réalise par le couplage de deux cavités, nasale et buccale, qui a pour effet acoustique d'apporter des modifications sur les résonances des sons telles que l'introduction des résonances nasales sur des résonances orales existant ou la baisse de l'énergie dans le spectre, etc. Ces changements rendent les analyses acoustiques difficiles et complexes, qu'elles soient manuelles ou automatiques car ils sont très sensibles à diverses variations telles que l'accentuation, l'articulation du son, la langue, etc.

2.2. Physiologie des voyelles nasales

La cavité nasale est un organe relativement rigide en comparaison des autres cavités (Honda et al., 1994 ; Rose, 2000 ; Amelot, 2004). Le port vélopharyngé a un rôle pour connecter deux cavités, nasale et buccale, et l'ouverture ou la fermeture s'effectue par plusieurs composantes : 1) principalement par le palais mou constitué de cinq muscles différents pour la partie antérieure; 2) la paroi latérale (lateral pharyngeal wall) qui s'occupe de la partie latérale; 3) la paroi postérieure (posterior pharyngeal wall) s'occupant de la partie postérieure; 4) la langue (Serrurier, 2006). Les trois premiers organes correspondent au sphincter vélopharyngé qui fonctionne comme une valve de muscles en trois dimensions (Amelot et al., 2013). Ce mécanisme, qui diffère d'un locuteur à l'autre, nécessite la coordination des mouvements articulatoires de différents muscles (Serrurier et al., 2005).



a) Sagittal view b) Oblique anterior view

Figure 2: Directions schématiques d'action des muscles impliqués dans le mouvement du velum et le port vélopharyngé d'après Serrurier (2008); 1: Tensor veli palatini; 2: Levator veli palatini; 3: Palatoglossus; 4: Palatopharyngeus; 5: Pharyngeal superior constrictor

Le voile du palais est constitué d'une dizaine de muscles ; parmi divers muscles du velum responsables du mécanisme d'ouverture et de fermeture[1], deux muscles considérés comme les plus impliqués dans le mouvement du velum peuvent être repérés :

1. le muscle appelé "levator veli palatini" qui fonctionne comme élévateur du velum en relevant la région médiale du velum vers les trompes d'Eustaches ;

[1] Selon Amelot, Levator palatini, péristaphylin externe (tensor veli), pharyngo-staphylin, palato-glosse, palato-staphylin, fibre accessoire (muscle occipito-staphylin) (Amelot, 2004) ;
D'après Serrurier, Levator veli palatini, tensor veli palatini, uvulae muscle, palatoglossus muscle (palato-glosse), palatopharyngeus muscle (Serrurier, 2006) ;

2. le palato-glosse (ou glosso-staphylin), muscle du pilier antérieur du voile, assure la connexion entre la langue et le velum et est capable non seulement d'abaisser le velum, mais aussi d'abaisser et d'augmenter la base de la langue (Serrurier, 2006 ; Amelot, 2004).

Dans le conduit nasal, sont également présents les sinus paranasaux qui représentent un ensemble d'espaces bilatéraux et asymétriques remplis d'air dans le crâne facial (Dang, 1996 ; Havel, 2016). L'ouverture d'un organe comme l'ostium permet la communication entre la cavité nasale et les cavités paranasales constituées de quatre sinus : maxillaire, ethmoïdal, frontal, et sphénoïdal (Dang, 1996 ; Amelot, 2004). Bien que l'accessibilité faible des cavités paranasales rende les analyses difficiles, on s'attend à ce que ces sinus participent à la contribution des effets de nasalité en introduisant les zéros dans les résonances (Havel, 2016). Plusieurs chercheurs ont remarqué l'existence des anti-résonances ; par exemple, concernant l'anti-résonance pour le sinus maxillaire, Koyama l'observe entre 400 Hz et 1000 Hz en 1966, Lindqvist et al., entre 200 Hz et 800 Hz en 1976, Maeda autour de 450 Hz en 1982, etc. Cette variation peut être due aux conditions d'enregistrements différentes selon les expériences, aux conditions instrumentales (par exemple, l'absence de muqueuse dans le spécimen du locuteur) ou à la variation entre locuteurs (Dang et Honda, 1996).

Deux grandes classes de sons nasals existent : 1) L'écoulement d'air ne s'effectue qu'à travers la cavité nasale (Yuan et al, 1974), 2) les sons produits avec le voile du palais abaissé, mais sans obstruction dans la cavité buccale. Les sons de la première classe étant par exemple /n/ ou /m/ pour lesquels la fermeture de la cavité buccale est observée. Contrairement aux sons de la première classe, ceux de la seconde se produisent à travers la cavité buccale en couplage de la cavité nasale, comme par exemple pour les voyelles nasales (Dang et al., 2016). La cavité orale étant le passage majoritaire de l'air, les timbres de nasalité s'ajoutent aux sons de la cavité buccale (Pruthi et al, 2007).

Les particularités physiologiques du locuteur et leurs habitudes dans l'articulation influencent les sons au cours de la production de la parole (Amino, 2009). Surtout, les nasales se trouvent plus dépendantes des propriétés de locuteurs car la cavité nasale a une morphologie différente selon le locuteur et cette variation interlocuteur résulte la forme de la résonance propre au locuteur (Amelot, 2004).

Comme dit dans la section précédente, l'accès vers la cavité nasale se fait par l'abaissement du voile du palais, et une séquence phonémique composée d'une consonne nasale suivie d'une voyelle orale (par exemple [ma]) impliquera une remontée du voile du palais à la fin du [m] pour fermer l'accès à la cavité nasale et réaliser la voyelle orale [a]. Si la remontée du voile du palais prend plus de temps que prévu, elle génèrera des phénomènes de coarticulation qui pourront aboutir à des assimilations de nasalité ([mɑ̃]). Le voile du palais étant un organe peu mobile - environ 50 ms pour le mouvement d'abaissement (Amelot, 2004) -, ces phénomènes ne sont pas rares, et la nasalité est fréquemment citée dans les cas de changements diachroniques (apparition et disparition des voyelles nasales dans la langue) : ces évolutions sont généralement le fait de variations inter-locuteurs (Ohala, 1993) dans leur faculté à abaisser cet articulateur plus ou moins rapidement. Notons que le français se prête particulièrement à cet exemple de la coarticulation nasale puisqu'il possède à la fois voyelles et consonnes nasales (seuls 22% des langues possèdent cette spécificité (Maddieson et Abramson, 1987) et que la différence entre voyelle nasale et voyelle nasalisée par coarticulation doit théoriquement être maintenue (Vaissière, 1988).

Pour finir, il n'est pas rare que la fermeture de la cavité nasale soit incomplète chez certains locuteurs, ce qui génère une voix en moyenne plus nasale que celles des locuteurs parvenant à effectuer une fermeture complète (Basset et al., 2001). Une étude de la nasalité en tant que caractéristique de locuteur, y compris dans le cadre de la coarticulation prend donc tout son sens ici.

La coordination des mouvements articulatoires du voile du palais et des parois pharyngées résulte en un mécanisme d'ouverture et de fermeture du port vélopharyngé qui a un rôle d'assurer la connexion entre la cavité buccale et la cavité nasale. Cette dernière étant relativement rigide et présentant une morphologie propre au locuteur, les variations interlocuteurs peuvent être abordées par l'étude de la nasalité dans la parole.

2.3. Qualité de voix nasale et caractérisation du locuteur

Lors de la production de parole, un son émis porte d'une part une information linguistique que le locuteur souhaite transmettre au destinataire, et d'autre part une information sur le locuteur tels que les traits physiologiques et comportementaux (Denes et Pinson, 1993 ; Suzuki et al., 1990). Puisque le locuteur montre ses caractéristiques à travers tous les sons émis de manière permanente, les humains sont capables d'approcher l'identité du locuteur seulement par les sons de la parole en recevant d'une part une information sur le contenu linguistique, d'autre part sur l'identité du locuteur (Amino et al., 2008) et que la qualité de voix, un des composants non segmentaux, peut être idiosyncratique et caractériser l'individualité du locuteur (Abercrombie, 1967). Le terme "qualité de voix" peut impliquer plusieurs notions (Nolan, 2007) tels que le type de phonation, la voix tendue/relâchée ou la nasalité. Par exemple, le type de phonation représente la vibration des plis vocaux comme dans la voix craquée, ou la voix soufflée. Nous pouvons distinguer la qualité de voix en deux modes par la notion de contrôle : 1) sous le contrôle du locuteur comme le chuchotement, 2) hors contrôle comme effet déterminé par la physiologie d'un locuteur.

Parmi différents types de qualité de voix, nous souhaitons nous intéresser à la nasalité qui est déterminé essentiellement par la morphologie des cavités propres à un locuteur et qui est connue comme un des indices pour la caractérisation du locuteur. La caractérisation du locuteur par la nasalité peut être envisagée dans le cadre de la reconnaissance du locuteur dans le but de créer des interactions entre homme et machines (Kahn, 2011). Notons que la reconnaissance du locuteur dans un cadre criminalistique (« phonétique forensique ») nécessite également une compréhension des traits utiles à la caractérisation audio d'un locuteur et compréhensibles par un jury non expert, par exemple, dans le but de déterminer si deux enregistrements distincts ont été produits par le même locuteur (Rose, 2000) et de reconnaître une personne parmi les suspects au moyen d'un enregistrement de parole (Kahn, 2011). Ces travaux peuvent également être utilisés en clinique pour détecter acoustiquement des productions anormalement nasales, d'ailleurs, les études sur la transcription automatique de la parole pathologique ont été abordées pour donner le feed-back aux locuteurs dysarthriques dans son amélioration articulatoire (Laaridh, 2017).

La reconnaissance de la parole date des années 1950 n'a pas connu une grande évolution, ce n'est qu'à partir de la fin des années 1960 que les systèmes ont pu traiter les paroles continues

(Gelly, 2017). Les réseaux de neurones existaient déjà dans les années 1980, mais sans être efficaces car ils nécessitent du grand corpus de parole déjà transcrit et de la capacité de la machine pour entraîner le système. Dans les années 2000, les compagnies d'évaluation des performances des systèmes de traitement automatique de la parole ont été proposées et organisées par le National Institute of Standard and Technology permettent aux modèles de devenir plus performants en continuant à diminuer le taux d'erreurs et à progresser (Kahn, 2011 ; Gelly, 2017). Depuis 2010, la plupart des systèmes de reconnaissance de parole se basent sur le réseau de neurones grâce à l'évolution des ordinateurs puissants et permettant d'utiliser les méthodes très lourdes comme les réseaux de neurones profonds, mais aussi à l'augmentation des données. Plusieurs études ont été abordées sur la paramétrisation telles que les analyses cepstrales, ou le niveau lexical, etc. pour reconnaître le locuteur (Kahn, 2011).

Les réseaux de neurones ont essayé de se référer à l'humain, comme pour les neurones, leur comportement attendu était celui des neurones du cerveau humain. Une autre référence à l'humain est le caractère écologique ; les systèmes de reconnaissance peuvent obtenir des résultats spectaculaires mais nécessitent beaucoup de temps de calcul alors que l'humain peut reconnaître les locuteurs dès l'enfance, par exemple, un enfant arrivant à caractériser et à reconnaître les personnes à partir d'un audio (Kahn, 2011). Mais lorsque l'humain ne connaît pas les locuteurs, ses capacités à reconnaître les locuteurs diminuent. Surpassant déjà les capacités humaines de reconnaissance du locuteur, le domaine de la reconnaissance automatique des locuteurs ne s'arrête pas à progresser en essayant de trouver les indices idiosyncratiques les plus pertinents que les systèmes peuvent utiliser au cours de l'entraînement, par exemple, la fréquence fondamentale, l'information sur le locuteur, le lexique utilisé par le locuteur, etc (Kahn, 2011).

Dans le cadre de la reconnaissance automatique du locuteur, il a fréquemment été observé (Kahn et al., 2011) que les nasales sont plus pertinentes pour caractériser le locuteur que d'autres phonèmes. L'explication apportée est que la différence des caractéristiques acoustiques entre locuteurs est due à la morphologie différente des cavités d'un locuteur à l'autre et au mouvement du velum étant difficile à contrôler de manière intentionnelle ; puisque les humains ne peuvent modifier les cavités nasales facilement ou volontairement, il n'est pas fréquent que les caractéristiques acoustiques des sons nasals changent (Amino, 2006). Au contraire, le

changement au niveau du larynx est facilement réalisable par les locuteurs qui souhaitent déguiser leurs voix, et induisent les erreurs dans les tâches de l'identification du locuteur (Animo, 2008). L'efficacité de l'utilisation des phonèmes nasals pour caractériser les locuteurs ont été prouvés par des tests de perception dans lesquels les humains reconnaissent mieux la voix des locuteurs familiers avec eux lorsqu'ils entendent les nasales que les orales (Amino, 2006). En plus des nasales, les voyelles sont aussi considérées comme un indice non seulement pour les humains mais aussi pour les machines pour identifier les locuteurs (Amino, 2005). Cela implique que les voyelles nasales tendent à contenir les informations sur le locuteur en comparaison avec les orales.

En outre, la nasalisation peut intervenir aux trois niveaux : linguistique, paralinguistique et extralinguistique (Rose, 2000). D'un point de vue linguistique, un contraste binaire est observé entre les orales et les nasales à travers le mécanisme vélopharyngien qui est un séparateur entre la cavité nasale et d'autres (Vaissière, 1995) et a été illustré à l'aide d'un exemple en français parisien : /o/ dans "beau" et /õ/ dans "bon" (Styler, 2017). La nasalité a une fonction paralinguistique qui s'explique par le fait qu'un locuteur de statut inférieur emploie une nasalité pour s'adresser à un destinataire de statut supérieur dans une langue bolivienne Cayuvana (Rose, 2000). Enfin, au niveau extralinguistique, "twang nasal" est utilisé comme caractéristique par certains locuteurs, et ces derniers sont relativement plus faciles à identifier dans les tâches de reconnaissance du locuteur (Rose, 2000).

Les sons émettent non seulement le contenu linguistique que le locuteur souhaite transmettre, mais également des informations sur le locuteur ; c'est pourquoi les humains peuvent approcher l'identité du locuteur seulement par les sons de la parole. La nasalité est considérée dans notre étude comme une des facettes de la qualité de voix du locuteur au même titre que le type de phonation. Notons que les humains sont souvent peu qualifiés pour décrire et discriminer la qualité de voix, mais nous pensons qu'une analyse acoustique avec un apprentissage machine ne possède pas ces limites.

2.4. Apprentissage automatique et profond

Le domaine du traitement automatique de la parole est exploré depuis le XXième siècle, l'une des premières études étant portée sur l'identification des chiffres prononcés en 1952. L'usage des réseaux de neurones est observé depuis les années 1980, et celui des réseaux de neurones profonds a explosé dans les années 2010 grâce à l'évolution des machines comme GPU (Gelly, 2017) et des quantités de données. Les réseaux de neurones profonds inspirés par le comportement des neurones du cerveau humain représentent un ensemble de neurones ou perceptrons interconnectés, et dans chaque neurone se trouvent la couche d'entrée, des couches cachées et la couche finale (O' Shea et Nash, 2015).

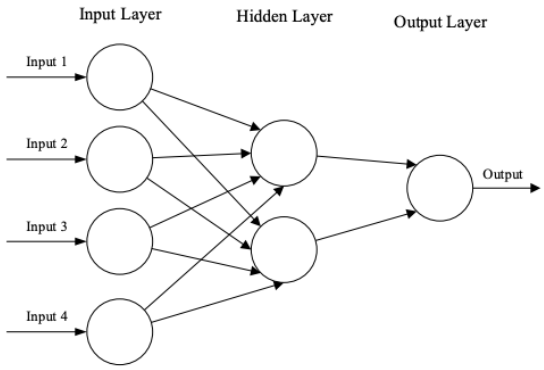


Figure 3: Exemple d'un réseau de neurones comportant une couche d'entrée, une couche cachée et une couche de sortie d'après O' Shea et Nash (2015)

L'entraînement consiste à extraire les paramètres à partir des données d'entrée et à prendre la décision au moyen de différentes couches et des fonctions (comme activation, loss, optimizer) existant dans un neurone (O' Shea et Nash, 2015). Les paramètres choisis de façon aléatoire permettent aux réseaux d'évoluer durant l'entraînement de manière à généraliser les variations entre les sorties (Yamashita et al., 2018).

Les réseaux de neurones profonds sont réputés pour leurs capacités phénoménales d'extraction de paramètres (Hinton et al., 2012 ; Lozano-Diez et al., 2018), et, parmi eux, les réseaux de neurones convolutifs sont considérées les plus adaptés pour traiter des données ayant un motif de grille comme les images depuis que leurs scores se sont avérés très élevés en reconnaissance de l'objet au concours intitulé "ImageNet Large Scale Visual Recognition Competition (ILSVRC)"

en 2012 (Yamashita et al., 2018). Comme les réseaux de neurones convolutifs ont recours à des images comme le type de données d'entrée, dans le domaine de la reconnaissance de la parole, les sons sont utilisés sous forme de spectrogramme.

2.4.1. Terminologie

Nous souhaitons introduire une explication des quelques termes clé pour mieux comprendre les prochaines parties.

- Paramètres : une variable qui est automatiquement apprise durant l'entraînement du réseau (O' Shea et Nash, 2015).
- Filtre (Kernel) : un ensemble de paramètres que le réseau peut apprendre (Yamashita et al., 2018).
- Activation : une transformation de la somme pondérée par une fonction non-linéaire, par exemple, la fonction d'activation "Relu" (The rectified linear activation function) qui transforme toutes les valeurs négatives en zéro et qui rendent les valeurs telles qu'elles lorsque ces dernières sont supérieures à zéro (Maas et al., 2013).
- Sur-apprentissage (Overfitting) : un modèle trop complexe a appris des paramètres non pertinents et spécifiques à un ensemble de données, par conséquent, la généralisation des réseaux se réduit sur un autre ensemble de données (O' Shea et Nash, 2015). L'augmentation des données par la transformation telle que la translation, l'effacement aléatoire peut être utile pour éviter le phénomène de sur-apprentissage (Yamashita et al., 2018).

2.4.2. Réseaux de neurones convolutifs

Comme mentionnés ci-dessus, les réseaux de neurones convolutifs (Convolutional Neural Networks) font partie des apprentissage automatique et profond (deep learning) et sont adaptés aux tâches de classification des images en permettant de déterminer les paramètres de manière aléatoire et en les optimisant progressivement à travers les différentes couches existant dans l'architecture. Il s'agit d'un apprentissage supervisé, les réseaux de neurones de convolution

nécessitent d'être fournis des étiquettes ou des réponses sur les données d'entrée ; cette méthode d'entraînement permettant de minimiser les erreurs de classification (O' Shea et Nash, 2015). Les réseaux de neurones convolutifs se différencient d'autres méthodes par ses caractéristiques spécifiques : ils n'ont pas recours à l'extraction des paramètres de la part des humains contrairement aux machines learning, mais plutôt à l'ensemble de données très quantitatives pour chercher eux-mêmes les paramètres qui peuvent être apprises (Yamashita et al., 2018).

Trois types de couches sont globalement observés dans l'architecture des réseaux de neurones convolutifs : couche de convolution, couche de pooling et couche dense (ou couche de fully-connected). Les couches de convolution et de pooling dans lesquelles les réseaux sont interconnectés localement peuvent être répétées et les couches denses où les réseaux sont interconnectés globalement viennent à la fin de l'entraînement. Le résultat obtenu dans une couche est utilisé en tant que l'entrée pour la prochaine couche (Hinton et al., 2012), et à chaque passage dans un neurone, la non-linéarité montrant le comportement similaire à la perception humaine (Wieser, 2018 ; Trigeorgis et al., 2016) est introduite au moyen de la fonction d'activation. La dernière couche de convolution ou de pooling permet d'obtenir un feature map à une dimension que le réseau a gardé comme modèle de l'image en entrée.

2.4.3. Couche de convolution

La couche de convolution permet l'extraction des traits ou des caractéristiques dans une image en optimisant la taille de l'image pour réduire la complexité du modèle (O' Shea et Nash, 2015). Dans cette couche visant à obtenir un ensemble des caractéristiques aussi appelé comme feature map, un filtre ou un kernel qui est un array des nombres traverse tout le long d'une image avec un pas (stride) défini (Hinton et al., 2012). Le filtre de couche convolutive est souvent de 3x3 (trois pixels en hauteur et en largeur) avec le stride d'un pixel, ce qui fait le chevauchement entre les régions de l'image (aussi appelées "tensor" selon Yamashita et al., 2018). L'application du filtre s'effectue sur chaque région de l'image qui est également un array des nombres de la même taille que le filtre (O' Shea et Nash, 2015). Une caractéristique est calculée en multipliant des valeurs élément par élément du filtre et du tensor et l'addition de ces dernières comme l'illustre la figure 4 présentée ci-après. Cette méthode peut risquer de perdre les informations sur les

bords, le zéro-padding permettant d'entourer l'image par des suites de zéros peut être considérable pour contourner l'éventuel problème (Yamashita et al., 2018).

Le fait que le même filtre passe sur toutes les régions de l'image introduit la notion de "partage des paramètres", "Parameters sharing" selon (O' Shea et Nash, 2015) ou "Weight sharing" selon (Yamashita, 2018). Cette notion montre qu'un paramètre utile dans une région de l'image est susceptible d'être également utile dans une autre région en permettant de rendre le réseau invariant à la translation et d'augmenter l'efficacité du modèle par la réduction des paramètres à apprendre.

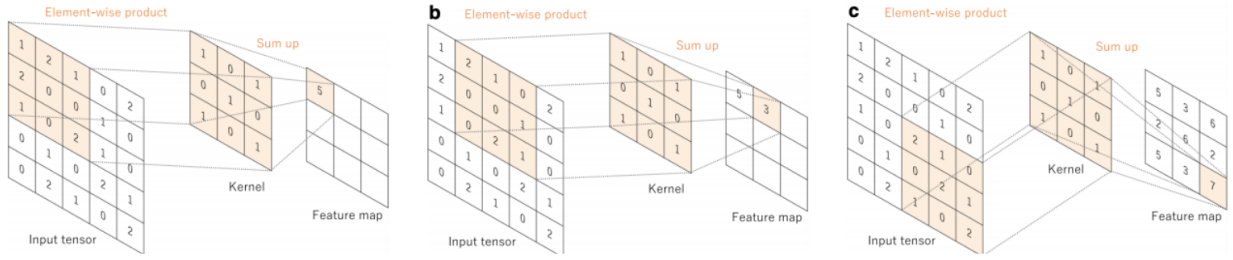


Figure 4: Exemple de l'opération de convolution sur une image avec un filtre de 3x3 et un stride de 1 d'après Yamashita et al. (2018)

2.4.4. Couche de pooling

La couche de pooling cherche à transformer les représentations en plus petite taille et à réduire le nombre des paramètres que le modèle peut apprendre (O' Shea et Nash, 2015). Elle permet également la recentration ou la localisation de l'élément important dans une image pour que le modèle soit résistant à la translation.

Parmi les diverses opérations de pooling, max-pooling est le plus généralement utilisé dans la pratique et permet de conserver la valeur la plus élevée pour la prochaine couche (Hinton et al., 2012). Le kernel de 2x2 avec un pas de deux pixels est couramment utilisé sur l'ensemble des caractéristiques (feature map) obtenu dans le cadre d'une couche de convolution précédente, ce qui n'implique pas de chevauchement entre les régions en entrée (O' Shea et Nash, 2015).

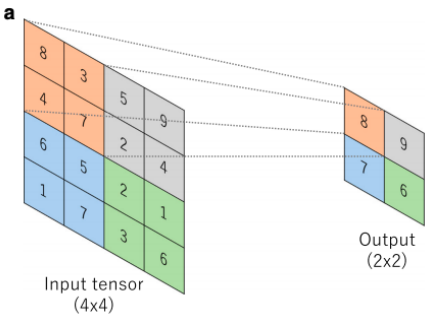


Figure 5: Exemple de l'opération de max-pooling avec un filtre de 2x2 et un stride de 2 d'après Yamashita et al. (2018)

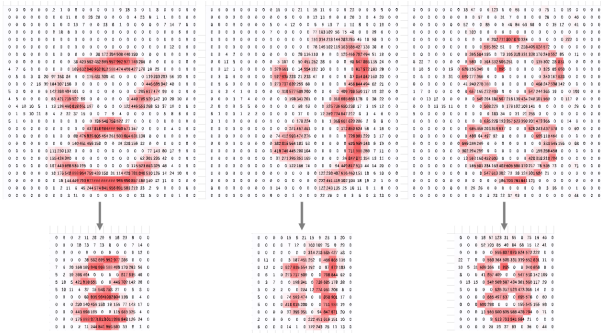


Figure 6: Exemple de résultats obtenus après le passage dans une couche de pooling d'après Yamashita et al. (2018)

2.4.5. Couche dense

Après l'extraction des paramètres dans les couches de convolution et le sous-échantillonnage dans les couches de pooling, les couches denses sont mises en œuvre avec pour objectif de chercher l'information et de donner un score tel que les probabilités d'appartenance selon chaque classe (voir la figure 5). Dans ces couches denses (ou fully-connected) qui sont analogues du réseau de neurones artificiel traditionnel, les neurones sont directement connectés entre eux par le poids désigné, ce qui fait augmenter le coût de calcul (O' Shea et Nash, 2015 ; Yamashita et al., 2018). Dans le but de fournir les informations déjà simplifiées aux couches denses et de réduire le temps de calcul, les couches denses se positionnent généralement à la fin de l'entraînement.

La non-linéarité introduite dans les couches peut rendre le modèle très expressif. Pour éviter le sur-apprentissage du réseau, une technique appelée "Dropout" peut être envisagée visant à "oublier" ou à remettre le poids à zéro pour un certain nombre de neurones choisis aléatoirement (Hinton, 2012 ; Srivastava, 2013).

Les réseaux de neurones convolutifs ont la réputation d'être très performants dans leurs capacités d'extraction dans les images, l'utilisation des données audio s'effectue donc sous forme de spectrogramme dans le cadre de la reconnaissance de la parole. Les couches de convolution et de pooling permettent de réduire le nombre de paramètres que le modèle peut apprendre et de localiser les éléments pertinents, ce qui rend le modèle invariant à la translation et les couches denses permettent d'obtenir les scores.

3. Corpus

La collecte des données du corpus utilisées dans le cadre de nos travaux a été principalement effectuée au cours de mon stage réalisé au sein du Laboratoire de Phonétique et de Phonologie à Paris (UMR 7018, CNRS et Université Sorbonne Nouvelle) sous la direction de C. Gendrot. Le corpus utilisé dans notre recherche est intitulé NCCFr (The Nijmegen Corpus of Casual French) et comporte au total plus de 36 heures de conversation amicale de 46 locuteurs français. La réalisation des phones a été annotée à l'aide de symboles proches de SAMPA (Speech Assessment Methods Phonetic Alphabet) et l'alignement des mots ont été établies par le LIMSI (Laboratoire Interdisciplinaire des Sciences du Numérique) et notamment Mme. Adda-Decker du LPP (Laboratoire de Phonétique et Phonologie) parmi les auteurs. Nous avons également récolté les couches d'annotation concernant l'identifiant et le sexe du locuteur, les mots en phonèmes, et la syllabation.

Dans notre recherche, nous avons décidé d'exclure un locuteur dont l'enregistrement ne contenait pas le nombre des occurrences quantitatif, le corpus avec lequel nous avons procédé à nos expériences comporte donc quarante-cinq enregistrements dont vingt-quatre locuteurs masculins et vingt-et-un locuteurs féminins ainsi que les grilles d'annotations qui leurs sont associées. Les différentes annotations nous serviront notamment dans les expériences où nous souhaitons non seulement mesurer la nasalité de chaque locuteur en prenant compte du contexte d'une voyelle, mais aussi vérifier si le sexe aurait un impact dans cette mesure.

4. Méthodes

Dans cette section, nous présenterons les méthodes abordées dans le but de construire divers jeux de données que nous utilisons dans le cadre de nos expériences. L'extraction de ces dernières nécessitait l'utilisation du langage Praat permettant de mettre en œuvre un fichier audio et un fichier au format textGrid en même temps pour obtenir des informations citées dans la section précédente.

Nous introduirons, en deuxième lieu, l'ensemble de nos travaux réalisés sur l'hyper-paramétrage des systèmes de réseau de neurones ainsi que la généralisation des modèles selon les données entraînées et la création d'une mesure de coarticulation nasale. Le contexte d'une voyelle et le découpage des images en petits tronçons seront les items permettant d'observer le phénomène de coarticulation.

4.1. Extraction des données

4.1.1. Sélection des échantillons

Une voyelle orale et une voyelle nasale qui ont toutes les deux le même degré d'aperture ou la même antériorité et qui se différencient de la nasalité, ont été considérées comme une paire de voyelles à examiner dans notre travail. Ainsi, la voyelle orale /a/ et la voyelle nasale /ɑ̃/ qui sont du type ouvert ont été repérées ainsi que la paire de /ɛ/ et /ɛ̃/ étant toutes les deux mi-ouvertes et antérieures et qui se distinguent par la nasalité.

Dans la phase d'extraction des voyelles, certaines restrictions peuvent être présentes afin d'obtenir le même nombre de voyelles chez tous les locuteurs. Comme la voyelle /a/ est plus fréquente que la voyelle /ɑ̃/ et est souvent plus courte, seules les voyelles /a/ ayant une durée plus longue que la durée moyenne (environ 62 ms pour la voyelle /a/ dans notre étude) étaient donc sujets d'extraction. Le calcul de la durée moyenne des voyelles /a/ de tous les locuteurs était nécessaire avant de procéder à l'extraction. Ensuite, les voyelles ayant une durée plus longue que 250 ms ont été exclues (afin d'éviter les formes atypiques, hésitations, etc.) et la sélection des occurrences de voyelles s'est effectuée sous le contrôle du contexte de la voyelle. Les différentes

consonnes présentes dans le contexte sont par la suite regroupées dans des six grandes classes qui permettent 36 combinaisons possibles : pause, /m/, /n/, subdivision des consonnes non-nasales en labiales (/p/, /b/, /f/, et /v/), coronales (/t/, /d/, /s/, /z/, /ʒ/, /ʃ/, et /l/) et dorsales (/j/, /w/, /ɥ/, /k/, /g/, et /ʁ/).

Avec ces trois restrictions présentées, nous avons souhaité d'une part calculer le nombre d'occurrences de chaque voyelle chez tous les locuteurs, d'autre part connaître le nombre minimal d'occurrences d'une voyelle. Par exemple, dans la parole du locuteur n° 1, 300 occurrences de la voyelle /a/ et 300 occurrences de la voyelle /ɑ̃/ ont été repérées tandis que le locuteur n° 2 contient dans sa parole 400 occurrences de la voyelle /a/ et 200 occurrences de la voyelle /ɑ̃/. Dans ce cas, le nombre minimal d'occurrences d'une voyelle est de 200, et l'extraction de 200 voyelles orales et nasales ne pose aucun problème chez tous les locuteurs.

En outre, nous avons inclus une autre paire de phonèmes /ɛ/ et /ɛ̃/ dans le but de s'assurer que les modèles entraînés non seulement sur les voyelles /a/ et /ɑ̃/, mais aussi sur les voyelles /a/, /ɑ̃/, /ɛ/ et /ɛ̃/ montrent leur capacité dans la généralisation de la nasalité sur un autre phonème. Nous souhaitons noter qu'il est nécessaire de calculer la durée moyenne et le nombre minimum des voyelles à extraire sur la totalité des voyelles que nous désirons analyser, donc sur quatre voyelles /a/, /ɑ̃/, /ɛ/ et /ɛ̃/ en l'occurrence.

Trois types de jeu de données ont été sujets d'extraction et peuvent être présentés comme l'illustre le tableau suivant :

	Voyelles d'entraînement	Voyelles de test
N° 1	/a/ et /ɑ̃/	/a/ et /ɑ̃/
N° 2	/a/ et /ɑ̃/	/a/, /ɑ̃/, /ɛ/ et /ɛ̃/
N° 3	/a/, /ɑ̃/, /ɛ/ et /ɛ̃/	/a/, /ɑ̃/, /ɛ/ et /ɛ̃/

Table 3: Présentation des jeux de données

4.1.2. Création des spectrogrammes pour les réseaux de neurones convolutifs

L'extraction des occurrences a été réalisée à l'aide d'une fonction du langage Praat "extract part" en lui fournissant comme paramètres le temps de début et celui de fin de la voyelle. Les extraits sonores ont été utiles pour procéder à la concaténation des voyelles dans le but d'un test de perception (section 4.2.6). Quant au spectrogramme, nous avons opté pour le créer à partir du signal global au lieu de l'extrait du signal qui est susceptible de produire des effets de bord, la fonction "To Spectrogram…" avec les paramètres par défaut a été utilisée pour créer l'objet spectrogramme.

Lors de l'extraction du spectrogramme des voyelles sous forme d'une image, nous avons dû penser à la durée des voyelles qui diffère selon les occurrences. Le zéro-padding permet d'obtenir une image de la même taille et ainsi d'éviter la compression ou l'expansion de certaines images a donc été utilisé. À partir d'une image de départ obtenue, une quantité de zéro proportionnelle à la taille de l'image a été insérée à la fin du son. Par exemple, lorsque nous avons un son à 100 ms, 150 ms de zéro ont été ajoutés pour obtenir à la fin de l'extraction une image de 250 ms. Dans le langage Praat, plusieurs fonctions ont été utilisées pour procéder au zéro-padding. En résumé, lorsque la voyelle rencontrée a un label identique à "a" ou "A" dans la grille d'annotation du corpus, que le son à gauche et celui à droite de la voyelle font partie des six grandes classes présentées, que la voyelle ait une durée supérieure à la durée moyenne des voyelles /a/ (environ 62 ms dans notre étude) et inférieure à la durée maximum (250 ms) et que le nombre de voyelles extraites jusqu'à présent ne dépasse pas le nombre minimum des voyelles, nous avons sélectionné l'objet spectrogramme du signal global puis calculé la quantité de zéros qui sera proportionnelle à la taille de l'image. Le calcul du facteur du zéro-padding est visible dans la figure 25 en annexe.

Ensuite, cette valeur a été utilisée comme deuxième paramètre de la fonction "Select outer viewport..." permettant d'inclure une marge. Cette dernière nécessite quatre paramètres pour fonctionner : le premier et le deuxième paramètre étant un intervalle horizontal à gauche et à droite respectivement, et le troisième et le dernier étant un intervalle vertical en haut et en bas

respectivement. Comme nous avons souhaité que des zéros suivent le son traité, les intervalles verticaux n'ont pas été modifiés ainsi que le premier intervalle. Seul le deuxième intervalle change en fonction de la durée du son traité. La figure 7 présente ce qui a été obtenu après l'utilisation de la fonction "Select outer viewport...".

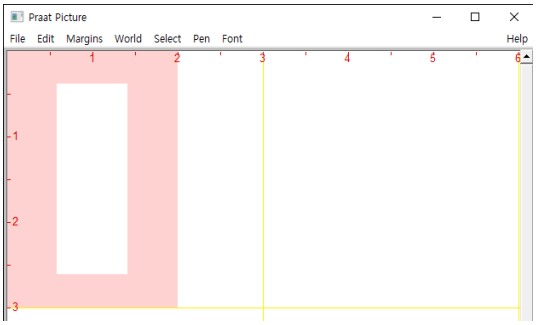


Figure 7: Fonction outer view port pour un son à 100 ms: 0, 2, 0, 3.

Comme l'illustre la figure 7 présentée ci-dessus, nous avons préparé une fenêtre dans laquelle se trouvera l'extrait du spectrogramme du son. À l'aide de la fonction "Paint…", nous avons indiqué le début et la fin de l'extraction par le temps de début et celui de fin de la voyelle en question. De plus, la présence des axes de temps et de fréquence peut être spécifiée au moyen du dernier paramètre de la fonction. Il est à noter que nous avons opté pour une image sans axes dans le cadre de notre étude dans le but que le système de réseaux de neurones ne les prenne en compte pour faire la tâche de classification.

Lorsque l'extrait de spectrogramme du son a bien été dessiné dans la fenêtre, des suites de zéros ont été insérées à la fin de l'extrait. Autrement dit, la fenêtre a été élargie vers la droite jusqu'à ce qu'elle ait atteint la même taille que celle de l'extrait du son à 250 ms, ceci est illustré dans les figures 8 et 9 ci-après.

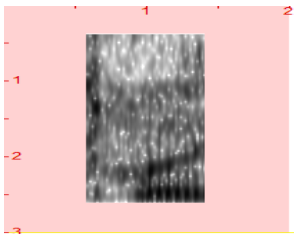


Figure 8: Fonction Paint sans axes

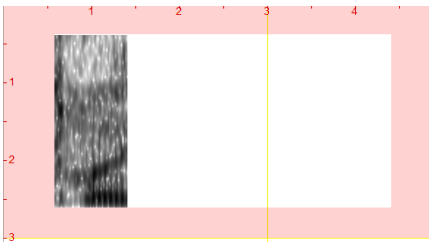


Figure 9: Zéro-padding sur le spectrogramme

Ainsi, la taille obtenue a été identique pour toutes images grâce à l'entourage des zéros autour de l'extrait. La sauvegarde des images s'est effectuée en format PNG (300 DPI).

Plusieurs informations ont été gardées dans le nom du fichier comme l'identifiant du locuteur, l'étiquette de la voyelle, le numéro d'intervalle où se trouve la voyelle dans la grille d'annotation, le contexte et le temps de début et de fin. Celles-ci nous ont été utiles dans la phase d'analyse des résultats. 80% de données extraites ont été sauvegardées dans le répertoire nommé "train" et le reste dans celui nommé "test". Chaque répertoire contient deux sous-répertoires "nasal" et "non_nasal" dans lesquels se trouvent respectivement les images de voyelles nasales et celles de voyelles orales.

4.1.3. Découpage en trois tronçons

Enfin, le découpage d'une image en trois parties égales ("tronçons") représentant le début, le milieu et la fin du segment a été réalisé dans le but d'observer l'évolution de la détection de nasalité au cours d'un phonème. Les restrictions ont été les mêmes, néanmoins, une modification a été apportée dans la phase d'extraction. Par exemple, la durée d'un tronçon a été calculée à partir de la durée du son en découpant chaque extrait en trois parties égales, et a été utilisée pour obtenir la valeur du facteur de zéro-padding. Le numéro du tronçon a été également inséré dans le nom du fichier, cette information a été utilisée dans le but d'étiqueter les images lors de la classification.

Trois jeux de données ont été établis à partir des voyelles /ɑ̃/ et /ɛ̃/ contreparties nasales des voyelles orales /a/, /ɛ/. Le calcul sur la durée moyenne (62 ms pour la voyelle /a/ dans notre étude) est le nombre minimum d'extractions était nécessaire pour restreindre le nombre des voyelles orales qui était souvent plus fréquentes et courtes que les voyelles nasales dans la production de parole. Cette restriction a été envisagée pour que tous les locuteurs aient le même nombre de voyelles extraites et que la durée des voyelles soit plus ou moins longue et similaire entre elles. Ces voyelles extraites ont été enregistrées sous forme de spectrogrammes et le découpage des spectrogrammes en trois parties égales en termes de durée a également été réalisé. La mise en œuvre du zéro-padding permet de ne pas perdre l'information sur les bords de l'image lors du passage dans la couche de convolution des réseaux de neurones convolutifs.

4.2. Réseau de neurones convolutif

4.2.1. Présentation des systèmes CNN

Nos réseaux de neurones convolutifs contiennent six couches et sont structurés ainsi :

- une couche de convolution dont la fonction d'activation est modifiée d'un modèle à l'autre dans nos expériences afin d'observer la performance du modèle
- une couche de pooling permettant de réduire l'information des images
- une couche convolutive
- une couche de pooling
- deux couches denses.

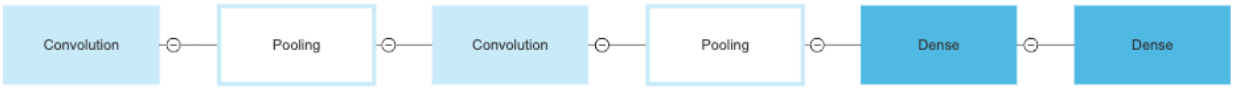


Figure 10: Représentation visuelle de l'architecture de nos réseaux de neurones convolutifs

Quatre variants de modèle ont été créés dans le cadre de notre expérience et dans chacun de ces derniers, la configuration (la fonction de perte, l'optimizer, la fonction d'activation de la première couche de convolution, etc.) a été modifiée dans le but d'observer la performance des modèles et de repérer les hyper-paramètres les mieux adaptés à nos données. Notons que tous nos modèles utilisaient le type de métrique "accuracy" pour présenter les résultats. Les hyper-paramètres et les fonctions dans un système de réseaux de neurones convolutifs avec lesquels nous avons obtenu les meilleurs résultats étaient les suivants :

- fonction d'activation de la première couche de convolution : "tanh"
- fonction de perte : "mean_squared_error"
- optimizer : "sgd"
- nombre d'époques : 115.

Dans la phase d'importation des données, pour la liste d'images et celle de étiquettes, nous avons implémenté une liste dans laquelle le nom du fichier a été ajouté. Comme il contenait

l'identifiant du locuteur, les contextes et les temps de début et de fin du son, chaque son a pu être identifié a posteriori.

Les modèles ont été sauvegardés à l'aide d'une fonction "*save*" et importés avec la fonction de keras "*tf.keras.models.load_model*". Cette méthode permettait de réduire le temps d'exécution et d'éviter d'importer les données d'entraînement à chaque expérience. L'enregistrement des modèles entiers ont été effectués avec les informations sur la configuration, la structure, et les valeurs de poids du modèle apprises lors de l'entraînement et ont pu être réutilisés a posteriori.

4.2.2. Analyse en items

Nous nous sommes ensuite intéressés à obtenir les scores selon chaque item à l'aide du modèle avec les hyper-paramètres choisis, selon le locuteur, le tronçon ou le segment. Pour ce faire, nous avons construit un jeu de données pour chaque item et procédé à la prédiction du modèle sur ces nouvelles données. A titre d'exemple, pour le locuteur n° 1, les images des voyelles de ce locuteur ont été repérées avec leur étiquette associée et stockées respectivement dans une liste d'images avec leurs étiquettes. Cette liste d'images a été utilisée par le modèle pour prédire la classe de chaque image, et les labels prédits ont été comparés avec les vrais labels des images avec la fonction de Keras "classification_report(y_test, y_pred)" permettant de calculer la précision, le rappel et la f-mesure selon la catégorie. Il en est de même pour l'item "tronçon" et l'item "voyelle". Plusieurs ensembles de données ont été créés selon l'item traité à partir du jeu de données de test, ce qui aide à obtenir des scores sur chacun des ensembles d'item. Tous les résultats ont été sauvegardés dans un fichier csv à l'aide d'une librairie de python "pandas".

	precision_0	recall_0	f1_score_0	support_0	precision_1	recall_1	f1_score_1	support_1	f1_score_accura	support_accurac	precision_macro	recall_macro_av	f1_score_macro
03_12_07_nb1_	0.92	0.85	0.89	55	0.86	0.93	0.89	55	0.89	110	0.89	0.89	0.89
03_12_07_nb1_	0.82	0.91	0.86	55	0.9	0.8	0.85	55	0.85	110	0.86	0.85	0.85
04_12_07_nb1_	0.9	1	0.95	55	1	0.89	0.94	55	0.95	110	0.95	0.95	0.95
04_12_07_nb1_	0.98	0.98	0.98	55	0.98	0.98	0.98	55	0.98	110	0.98	0.98	0.98
04_12_07_nb2_	0.98	0.89	0.93	55	0.9	0.98	0.94	55	0.94	110	0.94	0.94	0.94
04_12_07_nb2_	0.94	0.84	0.88	55	0.85	0.95	0.9	55	0.89	110	0.9	0.89	0.89
04_12_07_nb3_	0.92	0.82	0.87	55	0.84	0.93	0.88	55	0.87	110	0.88	0.87	0.87
04_12_07_nb3_	0.89	0.93	0.91	55	0.92	0.89	0.91	55	0.91	110	0.91	0.91	0.91
05_12_07_nb1_	0.96	0.84	0.89	55	0.85	0.96	0.91	55	0.9	110	0.91	0.9	0.9
05_12_07_nb1_	0.98	0.93	0.95	55	0.93	0.98	0.96	55	0.95	110	0.96	0.95	0.95
14_11_07_nb1_	0.92	1	0.96	55	1	0.91	0.95	55	0.95	110	0.96	0.95	0.95
14_11_07_nb1_	0.89	0.91	0.9	55	0.91	0.89	0.9	55	0.9	110	0.9	0.9	0.9

Figure 11: Résultat pour l'item "locuteur"

	precision_0	recall_0	f1_score_0	support_0	precision_1	recall_1	f1_score_1	support_1	f1_score_accura	support_accurac	precision_macro	recall_macro_av	f1_score_macro
A	0.00	0.00	0.00	0	1.00	0.93	0.96	1260	0.93	1260	0.50	0.47	0.48
E	1.00	0.68	0.81	1260	0.00	0.00	0.00	0	0.68	1260	0.50	0.34	0.40
I	0.00	0.00	0.00	0	1.00	0.65	0.79	1260	0.65	1260	0.50	0.32	0.39
a	1.00	0.93	0.96	1260	0.00	0.00	0.00	0	0.93	1260	0.50	0.46	0.48

Figure 12: Résultat pour l'item "voyelle" (/a/, /ɑ̃/, /ɛ/ et /ɛ̃/ notés "a", "A", "E" et "I" respectivement dans le tableau)

	precision_0	recall_0	f1_score_0	support_0	precision_1	recall_1	f1_score_1	support_1	f1_score_accura	support_accurac	precision_macro	recall_macro_av	f1_score_macro	support_macro_	precision_weigh	recall_weighted	f1_score_weig
1	0.76	0.7	0.73	2520	0.72	0.78	0.75	2520	0.74	5040	0.74	0.74	0.74	5040	0.74	0.74	0.7
2	0.79	0.72	0.75	2520	0.74	0.81	0.78	2520	0.77	5040	0.77	0.77	0.77	5040	0.77	0.77	0.7
3	0.81	0.68	0.74	2520	0.72	0.84	0.78	2520	0.76	5040	0.77	0.76	0.76	5040	0.77	0.76	0.7

Figure 13: Résultat pour l'item "tronçon" (1 pour première tiers, 2 pour deuxième tiers, 3 pour dernière tiers d'une image)

4.2.3. Analyse sur l'influence sur la classification

Nous nous sommes interrogés sur l'influence de certaines conditions sur la classification : les contextes précédents et suivants, la durée, la fréquence fondamentale, l'intensité du son et le sexe du locuteur.

4.2.3.1. Contextes et durée

La probabilité d'appartenance de chaque occurrence dans une catégorie permet d'étudier le nombre d'occurrences mal identifiées lors de la classification en tenant compte des contextes phonémiques de chaque occurrence et de détecter la nasalité pour chaque locuteur. La probabilité de chaque occurrence a été calculée selon le locuteur, et la liste des noms de fichiers créée dans le cadre de l'importation des données permettait de retrouver à quel locuteur chaque son appartenait. Ainsi, un fichier au format csv présenté comme dans la figure 14 ci-dessous a été créé dont chaque ligne contient les informations sur le son, les contextes, les temps de début et de fin, et les probabilités d'appartenance à chaque catégorie.

	son	contexte_gauche	contexte_droit	contextes	proba_oral	proba_nasal
1	a	dorsal	coronal	dorsal_coronal	0.998488	0.001512
2	a	labial	coronal	labial_coronal	0.998083	0.001917
3	a	labial	coronal	labial_coronal	0.989211	0.010789
4	a	dorsal	coronal	dorsal_coronal	1	0
5	a	labial	coronal	labial_coronal	1	0
6	a	dorsal	coronal	dorsal_coronal	1	0
7	a	n	coronal	n_coronal	0.999997	0.000003
8	a	labial	dorsal	labial_dorsal	0.999991	0.000009
9	a	coronal	coronal	coronal_coronal	0.999999	0.000001
10	a	dorsal	pause	dorsal_pause	0.899188	0.100812
11	a	coronal	coronal	coronal_coronal	0.998945	0.001055
12	a	pause	coronal	pause_coronal	0.999989	0.000011
13	a	coronal	m	coronal_m	0.966145	0.033855
14	a	n	coronal	n_coronal	1	0
15	a	dorsal	dorsal	dorsal_dorsal	0.96946	0.03054

Figure 14: Analyse de probabilité d'appartenance de chaque occurrence

A partir de ces résultats obtenus présentés dans la figure 14 ci-dessus, nous étions capables de calculer les vrais positifs, faux positifs, vrais négatifs et faux négatifs selon les locuteurs et déterminer dans quels contextes notre modèle de réseaux de neurones convolutionnels a une forte ou faible chance de se tromper lors de la prédiction.

Nous avons distingué deux types de faux négatifs dans notre expérience : les voyelles orales se trouvant dans un contexte oral qui ont été détectées comme nasales par les réseaux de neurones convolutifs, et celles ayant un contexte nasal qui ont été prédites comme nasales selon le même système. Notre hypothèse était que ces dernières ont été coarticulées, et qu'il paraissait logique qu'elles soient plus nasalisées. Nous avons analysé les voyelles orales dans un contexte oral qui ont été détectées comme nasales par le système, ceci nous permettait d'estimer dans quel contexte la voyelle orale a tendance à être détectée comme nasale. L'influence du locuteur sur la détection de nasalité pourrait à son tour être interprétée comme un indice de nasalité global du locuteur.

Nous avons également souhaité déterminer si la durée a un impact sur la bonne classification par notre modèle CNN. Une fois toutes les durées obtenues à partir de la segmentation en phones, un autre calcul a permis de classer catégoriellement un son comme long, égal ou court en fonction du locuteur (par exemple, la durée de /a/ en moyenne était d'environ 96 ms et celle de /ã/ était de

93 ms pour le locuteur ‘26_11_07_nb1_1_16’ dans notre étude). Faisant le lien avec les probabilités d’appartenance à une classe, nous avons initialisé un compteur pour compter le nombre de faux positifs et celui de faux négatifs. Autrement dit, si le son /a/ est prédit comme “nasal” selon le modèle, il a été traité comme un faux négatif tandis qu’une voyelle nasale détectée comme “orale” a été traitée comme un faux positif. En comparant deux nombres de mauvaises classifications, une tendance telle que “court”, “long”, “null” ou “égal” a été observée (la fréquence fondamentale au milieu de la voyelle /a/ en moyenne était d’environ 110 Hz et celle de /ɑ̃/ était d’environ 101 Hz ; l’intensité au milieu de la voyelle /a/ en moyenne était de 60 dB et celle de /ɑ̃/ était 52 dB pour le locuteur ‘26_11_07_nb1_1_16’ dans notre étude).

4.2.3.2. Fréquence fondamentale et intensité

Les valeurs de f0 et d'intensité ont été calculées à l'aide d'un script Praat. Dans le but de faciliter la tâche d'extraction, un fichier au format txt a été tout d'abord envisagé pour chaque locuteur. Dans ce dernier, existent trois colonnes représentant le segment, le temps de début et le temps de fin et chaque ligne correspond à un son du locuteur en question. Les colonnes ont été segmentées par une tabulation comme illustré dans la figure 15 présentée ci-dessous.

```
segment debut   fin
a       3644.3590000000004      3644.429
a       3634.2690000000002      3634.3390000000004
a       3658.2990000000004      3658.4190000000003
a       3784.138        3784.238
a       3793.868        3793.968
a       3748.926        3749.0359999999996
a       3790.006        3790.0759999999996
a       3722.8940000000002      3722.974
```

Figure 15: Fichier txt utilisé dans la phase d’extraction de f0 et intensité de chaque son

Ensuite, le son a été extrait avec une marge de 50 ms au début et à la fin du segment en préservant le temps réel et converti en objet *Pitch* et *Intensity* dans le but d’obtenir les valeurs de f0 au milieu de la voyelle, de f0 en moyenne, d’intensité au milieu de la voyelle et d’intensité en

moyenne. L'ajout des marges a été effectué dans cette phase d'extraction afin que la séquence ne soit pas trop courte à extraire.

L'analyse de la f0 et de l'intensité a été réalisée d'une part au milieu du segment à l'aide d'une fonction "*Get value at time...*" en spécifiant le temps, d'autre part sur la totalité de la durée du segment à l'aide d'une fonction "*Get mean…*". Notre programme Praat a permis d'obtenir un fichier au format csv en sortie contenant plusieurs colonnes : segment, temps de début, temps de fin, f0 au milieu, f0 en moyenne, intensité au milieu et intensité en moyenne. Trois premières nous ont été utiles pour retrouver la probabilité d'appartenance du son dans les fichiers que nous avons créés dans le cadre de l'étude de contextes. Le reste des colonnes étaient sujets de comparaison afin d'établir si, par exemple, un son ayant une f0 élevée est difficilement détectable comme correctement comparé à un son dont la f0 est basse.

son	temps_debut	temps_fin	f0_milieuV	f0_mean	intensite_milieuV	intensite_mean
a	3644.359	3644.429	172.1295414	170.6273605	62.25156974	61.42857983
a	3634.269	3634.339	147.0351021	145.9445299	64.52255175	63.2852496
a	3658.299	3658.419	114.6403436	114.8604454	66.69482322	64.20656671
a	3784.138	3784.238	145.5926358	145.2543834	58.76133565	58.06278423
a	3793.868	3793.968	--undefined--	111.9467671	48.24424829	48.04287881
a	3748.926	3749.036	148.5618815	148.1410223	66.0336625	66.21268662

Figure 16: Analyse selon f0 et intensité de chaque occurrence

4.2.4. Analyse sur la généralisation des CNNs

Afin d'observer la capacité de généralisation du modèle, le système a été à nouveau entraîné avec un jeu de données de quarante locuteurs et testé sur les données de cinq locuteurs. Pour un modèle donné, quatre tests ont été réalisés avec à chaque fois une sélection aléatoire de cinq locuteurs. Nous avons cherché à obtenir les divers scores tels que la précision, le rappel et la f-mesure selon chaque classe et procédé à une comparaison entre ces derniers et ceux obtenus par le modèle entraîné sur quarante-cinq locuteurs.

À chaque test, un jeu de données d'entraînement et celle de test ont été à nouveau initialisées en fonction des cinq locuteurs choisis, cependant, la modification n'a pas été apportée au niveau de la configuration et la structure du modèle.

4.2.5. Analyse en mesures acoustiques

Les analyses acoustiques ont été effectuées à l'aide de VoiceSauce (http://www.phonetics.ucla.edu/voicesauce/). Ce logiciel a été choisi car il effectue des mesures acoustiques d'amplitude d'harmoniques et de formants (notamment A1 et h1-A1) qui sont reconnues pour être pertinentes dans le cadre d'une analyse acoustique de la nasalité.

Les paramètres choisis pour l'analyse sont résumés par la capture d'écran du logiciel présentée dans la figure 17 ci-dessous. La segmentation n'a pas été extraite quant à elle, mais l'étiquette du phonème, le contexte phonétique immédiat, ainsi que le nom du locuteur ont été indiqués dans le nom du fichier pour les analyses à venir.

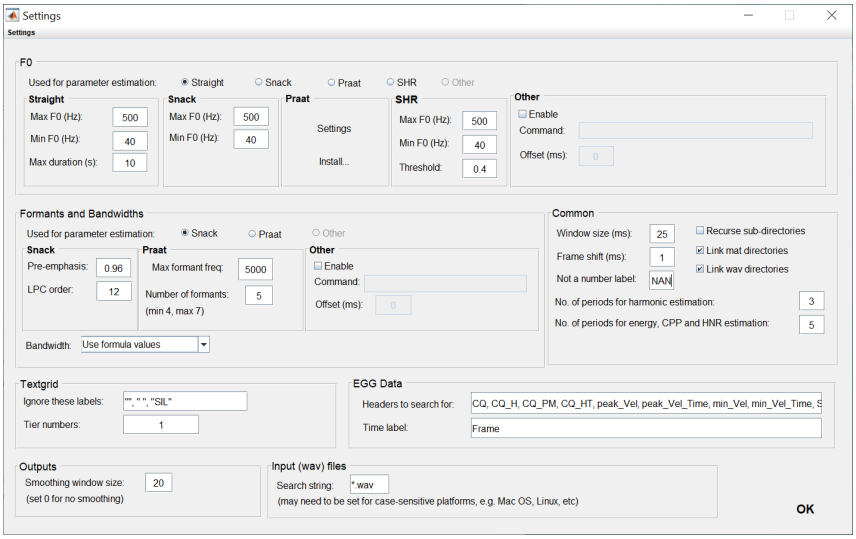


Figure 17: Paramétrage dans le logiciel VoiceSauce pour l'analyse en mesures acoustiques

Puisque le logiciel a des difficultés pour traiter les fichiers longs, les voyelles utilisées dans la phase test des réseaux de neurones convolutifs ont été extraites et analysées individuellement. Afin de ne pas être affectées par les effets de bords ainsi que par la taille de la fenêtre d'analyse, les voyelles ont été extraites avec une marge de 35 ms avant et après le début de la segmentation initiale de la voyelle. Les analyses ainsi obtenues ont été moyennées en une seule valeur après avoir retiré les valeurs obtenues sur les marges de précaution avant et après la segmentation.

4.2.6. Analyse perceptive

Après nous être appuyés sur le test de classification sur la nasalité des locuteurs, nous nous sommes intéressée à le valider de manière perceptive. Pour cela faire, sur une sélection des locuteurs, quatre groupes de voyelles /a/ et /ɑ̃/ ont été repérés :

- Voyelles nasales correctement identifiées ;
- Voyelles orales correctement identifiées ;
- Voyelles nasales mal identifiées ;
- Voyelles orales mal identifiées.

L'analyse perceptive s'est réalisée de manière semi-automatique car d'une part les sons ont été extraits et concaténés à l'aide d'un script Praat, d'autre part nous avions recours aux mains-d'œuvre de la part des humains pour la sauvegarde du fichier audio et le test de perception.

Extraction et concaténation des fichiers audio

Un fichier de texte contenant les sons et les temps de début et de fin associés a été préparé à partir des fichiers csv de chaque locuteur, ces temps permettaient de trouver non seulement le son, mais aussi les étiquettes du son et de ces contextes dans le fichier original. Une fois que nous avons réussi à repérer le moment du fichier auquel appartenait chaque son du groupe de voyelles, tous les sons du groupe ont été extraits avec une marge de 50 ms au début et à la fin de la voyelle et renommés avec les informations sur le nom du fichier original, les contextes et les temps. Les sons extraits ont été utilisés d'une part pour créer un test de perception, d'autre part pour mener une analyse qualitative dans le but d'étudier plus finement les erreurs de classifications à l'aide des mesures acoustiques.

Quant au test de perception, nous avons sélectionné quatre voyelles pour chaque groupe de voyelles présenté ci-dessus, ce qui fait seize voyelles au total. Les voyelles choisies devaient avoir une durée longue, être de même durée (soit entre 150 ms et 250 ms) et représentatives (la voyelle /a/ qui est réalisée comme telle). Le test de perception a été créé à l'aide d'un logiciel

gratuit intitulé “Google Form” permettant de créer un questionnaire en ligne. Dans le cadre de l’expérience perceptive sous forme d’un questionnaire, deux questions générales nous aidant à comprendre la situation des participants ont été posées, par exemple, si sa langue d’origine est le français et si le participant avait une expérience de suivre un ou des cours de linguistique. Parce que chaque voyelle a été répétée deux fois, trente-six voyelles ont été entendues une après l’autre par les participants. Pour chaque voyelle, deux questions ont été posées pour demander quelle voyelle entend le participant entre une voyelle orale et celle nasale, et aussi entre la voyelle /a/, la voyelle /ɑ̃/ et autres.

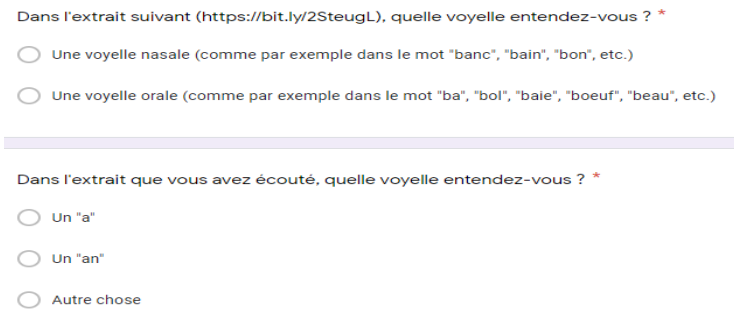


Figure 18: Exemple de question du test de perception

Dans le cadre d’une analyse qualitative, une concaténation des extraits a été réalisée. Lors de la concaténation des extraits d’enregistrement, une grille d’annotation a été créée et les noms des extraits y ont été insérés à l’aide de la fonction “*Concatenate recoverably*”. Elle permet également d’ajouter des intervalles à la fin de chaque extrait et d’y insérer le nom de l’extrait de façon automatique comme l'illustre la figure 19.

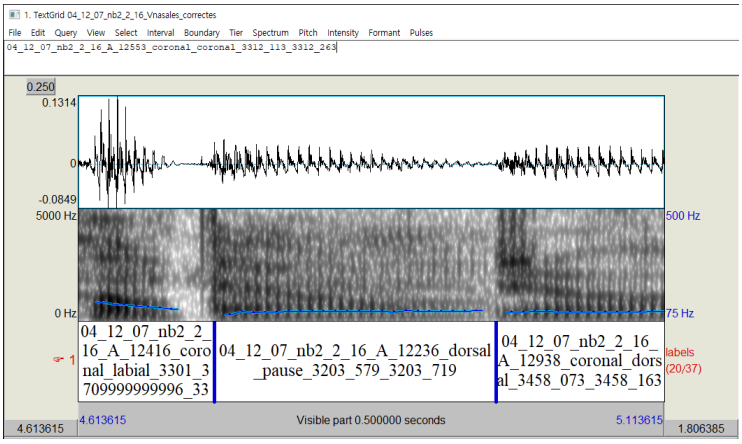


Figure 19: TextGrid créé à l'aide d'une fonction "concatenate recoverably"

Au moyen des réseaux de neurones convolutifs, diverses analyses ont été menées :

- en item comme le locuteur, la voyelle et le tronçon, les scores ont été obtenus pour chaque ensemble d'item pour savoir, par exemple, quels sont les locuteurs identifiés comme nasals
- sur l'influence des conditions telles que contextes, sexe, durée, f0 ou intensité sur la classification du système dans le but d'établir dans quels contextes le système tend à produire les erreurs.

Les mesures acoustiques H1c et H1A1c ont été étudiées dans le cadre de la vérification de la nasalité, et le test de perception et l'analyse acoustique pour comparer la perception humaine avec les performances du réseau ont également été entreprises.

5. Résultats et discussion

5.1. Analyse en items

L'analyse en différentes catégories aide à trouver des indices de nasalité dans des ensembles d'items. Avec l'item *Segment*, nous souhaitons en section 5.1.1. établir si le modèle de réseaux de neurones convolutifs entraîné sur une paire de voyelles /a/ et /ã/ est généralisable sur un ensemble de sons /a/, /ã/, /ɛ/ et /ɛ̃/ et comparer les performances du système selon de différentes jeux de données présentés en section 4.1.2.5. En outre, avec l'item *tronçon*, puisqu'un segment a été découpé en trois tronçons représentant le début, le milieu et la fin, nous regarderons en section 5.1.2. l'évolution de la nasalité au cours du phonème et étudierons quels tronçons ont les meilleurs scores selon les divers jeux de données. Enfin, l'item *locuteur* nous permettrait de savoir pour quels locuteurs notre système CNN a une forte chance de se tromper dans le cadre de la prédiction, ou au contraire, quels sont les locuteurs qui ont les meilleurs résultats obtenus à partir du système. La section 5.1.3. consiste à repérer ces deux types de locuteurs avec deux statistiques dits faux positifs et faux négatifs en nous concentrant sur une paire de voyelles /a/ et /ã/.

5.1.1. Segment

L'analyse sur les voyelles a été effectuée à l'aide des trois modèles CNN et les données d'entraînement et celles de test sont des images entières non découpées :

- Modèle entraîné avec une paire de voyelles /a/ et /ã/, et testé sur une paire de voyelles /a/ et /ã/ ;
- Modèle entraîné avec une paire de voyelles /a/ et /ã/, et testé sur deux paires de voyelles /a/, /ã/, /ɛ/ et /ɛ̃/ ;
- Modèle entraîné avec deux paires de voyelles /a/, /ã/, /ɛ/ et /ɛ̃/ et testé sur deux paires de voyelles /a/, /ã/, /ɛ/ et /ɛ̃/.

	F-mesure non nasal	F-mesure nasal	Accuracy
/a/	97%		94%
/ɑ̃/		96%	93%

Table 4: Modèle entraîné avec /a/-/ɑ̃/ et testé sur /a/-/ɑ̃/

	F-mesure non nasal	F-mesure nasal	Accuracy
/a/	96%		93%
/ɑ̃/		96%	93%
/ɛ/	81%		68%
/ɛ̃/		79%	65%

Table 5: Modèle entraîné avec /a/-/ɑ̃/ et testé sur /a/-/ɑ̃/ et /ɛ/-/ɛ̃/

Le premier modèle entraîné avec les voyelles /a/ et /ɑ̃/ et testé sur les mêmes voyelles obtient une exactitude de 94% pour identifier la voyelle /a/ et 93% pour la voyelle /ɑ̃/. La f-mesure pour la catégorie "*non nasal*" est de 97% voyelle /a/ et la f-mesure pour la catégorie "*nasal*" est de 96% pour la voyelle /ɑ̃/. Quant au deuxième modèle entraîné avec la même paire de voyelles que le premier mais testé sur deux paires de voyelles /a/-/ɑ̃/ et /ɛ/-/ɛ̃/, l'exactitude et la f-mesure rapportent une baisse de 1% pour la voyelle /a/ et restent identiques pour la voyelle /ɑ̃/. Les voyelles insérées dans le cadre du test ont des scores moins élevés comparés aux deux dernières. Le modèle entraîné sur la paire /a/-/ɑ̃/ obtient une exactitude de 68% pour détecter la voyelle /ɛ/ et 65% pour la voyelle /ɛ̃/. Une f-mesure pour la classe "*non nasal*" est de 81% pour la voyelle /ɛ/ et une f-mesure pour la classe "*nasal*" est de 79% pour la voyelle /ɛ̃/.

	F-mesure non nasal	F-mesure nasal	Accuracy
/a/	92%		85%
/ɑ̃/		98%	94%
/ɛ/	98%		97%
/ɛ̃/		89%	80%

Table 6: Modèle entraîné avec /a/-/ɑ̃/ et /ɛ/-/ɛ̃/, et testé sur /a/-/ɑ̃/ et /ɛ/-/ɛ̃/

Le dernier modèle a été entraîné avec deux paires de voyelles /a/-/ɑ̃/ et /ɛ/-/ɛ̃/ et testé sur les mêmes paires. Pour ce dernier, l'exactitude obtenue pour identifier la voyelle /a/ est de 85% et de 94% pour la voyelle /ɑ̃/. La f-mesure pour la classe "*non nasal*" est en baisse de 4% pour la voyelle /a/ comparé au deuxième modèle tandis que celle pour la classe "*nasal*" rapporte une augmentation de 2% pour la voyelle /ɑ̃/. L'exactitude et la f-mesure pour la classe "*non nasal*" obtenues pour la voyelle /ɛ/ sont de 97% et de 98% respectivement tandis que la voyelle /ɛ̃/ obtient une exactitude de 80% et une f-mesure de 89% pour la catégorie "*nasal*". Une hausse de deux mesures sont observés pour les voyelles /ɛ/ et /ɛ̃/ par rapport au deuxième modèle.

Les modèles entraînés et testés sur les mêmes voyelles montrent leur capacité de classification (le taux de bonnes classifications varie de 93% à 94% pour le modèle 1, de 80% à 97% pour le modèle 3). Le modèle entraîné sur une paire de voyelles /a/-/ɑ̃/ se trouve généralisable sur deux paires de voyelles /a/-/ɑ̃/ et /ɛ/-/ɛ̃/ bien que le taux de bonnes classifications soit plus bas que les deux autres modèles (le taux d'exactitude varie de 65% à 93% selon les voyelles).

5.1.2. Tronçon

L'analyse sur les tronçons s'est effectuée sur trois jeux de données dans lesquels se trouvent les images découpées en trois tronçons représentant le début, le milieu et la fin d'un son. La différence entre les modèles présentés dans la section précédente et ceux utilisés dans cette section porte seulement sur le type de données. Les modèles pour l'item *Segment* sont entraînés

et testés sur les images entières et non découpées tandis que ceux que nous avons utilisés pour l'item *Tronçon* sont entraînés et testés sur les images découpées.

Le modèle noté "*M1"* dans le tableau 4 ci-après a été entraîné et testé sur la paire de voyelles /a/-/ɑ̃/, le modèle "*M2*" entraîné avec /a/-/ɑ̃/ et testé sur /a/-/ɑ̃/ et /ɛ/-/ɛ̃/, et le dernier entraîné et testé sur /a/-/ɑ̃/ et /ɛ/-/ɛ̃/.

	Premier tiers			Deuxème tiers			Dernier tiers		
	F-mesure nasal	F-mesure oral	accuracy	F-mesure nasal	F-mesure oral	accuracy	F-mesure nasal	F-mesure oral	accuracy
M1	83%	83%	83%	89%	88%	89%	86%	85%	86%
M2	73%	74%	73%	76%	76%	76%	77%	75%	76%
M3	80%	79%	79%	84%	83%	83%	84%	82%	83%

Table 7: Scores de classification obtenus par les trois modèles selon le tronçon

Pour le premier tiers du son, l'exactitude obtenue fluctue entre 73% et 83% en fonction des modèles. Le modèle 1 obtient une f-mesure de 83% pour deux catégories données tandis que le modèle 2 montre une f-mesure de 73% pour la catégorie "*nasal*" et de 74% pour "*non nasal*". Concernant le modèle 3, une f-mesure de 80% est observée pour la catégorie "*nasal*" et celle de 79% pour la catégorie "*non nasal*". Parmi les trois modèles, le modèle 1 entraîné et testé sur les voyelles /a/ et /ɑ̃/ obtient les meilleurs scores dans les trois mesures et le modèle 2 entraîné avec la paire de voyelles /a/-/ɑ̃/ et testé sur deux paires de voyelles /a/-/ɑ̃/ et /ɛ/-/ɛ̃/ rapporte des scores les plus bas.

En ce qui concerne le deuxième tiers correspondant au milieu du son, l'exactitude obtenue à partir de trois modèles se situe entre 76% et 89%. Il en est de même pour les f-mesures pour la catégorie "*nasal*" et "*non nasal*" qui varient de 76% à 89% selon les modèles. Le modèle 1 montre une meilleure performance parmi les modèles utilisés en obtenant environ 89% pour toutes les mesures tandis qu'elles sont égales à 76% pour le modèle 2.

Pour le dernier tiers correspond à la fin d'un son, l'exactitude varie de 76% à 86% selon le modèle utilisé. Le modèle 1 observe une f-mesure de 86% pour la catégorie "*nasal*" et de 85% pour "*non nasal*". Pour le modèle 2, une f-mesure pour la classe "*nasal*" est de 77% et celle pour

la classe "*non nasal*" est de 75% tandis que le modèle obtient 84% et 82% pour la catégorie "*nasal*" et "*non nasal*" respectivement.

Parmi les trois parties du son, le deuxième tiers correspondant au milieu de la voyelle présente dans l'ensemble le taux de bonnes classifications le plus élevé et stable variant de 76% à 89% selon les modèles. Le dernier tiers obtient les scores analogues que ceux du deuxième tiers. Le premier tiers représente le taux d'exactitude le plus bas dans tous les modèles abordés situé entre 73% et 83%.

5.1.3. Locuteur

Il s'agit d'étudier le nombre d'occurrences de voyelles mal identifiées en fonction du locuteur. L'analyse sur l'item *locuteur* s'est basée principalement sur les voyelles /a/ et /ɑ̃/, et pour chaque locuteur, 110 occurrences de voyelles dont 55 voyelles /a/ et 55 voyelles /ɑ̃/ ont été étudiés. Les voyelles orales identifiées comme nasales ont été traitées comme les faux négatifs, les voyelles nasales bien classées comme vrais négatifs et les voyelles nasales détectées comme orales comme les faux positifs. Rappelons qu'il existe deux types de faux négatifs dans notre expérience : une voyelle orale dans un contexte oral détectée comme nasale et une voyelle orale dans un contexte nasal détectée comme nasale. Le fait de distinguer deux types de faux négatifs permet d'analyser plus finement les mauvaises détections.

	FN dans un contexte oral	FN dans un contexte nasal	VN	FP
26_11_07_nb1_1_16	2	0	53	2
26_11_07_nb2_1_16	4	0	51	4
26_11_07_nb1_2_16	2	2	52	3
26_11_07_nb2_2_16	1	0	55	0
26_11_07_nb3_1_16	2	0	52	3
26_11_07_nb3_2_16	1	0	51	4

27_11_07_nb1_1_16	2	0	54	1
27_11_07_nb1_2_16	1	0	55	0
27_11_07_nb2_1_16	0	1	51	4
27_11_07_nb2_2_16	4	0	54	1
28_11_07_nb1_1_16	0	0	47	8
28_11_07_nb1_2_16	0	1	50	5
28_11_07_nb2_1_16	7	0	53	2
28_11_07_nb2_2_16	0	0	48	7
29_11_07_nb1_2_16	1	1	39	16
29_11_07_nb2_1_16	4	0	51	4
29_11_07_nb2_2_16	6	1	53	2
30_11_07_nb1_1_16	8	1	49	6
30_11_07_nb1_2_16	3	1	54	1
05_12_07_nb1_1_16	8	1	53	2
05_12_07_nb1_2_16	4	0	54	1
14_11_07_nb1_1_16	0	0	41	14
14_11_07_nb1_2_16	6	0	50	5
14_11_07_nb2_1_16	6	1	53	2
14_11_07_nb2_2_16	4	0	48	7
16_11_07_nb1_1_16	3	0	50	5
16_11_07_nb1_2_16	5	1	54	1
16_11_07_nb2_1_16	2	0	55	0
16_11_07_nb2_2_16	1	0	48	7
20_11_07_nb1_1_16	1	0	49	6
20_11_07_nb1_2_16	3	2	52	3
22_11_07_nb1_1_16	2	0	49	6
22_11_07_nb1_2_16	0	1	50	5
22_11_07_nb2_1_16	1	0	53	2
22_11_07_nb2_2_16	3	0	51	4
23_11_07_nb1_1_16	4	1	51	4
23_11_07_nb1_2_16	5	0	52	3
03_12_07_nb1_1_16	7	0	53	2
03_12_07_nb1_2_16	3	0	47	8
04_12_07_nb1_1_16	1	0	49	6
04_12_07_nb1_2_16	2	1	54	1
04_12_07_nb2_1_16	6	0	53	2
04_12_07_nb2_2_16	6	2	52	3
04_12_07_nb3_1_16	4	2	51	4

04_12_07_nb3_2_16	4	0	49	6

Table 8: Faux positifs (FP), vrais négatifs (VN) et faux négatifs (FN) pour les occurrences de voyelles /a/-/ɑ̃/ selon locuteur

Les locuteurs ayant le plus de vrais négatifs et faux positifs ont les identifiants suivants :

- 28_11_07_nb2_1_16 ;
- 29_11_07_nb2_2_16 ;
- 05_12_07_nb1_1_16 ;
- 14_11_07_nb2_1_16 ;
- 03_12_07_nb1_1_16 ;
- 04_12_07_nb2_2_16.

Soixante-deux occurrences de voyelles détectées comme nasales sont repérées pour le locuteur 05_12_07_nb1_1_16. Sur l'ensemble des occurrences, cinquante-trois voyelles nasales bien classées sont présentes tandis que le nombre de voyelles orales identifiées comme nasales est de neuf. Quatre locuteurs (28_11_07_nb2_1_16, 29_11_07_nb2_2_16, 14_11_07_nb2_1_16 et 03_12_07_nb1_1_16) obtiennent le même score : les vrais positifs sont de cinquante-trois et les faux négatifs sont de sept, ce qui fait un total de soixante. Le locuteur 04_12_07_nb2_2_16 obtient cinquante-deux voyelles nasales identifiées comme nasales et huit voyelles orales détectées comme nasales.

Dans ce paragraphe, nous avons retiré les voyelles orales détectées comme nasales dans un contexte nasal dans le but de ne pas étudier les voyelles co-articulées d'une nasale, mais plutôt les voyelles qui sont détectées comme nasales pour d'autres raisons. Le locuteur 05_12_07_nb1_1_16 obtient le plus d'occurrences de voyelles orales identifiées comme nasales et de voyelles nasales détectées correctement. Sur soixante-et-un occurrences, sont présentes cinquante-trois voyelles nasales identifiées comme nasales et huit voyelles orales détectées comme nasales dans un contexte oral. Les locuteurs 28_11_07_nb2_1_16 et

03_12_07_nb1_1_16 obtiennent cinquante-trois occurrences de voyelles nasales bien classées et sept occurrences de voyelles orales identifiées comme nasales dans un contexte oral.

Trois locuteurs contenant le moins de vrais négatifs et de faux négatifs dans un contexte oral ont été identifiés : 28_11_07_nb1_1_16, 29_11_07_nb1_2_16 et 14_11_07_nb1_1_16. Aucune voyelle oral détectée comme nasale n'est observée pour deux locuteurs 28_11_07_nb1_1_16 et 14_11_07_nb1_1_16 tandis qu'ils obtiennent respectivement quarante-sept occurrences et quarante-et-une occurrences de voyelles nasales bien classées. Un nombre de vrais négatifs égal à trente-neuf et une occurrences de voyelle orale détectée incorrectement sont observés pour le locuteur 29_11_07_nb1_2_16. Ces locuteurs correspondent aux locuteurs ayant le plus de voyelles orales et nasales détectées comme orales.

Les trois locuteurs (05_12_07_nb1_1_16, 28_11_07_nb2_1_16 et 03_12_07_nb1_1_16) pour lesquels les voyelles orales et nasales confondues identifiées comme nasales peuvent être considérés comme ayant une voix la plus identifiée comme nasale que d'autres locuteurs. Au contraire, les locuteurs (28_11_07_nb1_1_16, 29_11_07_nb1_2_16 et 14_11_07_nb1_1_16) peuvent être les moins identifiés comme nasals parmi l'ensemble de locuteurs dans le corpus car ils obtiennent le moins de voyelles identifiées comme nasales ainsi que le nombre voyelles détectées comme orales le plus élevé que pour d'autres locuteurs.

5.2. Influence sur la classification

Ce chapitre s'appuie sur les valeurs pouvant influencer la classification. Les contextes phonémiques, le sexe du locuteur, la durée, la fréquence fondamentale et l'intensité seront étudiés avec pour objectif de déterminer dans quelles conditions le système de réseaux de neurones convolutionnels identifie mieux les voyelles.

5.2.1. Contexte

En ce qui concerne les contextes phonémiques, rappelons que nous nous sommes concentrés sur les voyelles /a/ et /ɑ̃/ et que les contextes gauche et droit ont été regroupés dans les six grandes catégories : la pause, les consonnes nasales /m/ et /n/ en considérant que d'autres consonnes nasales sont moins fréquentes, la subdivision des consonnes non-nasales en labiales, coronales et dorsales. Les éléments de gauche et de droite sont séparés par un tiret bas dans les tableaux présentés ci-dessous. Les tableaux concernant les contextes sont des récapitulatifs de l'ensemble des résultats obtenus, ils contiennent principalement le classement de cinq contextes dans lesquels les faux négatifs sont les plus représentés.

	dorsal_pause	labial_dorsal	coronal_pause	dorsal_coronal	dorsal_dorsal
/a/ détecté comme nasal	37	17	13	12	10
/a/ détecté comme oral	140	159	165	168	66
Total	177	176	178	180	76
Proportion	20.9%	9.7%	7.3%	6.7%	13.2%

Table 9: Nombre de faux négatifs selon le contexte

Le contexte *dorsal_pause* comporte le plus de voyelles orales détectées comme nasales. Sur 177 occurrences totaux de voyelles /a/ de tous les locuteurs dans ce contexte, 37 voyelles mal identifiées ont été repérées, ce qui fait une proportion de 20.9 % sur la totalité des occurrences

dans le même contexte. 17 faux négatifs ont été repérés sur l'ensemble des voyelles nasales dans le contexte *labial_dorsal*, et 13 voyelles orales mal identifiées sur 178 occurrences dans le contexte *coronal_pause*. Un nombre de faux négatifs égal à 12 sur 180 occurrences dans le contexte *dorsal_coronal* correspond à une proportion de 6.7% et celui égal à 10 sur 76 occurrences avec comme éléments de gauche et de droit *dorsal* conduit à une proportion de 13.2%.

	coronal_coronal	dorsal_coronal	coronal_dorsal	m_coronal	coronal_labial
/ɑ̃/ détecté comme oral	28	26	20	15	12
/ɑ̃/ détecté comme nasal	397	334	166	199	151
Total	425	360	186	214	163
Proportion	6.6%	7.2%	10.8%	7%	7.4%

Table 10: Nombre de faux positifs selon le contexte

Quant aux voyelles /ɑ̃/, le contexte *coronal_coronal* obtient 28 voyelles nasales détectées comme orales sur 425 occurrences. Le nombre de faux positifs est le plus élevé dans un contexte *coronal_coronal* bien que la proportion des faux positifs dans ce dernier soit le plus bas. 26 faux positifs sur 360 sont observés dans un contexte *dorsal_coronal*, ce qui fait une proportion de 7.2%. 20 voyelles nasales identifiées comme orales ont été repérées sur 186 occurrences dans le contexte *coronal_dorsal*, ce nombre représente une proportion de 10.8%. Autrement dit, le système se trompe une ou deux fois tous sur dix occurrences de voyelles nasales dans le contexte *coronal_dorsal*. Les contextes *m_coronal* et *coronal_labial* obtiennent une proportion d'environ 7% avec respectivement 15 faux positifs et 12 faux positifs sur leur nombre total d'occurrences.

La voyelle /a/ est fréquemment mal identifiée lorsqu'elle se situe entre une consonne dorsale et une pause, ce qui fait une proportion de 20,9% sur l'ensemble des voyelles dans le même contexte. Le contexte "dorsal_dorsal" vient à la suite pour la voyelle /a/ avec une proportion de

13,2% sur la totalité des voyelles dans le même contexte. La voyelle /ɑ̃/, quant à elle, est incorrectement identifiée une fois sur dix occurrences lorsqu'elle se situe entre une consonne coronale et une consonne dorsale.

5.2.2. Sexe

Dans cette section, nous avons analysé d'une part selon le contexte, d'autre part selon le sexe du locuteur dans le but d'étudier si le sexe a un impact sur la classification. Rappelons que vingt-quatre locuteurs masculins et vingt-et-un locuteurs féminins ont participé à enregistrer leur voix.

		dorsal_pause	labial_dorsal	coronal_pause	dorsal_coronal	dorsal_dorsal	total
Femmes	/a/ identifiée comme nasal	11	5	6	3	2	27
	/a/ identifiée comme oral	52	75	72	84	36	319
	Total	63	80	78	87	38	346
	Proportion	17.5%	6.3%	7.7%	3.4%	5.3%	7.8%
Hommes	/a/ identifiée comme nasal	26	12	7	9	8	62
	/a/ identifiée comme oral	88	84	93	84	30	379
	Total	114	96	100	93	38	441
	Proportion	22.8%	12.5%	7%	9.7%	21.1%	14.1%

Table 11: Nombre de faux négatifs selon le contexte et le sexe pour les voyelles orales

Aussi bien pour les hommes que pour les femmes, une grande proportion de voyelles mal identifiées dans un contexte donné est attribuée au contexte *dorsal_pause*. Pour les femmes, 11 voyelles /a/ sur 63 occurrences du contexte sont trouvées, ce qui conduit à une proportion de 17.5% sur la totalité des occurrences du contexte. Pour les hommes, 26 voyelles orales sur 114 occurrences de voyelles /a/ du contexte sont obtenues, ce qui correspond à une proportion de 22.8%.

Une différence entre les femmes et les hommes est observée au niveau des proportions. Les hommes obtiennent plusieurs contextes (*dorsal_pause, labial_dorsal et dorsal_dorsal*) représentant une proportion élevée (22.8%, 12.5% et 21.1% respectivement) tandis que pour les femmes un seul contexte *dorsal_pause* représente une proportion supérieure à 10%.

Le nombre d'occurrences de voyelles /a/ identifiées comme nasales est plus élevé pour les hommes que pour les femmes. Sur 441 occurrences de voyelles /a/ dans tous les contextes présentés dans le tableau ci-dessus, 62 voyelles orales ont été identifiées comme nasales chez les hommes, ce qui correspond à une proportion de 14.1% sur la totalité des voyelles /a/ dans tous les cinq contextes. Pour les femmes, le nombre de faux négatifs égal à 27 sur 346 conduit à une proportion de 7.8% sur toutes les voyelles /a/.

		coronal_coronal	dorsal_coronal	coronal_dorsal	m_coronal	coronal_labial	total
Femmes	/ɑ̃/ identifiée comme oral	15	17	9	5	6	52
	/ɑ̃/ identifiée comme nasal	173	170	67	71	63	544
	Total	188	187	76	76	69	596
	Proportion	8%	9.1%	11.8%	6.6%	8.7%	8.7%
Hommes	/ɑ̃/ identifiée comme oral	13	9	11	10	6	49
	/ɑ̃/ identifiée comme nasal	224	164	99	128	88	703
	Total	237	173	110	138	94	752
	Proportion	5.5%	5.2%	10%	7.2%	6.4%	6.5%

Table 12: Nombre de faux positifs selon le contexte et le sexe pour les voyelles nasales

Pour les femmes, le contexte représentant le nombre de faux positifs le plus élevé est *dorsal_coronal* avec 17 voyelles nasales détectées incorrectement sur 187 occurrences du contexte. Ce nombre conduit à une proportion de 11.8% sur la totalité des voyelles /ɑ̃/ dans le contexte *dorsal_coronal*. Pour les hommes, le contexte *coronal_coronal* comporte le plus de voyelles nasales mal identifiées, 13 faux positifs sont repérés sur 237 occurrences de voyelles /ɑ̃/ du contexte, ce qui correspond à 10% des voyelles /ɑ̃/ dans ce contexte.

Aussi bien pour les hommes que pour les femmes, le contexte *coronal_dorsal* représente le meilleur score pour la proportion des voyelles nasales mal identifiées sur la totalité des voyelles nasales dans ce contexte. Pour les femmes, le nombre de faux positifs est égal à 9 sur l'ensemble des voyelles nasales dans le contexte *coronal_dorsal* et correspond à une proportion de 11.8%. Pour les hommes, parmi 110 occurrences de voyelles nasales, 11 voyelles sont mal identifiées et conduisent à une proportion de 10%.

	VP	FP	VN	FN	FN contexte nasal	nb total	proportion
Hommes	1212	97	1223	96	12	2640	7.76%
Femmes	1104	85	1070	43	8	2310	5.88%

Table 13: Statistiques selon le sexe du locuteur

Le tableau 13 présenté ci-dessus comporte quatre statistiques dits vrais positifs, faux positifs, vrais négatifs et faux négatifs ainsi que le nombre total d'occurrences de voyelles orales et nasales et la proportion selon le sexe du locuteur. Pour les hommes, le nombre de faux positifs et de faux négatifs conduit à une proportion de 7.76% de voyelles mal identifiées sur l'ensemble des voyelles. Pour les femmes, la proportion des voyelles détectées incorrectement est de 5.88%.

Le nombre de faux négatifs chez les femmes est égal à 51 (43 faux négatifs dans un contexte oral et 8 dans un contexte nasal) et correspond à la moitié du nombre de faux négatifs chez les hommes, soit 108 (96 faux négatifs dans un contexte oral et 12 dans un contexte nasal).

Le sexe a une influence sur la classification : une proportion des voyelles mal identifiées plus élevée est observée chez les hommes (7,76%) que chez les femmes (5,88%). Les voyelles orales ont tendance à être incorrectement détectées par le système lorsqu'elles sont produites par les hommes (14,1% sur toutes les voyelles orales contre 7,8% chez les femmes). Au contraire, les réseaux tendent plus à faire des erreurs face aux voyelles nasales réalisées par les femmes que par les hommes (8,7% contre 6,5% sur toutes les voyelles nasales).

5.2.3. Durée

L'analyse sur la durée a été effectuée sur trois jeux de données différents dans le but de déterminer si la durée joue un rôle important en ce qui concerne la classification du système de réseaux de neurones convolutifs. Nous verrons si le système a tendance à mal classifier un son si ce dernier est court en termes de durée.

Le tableau 14 présenté ci-après résume les résultats obtenus à partir de trois modèles et chacun de ces modèles utilise un jeu de données différent. Le jeu de données construit à partir d'une paire de voyelles /a/-/ɑ̃/ a été utilisé pour entraîner et tester le modèle noté "M1" et le modèle noté "M2" a été entraîné avec le jeu de donnée d'une paire de voyelles /a/-/ɑ̃/ et testé à l'aide d'un autre jeu de données créé avec des images de voyelles /a/-/ɑ̃/ et /ɛ/-/ɛ̃/. Le jeu de données de deux paires de voyelles /a/-/ɑ̃/ et /ɛ/-/ɛ̃/ a été mise en œuvre pour entraîner et tester le modèle noté "M3". Tous les jeux de données utilisés dans les trois modèles présentés dans le tableau 14 ont été constitués à partir des images non découpées tandis que ceux du tableau 15 ont été créés à partir des images découpées en trois tronçons.

Quatre termes dits "court", "égal", "long", et "null" se trouvent dans les tableaux. Trois premiers se réfèrent à la tendance des voyelles classées incorrectement. Par exemple, si une voyelle /a/ mal identifiée a une durée inférieure à la moyenne des durées de voyelles /a/ d'un locuteur, cette dernière est comptée comme courte. Ensuite, si le nombre de voyelles courtes mal classifiées est supérieur au nombre de voyelles longues détectées incorrectement, la tendance est considérée comme courte, c'est-à-dire, il y a une forte chance que le système se trompe lorsque les sons d'un locuteur donné sont courts. Au contraire, si le nombre de voyelles courtes identifiées incorrectement est inférieur au nombre de voyelles longues mal classées, la tendance est "longue". Si ces deux nombres sont égaux, la tendance de mauvaises classifications est "égale". Si un locuteur donné n'obtient aucune voyelle détectée incorrectement, la tendance est nulle.

		/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/
		Court				Egal				Long				Null			
Images entières	M1	32	15			5	8			5	19			3	3		
	M2	24	8	45	4	3	7	0	0	6	23	0	41	12	7	0	0
	M3	20	8	**9**	4	5	2	1	8	14	**14**	4	29	6	21	31	4

Table 14: Nombre de locuteurs selon la tendance d'erreurs calculée avec la durée

Le modèle 1 effectue des mauvaises prédictions pour trente deux locuteurs lorsque la voyelle /a/ est courte et pour cinq locuteurs lorsque la voyelle a une durée égale ou longue à la durée en moyenne. Et trois locuteurs ont été repérés pour lesquels le modèle 1 n'obtient aucune mauvaise classification. Quant à la voyelle /ɑ̃/, dix-neuf locuteurs ont une tendance à avoir les voyelles longues mal identifiées tandis que quinze locuteurs observent une tendance à obtenir des voyelles détectées incorrectement lorsqu'elles sont courtes. Le modèle 1 a des difficultés à classer correctement une voyelle /ɑ̃/ dont la durée est égale à la durée en moyenne chez huit locuteurs, et ne se trompe pas pour trois locuteurs dans le cadre de la prédiction.

En ce qui concerne le modèle 2, la tendance "courte" est observée pour les voyelles /a/ chez vingt-quatre locuteurs, et la tendance "longue" chez vingt-trois locuteurs. Le modèle a une forte chance de se tromper lors de la prédiction de la voyelle /ɛ/ courte car tous les locuteurs obtiennent de mauvaises classifications lorsque la voyelle /ɛ/ est courte comparé à la moyenne des durées de voyelle /ɛ/. Une tendance forte est étudiée chez quarante-et-un locuteurs lorsque la voyelle /ɛ̃/ a une durée longue.

La voyelle /a/ se situe dans une tendance plutôt "courte" pour le dernier modèle, le nombre de locuteurs concernés est de 20. Le modèle a une tendance à prédire correctement les voyelles /ɑ̃/ et /ɛ/ tandis qu'il se trompe souvent pour la voyelle /ɛ̃/ lorsque cette dernière est longue.

Nous observons le nombre d'erreurs le plus bas pour toutes les quatre voyelles et tous les modèles, ce qui peut néanmoins être dû au nombre des voyelles ayant une durée parfaitement égale à la moyenne. Le nombre de voyelles ayant une durée parfaitement égale à la moyenne doit être toujours inférieur à celui des voyelles courtes et longues. Si nous ne tenons pas compte de tendance "égale", des erreurs se produisent fréquemment pour les voyelles orales courtes et les voyelles nasales longues.

		/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/
		court				egal				long				null			
Images en trois tronçons	M1	8	7			1	2			36	36			0	0		
	M2	14	6	43	2	0	3	1	1	31	36	1	42	0	0	0	0
	M3	19	8	16	15	3	1	4	1	23	34	20	29	0	2	5	0

Table 15: Nombre de locuteurs selon la tendance d'erreurs calculée avec la durée (images découpées)

Les voyelles /a/ et /ɑ̃/ longues se trouvent difficiles à identifier selon le modèle 1, chez trente-six locuteurs le modèle obtient souvent de fausses classifications lorsque ces deux voyelles sont longues. Cette tendance est également observée pour le reste des modèles, sauf que le modèle 2 a une tendance à identifier incorrectement la voyelle /ɛ/ courte chez quarante-trois locuteurs.

Pour les images découpées, nous ne trouvons pas de particularité comparées aux images non découpées qui apportent une tendance observée selon la nasalité. Cependant, la tendance "longue" pour les voyelles nasales peut être vérifiée pour les voyelles nasales à l'aide des images découpées.

5.2.4. Fréquence fondamentale

Dans cette section, nous examinons si la fréquence fondamentale a un impact dans la classification. Les jeux de données utilisés sont les mêmes que pour l'analyse en durée de la section 5.2.3., seuls les termes de tendances semblent légèrement différents : "basse", "égale", "élevée", et "nulle". Le concept est identique sauf que nous nous référons à la hauteur de la fréquence fondamentale dans cette expérience. Rappelons que la fréquence fondamentale a été calculée non seulement au milieu de la voyelle, mais aussi en moyenne.

5.2.4.1. F0 au milieu de la voyelle

		/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/
		Bas				Égal				Élevé				Null			
Images entières	M1	**18**	11			6	7			**18**	**24**			**3**	**3**		
	M2	8	13	**17**	8	6	6	8	1	**19**	**19**	**20**	**36**	12	**7**	0	0
	M3	11	11	3	12	5	3	2	7	**23**	10	9	**22**	6	**21**	**31**	4

Table 16: Nombre de locuteurs selon la tendance d'erreurs calculée avec la f0 au milieu de la voyelle

Le tableau 16 ci-dessus présente les résultats obtenus à partir de la fréquence fondamentale au milieu de la voyelle et des images non découpées. La voyelle /a/ se situe entre la tendance "basse" et "élevée" car dix-huit locuteurs obtiennent de fausses classifications du modèle 1 lorsque la fréquence fondamentale de la voyelle /a/ est basse ou élevée. La voyelle /ɑ̃/ se trouve plus dans une tendance à être mal identifiée par le modèle lorsqu'elle a une fréquence fondamentale supérieure à la moyenne.

Une forte chance que le modèle 2 commet des erreurs est observée pour toutes les voyelles lorsque la f0 de ces dernières est considérée comme élevée. La voyelle /ɛ/ se trouve légèrement entre deux tendances : "basse" et "élevée" chez dix-sept locuteurs et vingt locuteurs respectivement.

Le modèle 3 ne produit pas d'erreurs pour les voyelles /ɑ̃/ et /ɛ/ chez vingt-et-un locuteurs et trente-et-un locuteurs respectivement. Deux voyelles /a/ et /ɛ̃/ tendent à être incorrectement identifiées chez vingt-trois locuteurs et vingt-deux locuteurs respectivement lorsqu'elles ont une fréquence fondamentale élevée qui est calculée au milieu d'elles.

Pour les images découpées en tronçons présentés dans le tableau 17 ci-dessous, toutes les voyelles dont la fréquence fondamentale calculée au milieu de la voyelle est élevée semblent difficiles à détecter correctement pour tous les modèles. Cette tendance forte est observée chez une vingtaine ou trentaine de locuteurs.

		/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/
		Bas				Égal				Élevé				Null			
Images en trois tronçons	M1	13	6			3	1			29	38			0	0		
	M2	11	8	12	5	5	4	1	2	29	33	32	38	0	0	0	0
	M3	13	15	5	10	7	3	7	5	25	25	28	30	0	2	5	0

Table 17: Nombre de locuteurs selon la tendance d'erreurs calculée avec la f0 au milieu de la voyelle (images découpées)

Pour les images découpées en trois tronçons, les erreurs tendent à être produites par le réseau sur toutes les voyelles orales et nasales confondues lorsque la fréquence fondamentale au milieu de la voyelle est élevée. La tendance similaire est observée pour les images. Les voyelles utilisées pour entraîner et tester le modèle 1 et 2 obtiennent des erreurs avec une fréquence fondamentale au milieu de la voyelle élevée plus fréquemment qu'avec celle basse.

5.2.4.2. F0 en moyenne

		/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/
		Bas				Égal				Élevé				Null			
Images entières	M1	**20**	15			10	5			12	**22**			3	3		
	M2	**13**	**14**	**27**	15	7	**11**	7	3	**13**	**13**	14	**27**	**12**	7	0	0
	M3	**16**	11	3	**16**	4	5	2	9	**19**	8	9	**16**	6	**21**	**31**	4

Table 18: Nombre de locuteurs selon la tendance d'erreurs calculée avec la f0 moyenne

Le modèle 1 entraîné et testé sur une paire de voyelle /a/ et /ɑ̃/ a des difficultés à classer correctement la voyelle /a/ pour vingt locuteurs lorsque cette dernière a une fréquence fondamentale basse. Contrairement à la voyelle /a/, la voyelle /ɑ̃/ a une tendance à être mal identifiée par le modèle quand elle a une f0 supérieure à la moyenne.

La voyelle /a/ se situe entre trois tendances (bas, élevé et null) selon le modèle 2 entraîné avec une paire de voyelles /a/-/ɑ̃/ et testé sur deux paires de voyelles /a/-/ɑ̃/ et /ɛ/-/ɛ̃/ car une dizaine de locuteurs se voit concerner par chacune de ces tendances. La voyelle /ɑ̃/ est attribuée de trois tendances (basse, égale, élevée) avec également une dizaine de locuteurs qui en sont affectées. Le modèle 2 obtient une tendance "basse" pour vingt-sept locuteurs dans le cadre de la prédiction de la voyelle /ɛ/ et une tendance "élevée" pour vingt-sept locuteurs lors de la prédiction de la voyelle /ɛ̃/.

Pour le dernier modèle, une tendance "basse-élevée" est observée lors de la classification des voyelles /a/ et /ɛ̃/ chez une quinzaine ou vingtaine de locuteurs. Les voyelles /ɑ̃/ et /ɛ/ semblent faciles à identifier selon le modèle 3, aucune erreur dans la classification n'est observée chez vingt-et-un locuteurs et trente-et-un locuteurs respectivement.

Les modèles entraînés avec des images découpées en trois tronçons présentés dans le tableau 19 ci-après montrent une tendance majoritaire : élevée. Dans la plupart des voyelles de tous les modèles, une vingtaine ou trentaine de locuteurs est observée dans cette tendance, les voyelles ayant une fréquence fondamentale en moyenne sont considérées comme difficiles à détecter

d'après tous les modèles existant dans le tableau ci-dessous. Cependant, la voyelle /a/ se situe entre deux tendances (basse et élevée) dans tous les modèles.

		/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/
		Bas				Égal				Élevé				Null			
Images en trois tronçons	M1	21	9			4	1			**20**	35			0	0		
	M2	**18**	11	16	8	5	7	5	1	22	27	24	36	0	0	0	0
	M3	21	16	6	11	6	4	8	6	**18**	23	26	28	0	2	5	0

Table 19: Nombre de locuteurs selon la tendance d'erreurs calculée avec la f0 moyenne (images découpées)

Pour les images non découpées, la fréquence fondamentale moyenne ne dégage pas une tendance forte pour chacune des voyelles étudiées. La tendance d'erreurs varie entre "bas" et "élevé" selon les modèles. Au contraire, pour les images découpées, les erreurs sont observées dans les voyelles lorsqu'elles ont une fréquence fondamentale moyenne excepté la voyelle /a/ ayant une tendance située entre "bas" et "élevé".

5.2.5. Intensité

L'analyse en intensité a mis en œuvre les mêmes modèles et le même concept que les analyses en durée et en fréquence fondamentale. Le but de l'analyse en intensité est d'établir si un son à intensité forte est considéré comme difficile à identifier d'après le système de réseaux de neurones convolutifs. L'intensité a été calculée, comme la fréquence fondamentale, au milieu de la voyelle et en moyenne. Notons que les mesures d'intensité ici sont dépendantes de la position du casque par rapport aux lèvres du locuteur, ainsi qu'à la calibration du locuteur et doivent être considérées avec la précaution qui s'impose.

5.2.5.1. Intensité au milieu de la voyelle

		/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/
		Bas				Égal				Élevé				Null			
Images entières	M1	**30**	9			5	6			7	**27**			3	3		
	M2	**24**	11	12	10	2	9	12	2	7	**18**	**21**	**33**	12	7	0	0
	M3	**26**	10	6	7	2	2	1	7	11	**12**	**7**	**27**	6	**21**	**31**	4

Table 20: Nombre de locuteurs selon la tendance d'erreurs calculée avec l'intensité au milieu de la voyelle

Le premier modèle obtient une tendance forte mais distincte pour deux voyelles /a/ et /ɑ̃/ : "basse" chez trente locuteurs et "élevée" chez vingt-sept locuteurs respectivement. Elle peut être observée dans les résultats du modèle 2, la voyelle /a/ a une tendance à être mal détectée lorsqu'elle a une intensité forte qui a été calculée au milieu de la voyelle tandis que le reste des voyelles (/ɑ̃/, /ɛ/ et /ɛ̃/) se situent dans une tendance dite "élevée". Cette tendance pour trois voyelles peut être observée à l'aide du nombre de locuteurs qui y sont impactés (18, 21 et 33 locuteurs respectivement).

Le dernier modèle obtient la même tendance dite "basse" pour la voyelle /a/ chez vingt-six locuteurs mais seule la voyelle /ɛ̃/ se trouve dans la tendance "élevée" avec vingt-sept locuteurs. Le modèle 3 n'obtient aucune faute lors de la prédiction de deux voyelles /ɑ̃/ et /ɛ/ chez vingt-et-un locuteurs et trente-et-un locuteur respectivement.

Quant aux images découpées en trois tronçons présentés dans le tableau 18 ci-dessous, la voyelle /a/ semble être difficile à classer correctement dans tous les modèles car chez la majorité de locuteurs, le modèle se trompe souvent sur les voyelles /a/ lorsqu'elles ont une intensité basse. Contrairement à la voyelle /a/, tout le reste des voyelles (/ɑ̃/, /ɛ/ et /ɛ̃/) s'observent dans une tendance dite "élevée" dans tous les modèles abordés dans cette expérience.

		/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/
		Bas				Égal				Élevé				Null			
Images en trois tronçons	M1	31	11			2	1			12	33			0	0		
	M2	30	13	12	6	6	3	8	2	9	29	25	37	0	0	0	0
	M3	23	15	9	7	4	8	9	6	18	20	22	32	0	2	5	0

Table 21: Nombre de locuteurs selon la tendance d'erreurs calculée avec l'intensité au milieu de la voyelle (images découpées)

Aussi bien pour les images non découpées que pour les images découpées, l'intensité au milieu de la voyelle montre qu'il est fréquent que la voyelle /a/ soit incorrectement identifiée lorsqu'elle a une intensité basse et, à l'inverse les autres voyelles /ɑ̃/, /ɛ/, /ɛ̃/ sont incorrectement identifiées lorsque l'intensité élevée.

5.2.5.2. Intensité en moyenne

		/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/
		Bas				Égal				Élevé				Null			
Images entières	M1	**30**	9			5	6			7	**27**			3	3		
	M2	**24**	12	**17**	8	3	9	11	5	6	**17**	**17**	**32**	12	7	0	0
	M3	**27**	9	6	7	2	4	1	7	10	**11**	**7**	**27**	6	**21**	**31**	4

Table 22: Nombre de locuteurs selon la tendance d'erreurs calculée avec l'intensité moyenne

Dans cette expérience, les voyelles /a/ dont l'intensité en moyenne est basse sont sujets difficiles à classer selon nos modèles. Le modèle 1 fait souvent des fausses prédictions de la voyelle /a/ chez trente locuteurs tandis qu'il en obtient moins pour la voyelle /ɑ̃/. Cette dernière se trouve

plutôt facile à identifier par le modèle 1 lorsqu'elle a une intensité en moyenne faible, mais au contraire, lorsque l'intensité en moyenne est élevée, le modèle tend à se tromper sur la voyelle /ɑ̃/.

Le modèle 2 obtient des résultats similaires à ceux obtenus à partir du modèle 1. Il identifie difficilement les voyelles /a/ lorsqu'elle a une intensité moyenne basse comme nous le voyons chez vingt-quatre locuteurs. Il en est de même pour la voyelle /ɛ/, le modèle a du mal à identifier face aux voyelles /ɛ/ chez 17 locuteurs lorsque ces voyelles ont une intensité basse. Les voyelles /ɑ̃/ et /ɛ̃/ font partie des sons que le modèle fait le plus souvent des fausses classifications lorsque leur intensité moyenne est élevée.

Le dernier modèle noté "*M3*" dans le tableau 19 ci-dessus explique la difficulté d'identification sur les voyelles /a/ dont l'intensité moyenne est basse et sur les voyelle /ɛ̃/ dont l'intensité est élevée. Deux voyelles /ɑ̃/ et /ɛ/ ne sont pas considérées comme difficiles à identifier, aucune erreur de classification n'est observée chez vingt-et-un locuteurs et trente-et-un locuteurs respectivement.

Les résultats obtenus à partir des modèles entraînés et testés avec les images découpées en trois tronçons se sont présentés dans le tableau 23 ci-dessous. Tous nos modèles considèrent les voyelles /a/ comme difficiles à détecter correctement lorsque l'intensité de ces sons est basse, ils obtiennent fréquemment les fausses détections chez une vingtaine ou une trentaine de locuteurs. Tous les autres sons tels que /ɑ̃/, /ɛ/ et /ɛ̃/ tendent à être mal identifiés par les modèles lorsqu'ils sont une intensité moyenne élevée excepté la voyelle /ɑ̃/ se situe entre deux tendances basses et élevée selon le modèle 3.

		/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/	/a/	/ɑ̃/	/ɛ/	/ɛ̃/
		Bas				Égal				Élevé				Null			
Images en trois tronçons	M1	33	11			2	1			10	33			0	0		
	M2	31	15	18	7	5	3	5	1	9	27	22	37	0	0	0	0
	M3	24	**18**	11	6	6	6	6	7	15	19	23	32	0	2	5	0

Table 23: Nombre de locuteurs selon la tendance d'erreurs calculée avec l'intensité moyenne (images découpées)

Pour les images non découpées comme pour les images découpées, lorsqu'une erreur est observée dans les voyelles /a/, il est fréquent que celles-ci aient une intensité moyenne basse. Au contraire, les autres voyelles tendent à être incorrectement identifiées lorsque leur intensité moyenne est élevée.

5.3. Généralisation du modèle

Dans cette nouvelle section, nous essayons de généraliser le modèle entraîné avec les données de quarante locuteurs sur les données de cinq locuteurs non vus par le système, et comparons les résultats obtenus avec ceux du modèle original. Rappelons que quatre tests ont été effectués pour un modèle donné et cinq locuteurs ont été choisis de manière aléatoire pour chaque test. Comme pour les analyses sur l'influence de la durée (section 5.2.3.) des fréquences fondamentales (section 5.2.4.) et de l'intensité sur la classification (section 5.2.5.), trois jeux de données ont été utilisés dans le cadre de cette expérience.

5.3.1. Modèles entraînés et testés sur deux voyelles

Le jeu de données constitué d'une paire de voyelles /a/-/ã/ a été mis en œuvre pour entraîner et tester un modèle. Deux types de jeu de données peuvent être distingués : jeu de données créé à partir des images non découpées, et des images découpées en trois tronçons. Le modèle dit original dans les tableaux qui suivent a été entraîné et testé sur l'ensemble des locuteurs dans le corpus, et les modèles notés "test" avec leur numérotation ont été entraînés avec les données de quarante locuteurs et testés sur celles de cinq locuteurs. Comme l'illustre le tableau 24

ci-dessous, la sélection de cinq locuteurs non vus par le système lors de l'entraînement diffère d'un modèle de test à l'autre.

	Cinq locuteurs de test
Test 1	29_11_07_nb1_2_16, 29_11_07_nb2_1_16, 29_11_07_nb2_2_16, 30_11_07_nb1_1_16, 30_11_07_nb1_2_16
Test 2	27_11_07_nb2_2_16, 28_11_07_nb1_1_16, 28_11_07_nb1_2_16, 28_11_07_nb2_1_16, 28_11_07_nb2_2_16
Test 3	26_11_07_nb1_2_16, 26_11_07_nb2_1_16, 26_11_07_nb2_2_16, 26_11_07_nb3_1_16, 22_11_07_nb2_1_16
Test 4	04_12_07_nb3_2_16, 03_12_07_nb1_1_16, 03_12_07_nb1_2_16, 04_12_07_nb1_1_16, 04_12_07_nb1_2_16

Table 24: Cinq locuteurs selon les modèles de test entraînés et testés sur deux voyelles

	F-mesure pour non-nasal	F-mesure pour nasal	accuracy
Modèle original	90%	93%	93%
Test 1	91%	90%	90%
Test 2	94%	94%	94%
Test 3	94%	95%	94%
Test 4	92%	92%	92%
Moyenne des 4 tests	92.75%	92.75%	92.5%

Table 25: Scores obtenus à partir du modèle normal et des modèles de test entraînés et testés sur deux voyelle

D'après le tableau 25 présenté ci-dessous, le modèle original entraîné et testé avec des images non découpées obtient une f-mesure de 90% pour la catégorie "*non_nasal*" et de 93% pour la catégorie "*nasal*", et une exactitude égale à 93%. Pour les tests, les f-mesures varient de 91% à 94% pour la classe "*non_nasal*" et de 90% à 95% pour la classe "*nasal*" et l'exactitude des tests se situe entre 90% et 94%.

Les résultats obtenus à partir des images découpées en trois tronçons sont résumés dans le tableau ci-dessous. Une f-mesure de 86% pour la catégorie "*non_nasal*" et pour celle "*nasal*" est

observée pour le modèle original avec une exactitude de 86%. La f-mesure varie en fonction du modèle de test de 82% à 84% pour la classe "*non_nasal*" et de 83% à 86% pour la classe "*nasal*". L'exactitude fluctue entre 82% et 85% selon le modèle de test.

	Cinq locuteurs de test
Test 1	30_11_07_nb1_1_16, 30_11_07_nb1_2_16, 29_11_07_nb2_1_16, 29_11_07_nb2_2_16, 29_11_07_nb1_2_16
Test 2	28_11_07_nb2_1_16, 28_11_07_nb2_2_16, 28_11_07_nb1_2_16, 27_11_07_nb2_2_16, 28_11_07_nb1_1_16
Test 3	26_11_07_nb3_1_16, 26_11_07_nb2_1_16, 26_11_07_nb1_1_16, 26_11_07_nb1_2_16, 23_11_07_nb1_2_16
Test 4	04_12_07_nb2_1_16, 04_12_07_nb1_1_16, 04_12_07_nb1_2_16, 03_12_07_nb1_2_16, 03_12_07_nb1_1_16

Table 26: Cinq locuteurs selon les modèles de test entraînés et testés sur deux voyelles (images découpées)

	F-mesure pour non-nasal	F-mesure pour nasal	accuracy
Modèle original	86%	86%	86%
Test 1	82%	83%	82%
Test 2	84%	85%	85%
Test 3	83%	86%	85%
Test 4	84%	84%	84%
Moyenne des 4 tests	83.25%	84.5%	84%

Table 27 : Scores obtenus à partir du modèle normal et des modèles de test entraînés et testés sur deux voyelles (images découpées)

Pour les images non découpées des voyelles /a/ et /ɑ̃/, les f-mesures pour chaque catégorie et les taux d'exactitude obtenus pour les modèles de test sont stables et similaires avec les résultats obtenus pour le modèle original. Il en est de même pour les images découpées, les modèles entraînés avec quarante locuteurs sont généralisables sur les données de cinq locuteurs non vues lors de l'entraînement.

5.3.2. Modèles entraînés avec deux voyelles et testés sur quatres voyelles

Le jeu de données utilisé pour entraîner le modèle dans cette section a été constitué de deux voyelles /a/-/ã/. Une différence entre les modèles construits en section précédente 5.3.1. et ceux de cette section a été apportée sur le jeu de données de test. Les modèles présentés dans cette section ont été testés sur quatre voyelles /a/, /ã/, /ɛ/ et /ɛ̃/.

	Cinq locuteurs de test
Test 1	29_11_07_nb1_2_16, 29_11_07_nb2_1_16, 29_11_07_nb2_2_16, 30_11_07_nb1_1_16, 30_11_07_nb1_2_16
Test 2	27_11_07_nb2_2_16, 28_11_07_nb1_1_16, 28_11_07_nb1_2_16, 28_11_07_nb2_1_16, 28_11_07_nb2_2_16
Test 3	26_11_07_nb1_2_16, 26_11_07_nb2_1_16, 26_11_07_nb2_2_16, 26_11_07_nb3_1_16, 22_11_07_nb2_1_16
Test 4	04_12_07_nb3_2_16, 03_12_07_nb1_1_16, 03_12_07_nb1_2_16, 04_12_07_nb1_1_16, 04_12_07_nb1_2_16

Table 28: Cinq locuteurs selon les modèles de test entraînés avec deux voyelles et testés sur quatre voyelles

	F-mesure pour non-nasal	F-mesure pour nasal	accuracy
Modèle original	80%	79%	80%
Test 1	85%	82%	84%
Test 2	85%	82%	84%
Test 3	89%	88%	89%
Test 4	89%	88%	88%
Moyenne des 4 tests	87%	85%	86.25%

Table 29: Scores obtenus à partir du modèle normal et des modèles de test entraînés avec deux voyelles et testés sur quatre voyelles

Une f-mesure de 80% pour la catégorie "*non_nasal*" et de 79% pour celle "*nasal*" est obtenue par le modèle original entraîné et testé sur quarante-cinq locuteurs et son taux d'exactitude est

égale à 80%. Les valeurs de f-mesures s'échelonnent entre 85% et 89% pour la classe de non nasalité et entre 82% et 88% pour celle de nasalité selon les modèles de test entraînés avec les données de quarante locuteurs et testés sur cinq locuteurs. Le taux d'exactitude varie entre 84% et 89% en fonction de ces modèles.

Dans le tableau 29, est affiché l'ensemble des résultats pour les modèles entraînés et testés avec des images découpées. Le modèle original obtient une exactitude de 75% ainsi qu'une f-mesure égale à 75% pour la classe dite "*non_nasal*" et à 76% pour celle dite "*nasal*". Selon le modèle de test, trois mesures ont un écart de 4-6% entre les valeurs supérieures et inférieures : f-mesure variant de 80% à 84% pour la classe "*non_nasal*" et de 77% à 83% pour "*nasal*", accuracy situé entre 79% et 83%.

	Cinq locuteurs de test
Test 1	30_11_07_nb1_1_16, 30_11_07_nb1_2_16, 29_11_07_nb2_1_16, 29_11_07_nb2_2_16, 29_11_07_nb1_2_16
Test 2	28_11_07_nb2_1_16, 28_11_07_nb2_2_16, 28_11_07_nb1_2_16, 27_11_07_nb2_2_16, 28_11_07_nb1_1_16
Test 3	26_11_07_nb3_1_16, 26_11_07_nb2_1_16, 26_11_07_nb1_1_16, 26_11_07_nb1_2_16, 23_11_07_nb1_2_16
Test 4	04_12_07_nb2_1_16, 04_12_07_nb1_1_16, 04_12_07_nb1_2_16, 03_12_07_nb1_2_16, 03_12_07_nb1_1_16

Table 30: Cinq locuteurs selon les modèles de test entraînés avec deux voyelles et testés sur quatre voyelles (images découpées)

	F-mesure pour non-nasal	F-mesure pour nasal	accuracy
Modèle original	75%	76%	75%
Test 1	81%	77%	79%
Test 2	80%	78%	79%
Test 3	83%	83%	83%
Test 4	84%	82%	83%
Moyenne des 4 tests	82%	80%	81%

Table 31: Scores obtenus à partir du modèle normal et des modèles de test entraînés avec deux voyelles découpées et testés sur quatre voyelles (images découpées)

Aussi bien pour les images non découpées que pour les images découpées, les modèles entraînés avec les voyelles /a/ et /ɑ̃/ de quarante locuteurs réussissent à généraliser les voyelles /a/, /ɑ̃/, /ɛ/ et /ɛ̃/ de cinq locuteurs non vus, à quel point que les mesures telles que f-mesure et accuracy sont meilleures que celles du modèle original.

5.3.3. Modèles entraînés et testés sur quatre voyelles

Quatre voyelles /a/, /ɑ̃/, /ɛ/ et /ɛ̃/ ont été choisies dans le but d'entraîner et de tester cinq modèles.

	Cinq locuteurs de test
Test 1	29_11_07_nb2_1_16, 29_11_07_nb2_2_16, 30_11_07_nb1_1_16, 30_11_07_nb1_2_16, 28_11_07_nb2_1_16
Test 2	28_11_07_nb2_2_16, 29_11_07_nb1_2_16, 27_11_07_nb2_2_16, 28_11_07_nb1_1_16, 28_11_07_nb1_2_16
Test 3	26_11_07_nb3_1_16, 26_11_07_nb3_2_16, 23_11_07_nb1_2_16, 26_11_07_nb1_1_16, 26_11_07_nb1_2_16
Test 4	04_12_07_nb2_1_16, 04_12_07_nb2_2_16, 03_12_07_nb1_1_16, 03_12_07_nb1_2_16, 04_12_07_nb1_1_16

Table 32: Cinq locuteurs selon les modèles de test entraînés et testés sur quatre voyelles

	F-mesure pour non-nasal	F-mesure pour nasal	accuracy
Modèle original	89%	89%	89%
Test 1	88%	88%	88%
Test 2	87%	87%	87%
Test 3	88%	87%	88%
Test 4	86%	87%	86%
Moyenne des 4 tests	87.25%	87.25%	87.25%

Table 33: Scores obtenus à partir du modèle normal et des modèles de test entraînés et testés sur quatre voyelles

Dans le tableau ci-dessus résumant des résultats obtenus à partir des images non découpées, trois mesures dites "f-mesure pour la catégorie de non nasalité", "f-mesure pour la catégorie de

nasalité" et "accuracy" sont toutes de 89% pour le modèle original. Les modèles de test entraînés sur un jeu de données constitué de quarante locuteurs et testés sur celui de cinq locuteurs aléatoires obtiennent une f-mesure variant de 86% à 88% pour la classe "*non_nasal*" et de 87% à 88% pour la classe "*nasal*". Un taux d'exactitude situé entre 86% et 88% est observé pour trois variantes de modèles de test.

Pour les images découpées en tronçons, selon le modèle 1, deux f-mesures égales à 81% et à 82% sont observées pour deux catégories dites "*non_nasal*" et "*nasal*" respectivement et le taux d'exactitude est de 82%. En fonction des modèles de test, est observée une variation de valeurs dans les trois mesures. La f-mesure fluctue entre 80% et 83% pour la classe de non-nasal et entre 78% et 82% pour celle de nasal tandis que le taux d'exactitude se situe entre 79% et 83%.

	Cinq locuteurs de test
Test 1	29_11_07_nb1_2_16, 29_11_07_nb2_1_16, 29_11_07_nb2_2_16, 30_11_07_nb1_1_16, 30_11_07_nb1_2_16
Test 2	27_11_07_nb2_2_16, 28_11_07_nb1_1_16, 28_11_07_nb1_2_16, 28_11_07_nb2_1_16, 28_11_07_nb2_2_16
Test 3	26_11_07_nb1_2_16, 26_11_07_nb2_1_16, 26_11_07_nb2_2_16, 26_11_07_nb3_1_16, 22_11_07_nb2_1_16
Test 4	04_12_07_nb3_2_16, 03_12_07_nb1_1_16, 03_12_07_nb1_2_16, 04_12_07_nb1_1_16, 04_12_07_nb1_2_16

Table 34: Cinq locuteurs selon les modèles de test entraînés et testés sur quatre voyelles (images découpées)

	F-mesure pour non-nasal	F-mesure pour nasal	accuracy
Modèle original	81%	82%	82%
Test 1	80%	81%	80%
Test 2	80%	78%	79%
Test 3	83%	82%	83%
Test 4	81%	80%	81%
Moyenne des 4 tests	81%	80.25%	80.75%

Table 35: Scores obtenus à partir du modèle normal et des modèles de test entraînés et testés sur quatre voyelles (images découpées)

Pour les images découpées comme pour les images non découpées des voyelles /a/, /ɑ̃/, /ɛ/ et /ɛ̃/, les résultats obtenus avec les modèles de test sont stables en termes de f-mesure selon les catégories et le taux d'exactitude. La généralisation par les modèles de test est donc observée sur les données non vues par le système.

5.4. Mesures acoustiques

Certaines mesures acoustiques aident à discriminer la nasalité d'un son comme vu dans la partie littérature Le but de l'analyse acoustique dans le cadre de notre travail est de vérifier la nasalité, les mesures acoustiques telles que H1c et H1A1c ont été mises en œuvre afin d'observer la différence entre deux voyelles /a/ et /ɑ̃/. L'étude s'est basée sur un jeu de données construit à partir des items /a/ et /ɑ̃/ et cinquante-cinq occurrences y sont présents pour chaque voyelle.

5.4.1. H1c

La réalisation d'une nasale provoque une amélioration des harmoniques, il s'agit d'un pôle nasal qui se situe dans une basse fréquence, soit entre 250 et 450 Hz en fonction du locuteur, ce dernier peut être le premier ou le second harmonique. Dans cette expérience, l'amplitude du premier harmonique est l'objet d'étude permettant de distinguer la voyelle orale /a/ de la voyelle nasale /ɑ̃/.

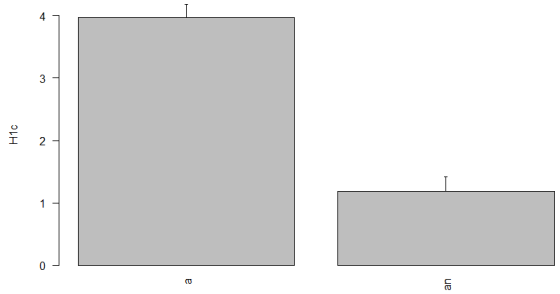


Figure 20: H1c selon la voyelle

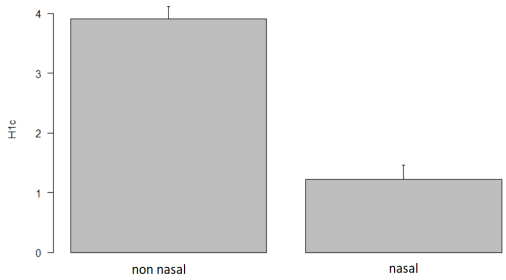


Figure 21: H1c selon la classe

À l'aide de deux figures 20 et 21, une tendance peut être observée entre les catégories à prédire et le paramètre H1c. Lorsque la classe d'une voyelle est détectée comme orale c'est-à-dire, un son est considéré comme oral, la valeur du paramètre H1c de ce son est similaire à une valeur d'une voyelle /a/. À l'inverse, lorsqu'un son est identifié comme nasal, la valeur de H1c de ce son est basse comme la valeur de la voyelle /ɑ̃/ attendue.

Nous pouvons construire deux hypothèses :

- Si un son oral est identifié comme nasal, la valeur de H1c descend ;
- Si un son nasal est identifié comme oral, la valeur de H1c augmente.

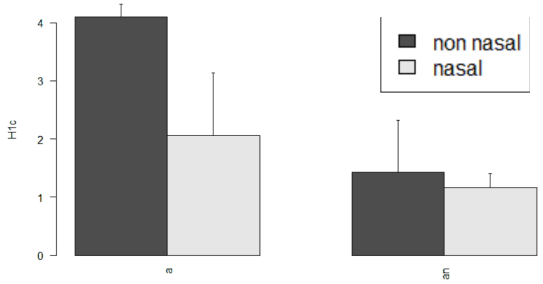


Figure 22: H1c selon la voyelle et la classe prédite

Une voyelle /a/ identifiée comme orale selon le système de CNNs obtient une valeur de H1c proche de la valeur de la voyelle /a/ comme attendu. Cependant, une voyelle /a/ détectée comme nasale montre une diminution de la valeur qui se rapproche plus de celle de la voyelle /ɑ̃/.

Une valeur de H1c proche de la valeur de voyelle /ã/ est observée pour une voyelle nasale détectée comme nasale, elle se situe entre 1 et 2. Une voyelle /ã/ ayant été classée incorrectement obtient une valeur un peu plus élevée que celle de voyelle /ã/, une augmentation est donc observée pour cette valeur.

5.4.2. H1A1c

La production de la nasalité provoque aussi une réduction de l'amplitude du premier formant d'une voyelle. À l'aide de l'amplitude de l'harmonique le plus élevé du premier formant, une mesure relative H1A1c a été envisagée. Soustraire l'amplitude du premier harmonique de celle du premier formant est nécessaire pour obtenir la différence.

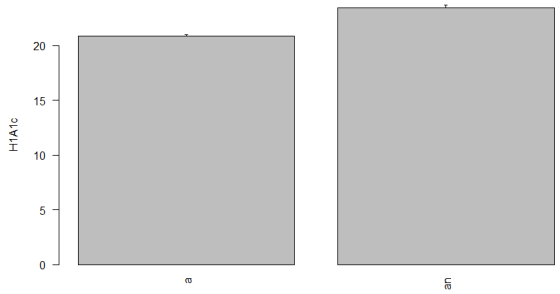


Figure 23: H1A1c selon la voyelle

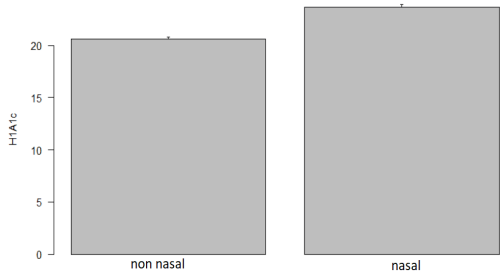


Figure 24: H1A1c selon la classe prédite

Les catégories "*nasal*" et "*non_nasal*" (ou oral) se trouvent dans la même tendance que la valeur de H1A1c de deux voyelles /a/ et /ã/. La valeur d'un son identifié comme oral semble proche de la valeur de voyelle /a/ tandis que la valeur d'un son détecté comme nasal se rapproche de la valeur de voyelle nasale /ã/.

Deux hypothèses peuvent être proposées à partir de ces résultats. La première est qu'un son oral détecté comme nasal aura une valeur de H1A1c plus élevée que ce qui est attendu et proche de la valeur de la voyelle /ã/. Et pour un son nasal considéré comme oral selon le système de CNNs, la valeur de H1A1c de ce son sera diminuée.

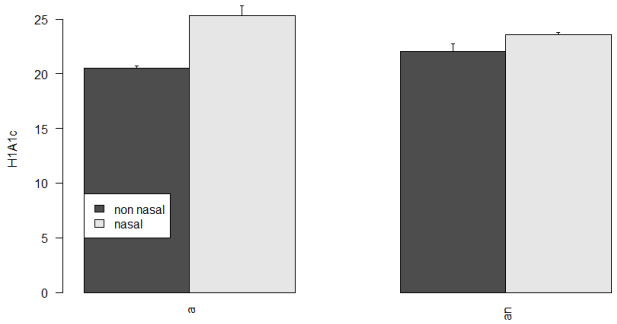


Figure 25: H1A1c selon la voyelle et la classe prédite

Une voyelle /a/ classée dans une catégorie "*non_nasal*" obtient une valeur de H1A1c égale à la valeur de voyelle /a/ comme ce qui est attendu. Au contraire, une voyelle orale /a/ identifiée comme nasale a une valeur de H1A1c plus élevée que la valeur de voyelle /a/ et proche de la valeur de la voyelle /ɑ̃/.

Quant à une voyelle nasale détectée correctement, une valeur de H1A1c élevée et proche de la valeur de voyelle /ɑ̃/ est observée. Une voyelle nasale identifiée comme orale obtient une valeur de H1A1c plus basse que ce qui est attendu, et plus proche de la voyelle /a/.

Pour la mesure H1A1c, une différence entre la voyelle orale /a/ et la voyelle nasale /ɑ̃/ est observée comme illustrée ci-dessus. Cette différence peut être due à la nasalité, mais aussi aux autres facteurs, par exemple, l'articulation qui diffère selon les voyelles et les contextes. Dans une suite donnée à cette étude, il sera donc nécessaire d'effectuer des mesures acoustiques en tenant compte des contextes de voyelle dans le but d'étudier si la mesure H1A1c change pour les différents types de voyelles comme les voyelles /a/ détectées correctement, les voyelles /a/ dans un contexte nasal ou les voyelles /ɑ̃/ détectées comme orales.

Nous avons vu que les mesures H1c et H1A1c permettent de vérifier la nasalité entre les voyelles orales et les voyelles nasales. Pour deux paramètres, les valeurs obtenues pour les catégories "nasal" et "non nasal" ressemblent à celles obtenues pour la voyelle /ɑ̃/ et la voyelle /a/ respectivement. De plus, quatre hypothèses que nous avons construites ont été bien répondues : la voyelle orale identifiée comme nasale aura une valeur analogue de la voyelle /ɑ̃/ et la voyelle nasale détectée comme orale obtiendra une valeur qui se rapproche de la voyelle /a/.

Dans cette section, nous souhaitons aborder les mêmes mesures acoustiques sur les locuteurs identifiés comme ayant une voix plus nasale que d'autres et sur les locuteurs identifiés comme ayant une voix la moins nasale en section 5.1.3. Trois locuteurs identifiés comme ayant une voix plus nasale que d'autres sont 05_12_07_nb1_1_16, 28_11_07_nb2_1_16 et 03_12_07_nb1_1_16.

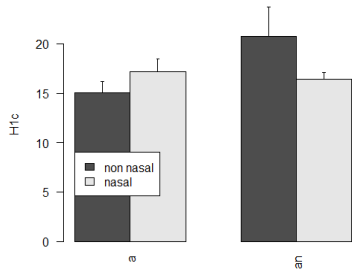


Figure 26: H1c selon la voyelle et la classe prédite pour le locuteur 05_12_07_nb1_1_16

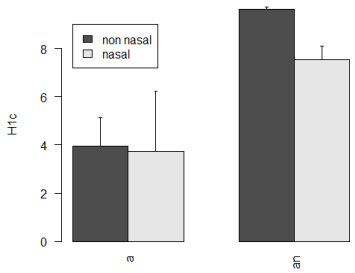


Figure 27: H1c selon la voyelle et la classe prédite pour le locuteur 28_11_07_nb2_1_16

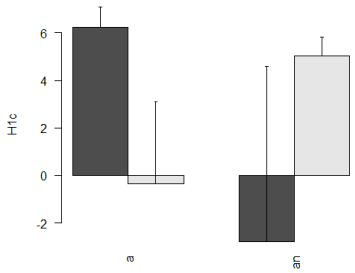


Figure 28: H1c selon la voyelle et la classe prédite pour le locuteur 03_12_07_nb1_1_16

Pour la mesure H1c, nous avons vu que la valeur de la voyelle /a/ se rapproche de 4 et celle de la voyelle /ã/ se situe autour de 1. De plus, la valeur diminue lorsqu'un son oral est identifié comme nasal et qu'au contraire la valeur augmente lorsqu'un son nasal est détecté comme oral.

Pour le locuteur 05_12_07_nb1_1_16, la valeur des voyelles orales et nasales se trouve très au-dessus des valeurs obtenues pour la moyenne des locuteurs. Lorsqu'un son oral est détecté comme oral, la voyelle /a/ obtient une valeur de H1c égale à 15 alors qu'un son oral identifié comme nasal, on observe une augmentation de la valeur. La voyelle /ã/ identifiée comme nasale montre une valeur de H1c aussi élevée que la voyelle /a/ bien identifiée. Pour la voyelle nasale classée comme orale, une augmentation de la valeur est observée.

Le locuteur 28_11_07_nb2_1_16 obtient, pour la voyelle /a/ correctement identifiée, une valeur similaire à la voyelle /a/ en moyenne et observe une diminution pour la voyelle orale détectée comme nasale. La voyelle nasale correctement classée obtient une valeur égale à 8 et la voyelle nasale détectée comme orale obtient une valeur plus élevée que ce qui a été attendu.

Pour le dernier locuteur 03_12_07_nb1_1_16, la valeur de la voyelle /a/ correctement identifiée est proche de la voyelle /a/ en moyenne, et nous observons que la valeur de la voyelle orale détectée comme nasale diminue et se rapproche de la valeur de la voyelle /ã/ en moyenne. La voyelle nasale identifiée comme nasale obtient une valeur un peu plus élevée que ce qui a été attendu pour la voyelle /ã/ correctement détectée et la voyelle nasale classée comme orale montre la diminution de valeur H1c contrairement à ce qui a été attendu pour la voyelle nasale identifiée comme orale.

En résumé, pour ces trois locuteurs, mis à part pour le locuteur 03_12_07_nb1_1_16 pour lequel des mesures très aléatoires sont obtenues, on observe en moyenne que les voyelles /ã/ ont des valeurs de H1c beaucoup plus élevées que la moyenne des locuteurs. Pour la voyelle /a/, à l'exception du locuteur 03_12_07_nb1_1_16 encore une fois, le /a/ détecté comme nasal a des valeurs de H1c équivalentes ou plus élevées que le /a/ détecté comme oral contrairement à ce qui était observé pour la moyenne des locuteurs.

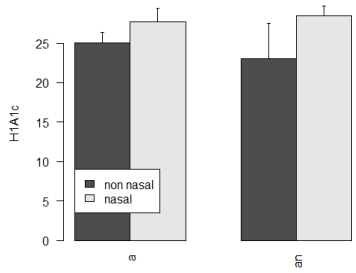


Figure 29: H1A1c selon la voyelle et la classe prédite pour le locuteur 05_12_07_nb1_1_16

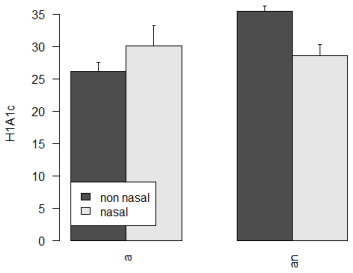


Figure 30: H1A1c selon la voyelle et la classe prédite pour le locuteur 28_11_07_nb2_1_16

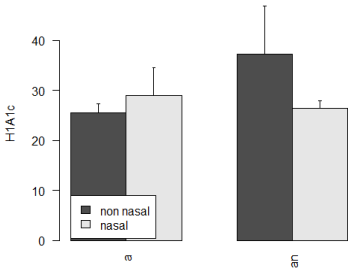


Figure 31: H1A1c selon la voyelle et la classe prédite pour le locuteur 03_12_07_nb1_1_16

Pour la mesure H1A1c, la valeur de la voyelle /a/ se rapproche de 20 et celle de la voyelle /ã/ est supérieure à 20. L'augmentation en valeur H1A1c est attendue pour une voyelle orale détectée comme nasale, et on s'attend à obtenir une valeur diminuée pour la voyelle nasale incorrectement identifiée.

Nous observons que, pour le locuteur 05_12_07_nb1_1_16, la voyelle /a/ identifiée comme non nasale obtient une valeur similaire à la voyelle /a/ en moyenne et la voyelle orale détectée comme nasale obtient une valeur H1A1c plus élevée que la voyelle orale bien identifiée comme attendue. La voyelle nasale classée comme nasale se rapproche de la valeur de la voyelle /ɑ̃/ en moyenne, et la voyelle nasale identifiée comme orale obtient une valeur plus basse et proche de la voyelle /a/ en moyenne.

Le locuteur 28_11_07_nb2_1_16 obtient une valeur analogue de la voyelle /a/ en moyenne pour la voyelle orale correctement identifiée et une valeur plus élevée et proche de la voyelle /ɑ̃/ en moyenne est attribuée pour la voyelle orale incorrectement identifiée. En ce qui concerne la voyelle /ɑ̃/ détecté comme nasale, une valeur proche de la voyelle /ɑ̃/ moyenne est observée alors que la voyelle nasale identifiée comme orale obtient une valeur plus élevée que la voyelle nasale correctement identifiée, ce qui n'a pas été attendu.

Le locuteur 03_12_07_nb1_1_16 observe la voyelle orale correctement identifiée proche de la voyelle /a/ en moyenne. La voyelle orale identifiée comme nasale obtient une valeur plus élevée que la voyelle orale bien identifiée, ce qui est attendu pour la voyelle orale incorrectement identifiée. La voyelle nasale identifiée comme nasale est proche de la voyelle /ɑ̃/ en moyenne alors qu'une forte augmentation est observée pour la voyelle nasale identifiée comme orale.

Pour ces trois locuteurs, les valeurs moyennes de H1A1c sont globalement plus élevées que la moyenne des locuteurs, ce qui coïncide avec une nasalité plus importante.

Les voyelles des locuteurs identifiés comme plus nasals que d'autres répondent partiellement à nos hypothèses émises sur les mesures acoustiques H1c et H1A1c. Lorsqu'une voyelle orale ou nasale est correctement identifiée, la valeur est généralement plus élevée en restant proche de la valeur moyenne obtenue à partir de l'ensemble de locuteurs. Pour les voyelles orales et nasales incorrectement identifiées, les valeurs varient en fonction des voyelles et du locuteur.

Nous nous intéressons ici aux locuteurs identifiés comme moins nasals que d'autres 28_11_07_nb1_1_16, 29_11_07_nb1_2_16 et 14_11_07_nb1_1_16.

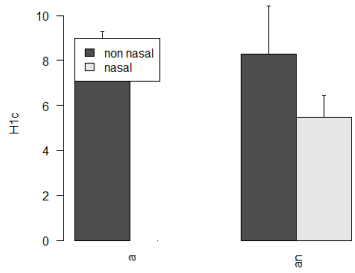


Figure 32: H1c selon la voyelle et la classe prédite pour le locuteur 28_11_07_nb1_1_16

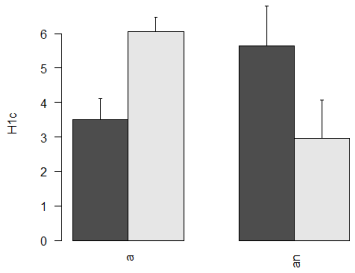


Figure 33: H1c selon la voyelle et la classe prédite pour le locuteur 29_11_07_nb1_2_16

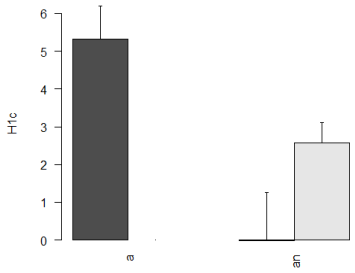


Figure 34: H1c selon la voyelle et la classe prédite pour le locuteur 14_11_07_nb1_1_16

Le locuteur 28_11_07_nb1_1_16 pour qui aucune voyelle orale n'est détectée comme nasale selon les réseaux, la valeur de H1c pour la voyelle /a/ détectée comme orale est supérieure à la voyelle /a/ en moyenne. La voyelle nasale /ã/ détectée comme nasale a une valeur plus élevée que la moyenne, et la valeur de H1c pour la voyelle nasale détectée comme non nasale est supérieure à celle du /ã/ détectée comme nasale. Cette augmentation est attendue pour la voyelle nasale identifiée comme orale.

Pour le locuteur 29_11_07_nb1_2_16, la voyelle orale détectée comme orale se situe autour de la valeur égale à 4, qui est proche de la voyelle /a/ en moyenne. La voyelle orale détectée comme nasale montre l'augmentation de sa valeur contrairement à ce qui a été attendu. La voyelle nasale correctement identifiée est proche de la voyelle /a/ moyenne tandis que la voyelle nasale incorrectement identifiée obtient une valeur plus élevée comme attendu.

Le dernier locuteur pour qui aucune voyelle orale détectée comme nasale n'est observée, la valeur de la voyelle /a/ est très augmentée comparé à la moyenne des locuteurs. Pour la voyelle nasale correctement détectée, elle se trouve généralement élevée comparée à la moyenne des locuteurs qui est égale à 1. La diminution est observée pour la voyelle nasale détectée comme orale.

Pour les trois locuteurs identifiés comme le moins nasals que d'autres, les voyelles orales correctement identifiées et les voyelles nasales détectées comme orales ont les valeurs de H1c globalement élevée à l'exception du locuteur 14_11_07_nb1_1_16 ayant une voyelle nasale détectée comme orale qui montre la diminution de valeur de H1c.

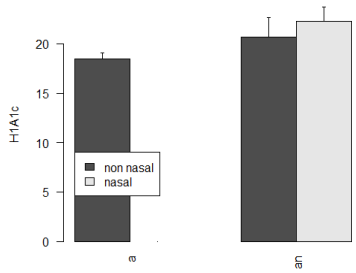


Figure 35: H1A1c selon la voyelle et la classe prédite pour le locuteur 28_11_07_nb1_1_16

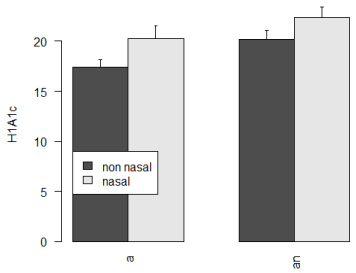


Figure 36: H1A1c selon la voyelle et la classe prédite pour le locuteur 29_11_07_nb1_2_16

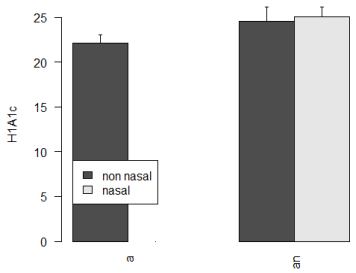


Figure 37: H1A1c selon la voyelle et la classe prédite pour le locuteur 14_11_07_nb1_1_16

Pour le locuteur 28_11_07_nb1_1_16, la voyelle orale correctement identifiée obtient une valeur proche de la voyelle /a/, moyenne des locuteurs. La voyelle nasale se situe entre deux voyelles /a/ et /ɑ̃/ en moyenne et la voyelle nasale identifiées comme orale se rapproche de la moyenne des voyelles /a/ des locuteurs, la valeur de H1A1c pour cette voyelle nasale a diminué comme attendu.

Le locuteur 29_11_07_nb1_2_16 obtient la valeur de la voyelle orale détectée comme non nasale proche de la voyelle /a/ en moyenne, il en est de même pour la voyelle nasale correctement identifiée qui se situe supérieure à 20 comme la voyelle /ɑ̃/ en moyenne. Pour la voyelle orale détectée comme nasale, la valeur H1A1c augmente alors que pour la voyelle nasale la valeur diminue comme attendu lorsque cette dernière est identifiée comme orale.

Le locuteur 14_11_07_nb1_1_16 obtient une voyelle orale identifiée comme orale qui se situe entre 20 et 25. Cette valeur est un peu élevée par rapport à la moyenne, mais elle reste autour de la moyenne. La voyelle nasale identifiée comme nasale est supérieure à 20 avec la voyelle nasale identifiée comme orale. Une diminution en valeur est observée pour cette voyelle.

Pour trois locuteurs, les valeurs de H1A1c sont autour de la moyenne des locuteurs, et les diminutions en valeur sont bien observées pour les voyelles nasales identifiées comme orales de trois locuteurs.

Les locuteurs identifiés comme le moins nasals que d'autres obtiennent les valeurs qui répondent à nos hypothèses émises sur les mesures acoustiques H1c et H1A1c et qui sont généralement situées autour de la moyenne des locuteurs.

5.5. Analyse perceptive

L'analyse perceptive permet de valider l'hypothèse selon laquelle les humains produisent des résultats semblables à ceux obtenus avec les réseaux de neurones convolutifs et de comprendre certaines erreurs dans la classification. En section 5.5.1, nous essayerons de déterminer si les voyelles orales et nasales incorrectement identifiées par le système de réseau de neurones convolutifs seraient détectées comme telles par la perception humaine. La section 5.5.2 consiste à amener une analyse à la fois perceptive et qualitative par nous-mêmes sur les mêmes voyelles correctement et incorrectement identifiées à l'aide des mesures acoustiques telles que la durée, la qualité de voix, l'intensité ou la f0.

5.5.1. Test de perception

Trente locuteurs natifs du français ont participé à notre test perceptif. Parmi l'ensemble des participants, vingt-sept personnes ont déjà suivi au moins un cours de linguistique et trois locuteurs n'ont pas eu d'expérience de le suivre. Deux exemples seront examinés pour chacun des groupes de voyelles incorrectement identifiées. Les tableaux présentés dans cette section sont un récapitulatif des réponses des participants sur deux voyelles de chaque groupe. Rappelons que chaque voyelle a été répétée deux fois pour vérifier la validité des réponses. Chacune des cinq colonnes (oral, nasal, /a/, /ã/ et autres) contient deux cases, la case gauche représente la réponse

des participants pour la première expérience sur une voyelle donnée et, dans la case droite, est présentée la réponse pour la même voyelle répétée (donc pour la deuxième expérience).

Seize extraits audio réalisés par neuf locuteurs distincts ont été utilisés pour créer le test de perception, le tableau ci-dessous est un récapitulatif de l'ensemble des locuteurs qui ont produit les voyelles utilisées dans le test de perception.

V nasales correctement identifiées	04_12_07_nb1_1_16, 04_12_07_nb1_1_16, 22_11_07_nb2_2_16, 29_11_07_nb1_2_16
V nasales incorrectement identifiées	04_12_07_nb1_2_16, 22_11_07_nb2_2_16, 26_11_07_nb3_1_16, 29_11_07_nb1_2_16
V orales correctement identifiées	04_12_07_nb2_2_16, 26_11_07_nb1_2_16, 28_11_07_nb2_1_16, 28_11_07_nb2_1_16
V orales incorrectement identifiées	04_12_07_nb2_2_16, 04_12_07_nb3_1_16, 26_11_07_nb1_2_16, 28_11_07_nb2_1_16

Table 36: Récapitulatif des locuteurs qui ont produit les voyelles utilisées dans le test de perception

5.5.1.1. Voyelles orales correctement identifiées par les CNNs

	Oral		Nasal		/a/		/ɑ̃/		Autres	
V orale 1	30	29	0	1	30	29	0	1	0	0
V orale 2	30	30	0	0	12	12	0	0	18	18

Table 37: Résultats pour deux voyelles orales correctes dans le test de perception

La voyelle 1 correctement classée par le système a été identifiée comme orale et perçue comme une voyelle /a/ à l'unanimité lors de la première expérience, et, lors de la deuxième expérience, selon vingt-neuf participants. Un locuteur a jugé la voyelle 1 comme étant nasale et une voyelle /ɑ̃/ et personne n'a entendu un son autre que les voyelles /a/ ou /ɑ̃/.

L'ensemble des participants ont perçu la voyelle orale 2 comme étant orale. La différence est observée entre deux voyelles orales correctes sur le choix entre /a/, /ɑ̃/ et "autres". Douze participants ont estimé entendre une voyelle /a/ pour la voyelle 2 dans les deux expériences et dix-huit locuteurs ont jugé que la voyelle /a/ été réalisée comme un autre son.

L'unanimité ou la majorité est observée pour le choix de la catégorie des voyelles orales correctes, presque tous les locuteurs ont estimé qu'il s'agissait d'une voyelle orale pour deux voyelles orales correctes malgré une variation éventuelle dans la réalisation.

5.5.1.2. Voyelles orales incorrectement identifiées par les CNNs

	Oral		Nasal		/a/		/ɑ̃/		Autres	
V orale 1	30	29	0	1	23	27	0	1	7	2
V orale 2	30	29	0	1	29	29	0	0	1	1

Table 38: Résultats pour deux voyelles orales incorrectes dans le test de perception

Pour la première voyelle orale détectée comme nasale selon le réseau de neurones convolutifs, tous les locuteurs l'entendent comme une voyelle orale dans le cadre de la première expérience et vingt-neuf pour la deuxième expérience. Un locuteur a jugé percevoir une voyelle nasale lors de la seconde expérience. La voyelle orale 1 a été perçue comme une voyelle /a/ à la majorité des participants (vingt-trois participants pour la première et vingt-sept pour la seconde expérience), et comme la réalisation d'une voyelle nasale /ɑ̃/ selon un participant lors de la deuxième expérience. Sept participants pour la première expérience et deux pour la deuxième ont estimé que la voyelle en question a été perçue comme une voyelle autre que la voyelle /a/ ou /ɑ̃/.

La seconde voyelle orale incorrectement classée selon le système a été identifiée comme étant orale par la plupart des locuteurs, soient trente locuteurs pour la première expérience et vingt-neuf pour la deuxième expérience sur la voyelle. Comme pour la voyelle orale 1, la voyelle orale 2 a été jugée comme une voyelle nasale selon un locuteur lors de deuxième expérience. Vingt-neuf personnes ont estimé entendre une voyelle orale /a/ dans l'ensemble des expériences, et un participant a entendu une voyelle autre que les voyelles /a/ ou /ɑ̃/. Personne n'a considéré la voyelle orale 2 comme la réalisation d'une voyelle /ɑ̃/.

La majorité des participants ont pu identifier les voyelles orales mentionnées dans le tableau 33 comme étant orales alors que celles-ci ont été détectées comme nasales par le système. Ces voyelles ont été perçues comme une voyelle /a/ plutôt que d'autres voyelles. Les variations qui peuvent être présentes dans la réalisation des voyelles orales n'empêchent pas la perception humaine de détecter la non-nasalité.

5.5.1.3. Voyelles nasales correctement identifiées par les CNNs

	Oral		Nasal		/a/		/ɑ̃/		Autres	
V nasale 1	2	1	28	29	1	0	19	21	10	9
V nasale 2	1	2	29	28	0	0	20	21	10	9

Table 39: Résultats pour deux voyelles nasales correctes dans le test de perception

La voyelle nasale 1 a été perçue comme orale selon peu de participants (deux pour la première expérience et un pour la seconde) et comme nasale pour la plupart des participants (vingt-huit pour la première expérience et vingt-neuf pour la seconde). La réalisation d'une voyelle /ɑ̃/ a été observée à une vingtaine de locuteurs (dix-neuf pour la première expérience et vingt-et-un pour la deuxième) alors que celle d'un son autre que les voyelles /a/ ou /ɑ̃/ a été observée par dix locuteurs lors de la première expérience et neuf lors de la deuxième expérience. La voyelle nasale 1 a également été entendue comme une voyelle /a/ selon un participant dans le cadre de la première expérience.

Quant à la voyelle nasale 2, peu de participants ont estimé que la voyelle entendue est orale (un pour la première expérience et deux pour la seconde) et la plupart l'ont jugée comme nasale (vingt-neuf et vingt-huit). Une vingtaine de locuteurs ont entendu la voyelle nasale 2 comme une voyelle /ɑ̃/ (vingt pour la première expérience et vingt-et-un pour la seconde) et une dizaine de participants l'ont perçue comme un son autre que les voyelles /a/ ou /ɑ̃/ (dix pour la première et neuf pour la seconde). Aucune des deux voyelles n'ont été entendues comme une voyelle /a/.

La plupart des participants ont montré leur capacité à identifier les voyelles nasales qui ont été également détectées comme nasales selon le réseau. certains ont estimé avoir entendu quelques variations des voyelles nasales, d'autre disent que la voyelle /ɑ̃/ a été perçue.

5.5.1.4. Voyelles nasales incorrectement identifiées par les CNNs

	Oral		Nasal		/a/		/ɑ̃/		Autres	
V nasale 1	0	1	30	29	0	0	30	30	0	0
V nasale 2	0	1	30	29	0	0	27	29	3	1

Table 40: Résultats pour deux voyelles nasales incorrectes dans le test de perception

En ce qui concerne la voyelle nasale 1 mal classée selon le système, elle a été entendue comme orale selon un locuteur lors de la deuxième expérience, et comme nasale à l'unanimité pour la première expérience et vingt-neuf pour la deuxième. La voyelle a été perçue comme une voyelle /ɑ̃/ à l'unanimité lors de deux expériences.

La voyelle nasale 2 a été identifiée comme nasale à l'unanimité lors de la première expérience et selon vingt-neuf participants lors de la seconde expérience. La majorité l'a jugé comme la réalisation d'une voyelle /ɑ̃/, soit vingt-sept locuteurs pour la première expérience et vingt-neuf pour la deuxième. La réalisation d'une autre voyelle a été observée par peu de participants (trois pour la première et un pour la seconde) et personne n'a perçu la voyelle nasale 2 comme une voyelle /a/ dans les deux expériences.

Aussi bien pour les voyelles nasales incorrectement classées que pour les voyelles nasales correctement classées, elles ont été correctement identifiées et entendues comme nasales selon la plupart des participants. Les deux voyelles nasales incorrectement identifiées par le système ont été entendues plus comme la réalisation d'une voyelle /ɑ̃/ pour les humains. Bien que la variation ait pu avoir lieu dans la réalisation des voyelles nasales /ɑ̃/, la perception humaine a pu réussir à reconnaître la nasalité dans ces voyelles nasales classées comme orales par le système.

5.5.2. Analyse qualitative

Nous souhaitons aborder une analyse qualitative sur les mêmes voyelles utilisées dans la section précédente afin d'essayer de comprendre l'erreur dans la classification. Une voyelle est suivie d'un numéro dans les tableaux présentés dans cette section, les voyelles ayant le même numéro ont été produits par le même locuteur ; par exemple, deux voyelles notées V1 dans le tableau Voyelles orales ont été réalisé par le locuteur '04_12_07_nb2_2_16' et V2 par le locuteur '28_11_07_nb2_1_16'. L'information sur le locuteur est fournie dans la légende du tableau. Le fait d'avoir le même locuteur pour deux voyelles nous donnera un point de comparaison.

5.5.2.1. Voyelles orales

	Sexe	Durée	Intensité max	Fréquence fondamentale max	Contextes	Qualité de voix	Probabilité à la classe "orale"
V1 correcte	Homme	250 ms	73 dB	130 Hz	labial_coronal		100%
V1 incorrecte		190 ms	65 dB	136 Hz	labial_pause	Soufflée	4%
V2 correcte	Homme	150 ms	68 dB	140 Hz	labial_dorsal		99%
V2 incorrecte		170 ms	60 dB	103 Hz	coronal_dorsal		0.1%

Table 41: Résultat pour deux locuteurs (V1 pour 04_12_07_nb2_2_16 et V2 pour 28_11_07_nb2_1_16)

Quant aux voyelles orales 1 produites par le locuteur 04_12_07_nb2_2_16, la voyelle correctement identifiée a une durée plus longue que la voyelle mal identifiée (250 ms pour la voyelle correcte et 190 ms pour incorrecte) ainsi que l'intensité maximum plus élevée (73 dB pour correcte et 65 dB pour incorrecte). Deux voyelles ont une valeur similaire pour la fréquence fondamentale (130 Hz et 136 Hz pour la voyelle correcte et incorrecte respectivement) et le même contexte gauche étant "labial". Puis la probabilité d'appartenance à la classe "non-nasal" varie de 4% à 100% selon le statut de la voyelle. Avec la pause qui suit la voyelle 1 incorrecte et la voix soufflée, nous pouvons supposer que le locuteur a abaissé le voile du palais vers la fin du mot et que la voix soufflée, même très faible, peut être comprise comme nasale selon le système.

Les voyelles 2 ont une probabilité qui varie de 0,1% à 99% (pour la voyelle incorrecte et correcte respectivement) et partagent le même élément de droit dans leurs contextes, soit “dorsal”. La voyelle 2 correcte a une durée plus courte (150 ms) que la voyelle 2 incorrecte (170 ms), ainsi que l’intensité et la fréquence fondamentale (140 Hz et 68 dB pour la voyelle correcte et 103 Hz et 60 dB pour la voyelle incorrecte). À partir des résultats obtenus pour ces voyelles 2, nous supposons que la fréquence fondamentale basse a posé une difficulté pour classifier la voyelle 2 incorrecte selon le système.

Nous avons repéré le contexte “pause” et la voix soufflée comme introduisant une difficulté au système lors de la classification des voyelles orales, ces derniers peuvent donner une impression de nasalité aux réseaux de neurones convolutifs. La voyelle orale peut être également identifiée comme nasale lorsqu’elle a une fréquence fondamentale basse.

5.5.2.2. Voyelles nasales

	Sexe	Durée	Intensité max	Fréquence fondamentale max	Contextes	Probabilité à la classe “nasale”
V1 correcte	Femme	190 ms	69 dB	441 Hz	coronal_pause	99%
V1 incorrecte		150 ms	70 dB	217 Hz	pause_dorsal	16%
V2 correcte	Femme	250 ms	72 dB	275 Hz	dorsal_coronal	99%
V2 incorrecte		170 ms	75 dB	252 Hz	coronal_pause	35%

Table 42: Résultat pour deux locuteurs (V1 pour 22_11_07_nb2_2_16 et V2 pour 29_11_07_nb1_2_16)

Les voyelles nasales 1 ont été réalisées par le locuteur 22_11_07_nb2_2_16 et n’observent pas une différence au niveau de l’intensité (69 dB pour la voyelle correcte et 70 dB pour incorrecte). La probabilité se fluctue entre 99% et 16% selon le statut de la voyelle. La voyelle 1 correcte a une durée de 190 ms qui est plus longue que la voyelle 1 incorrecte ayant une durée de 150 ms. La voyelle 1 correcte observe une fréquence fondamentale de 441 Hz avec comme contextes “coronal_pause” tandis que la voyelle 1 incorrecte obtient une fréquence fondamentale égale à 217 Hz avec comme contextes “pause_dorsal”. Nous supposons que la fréquence fondamentale n’a pas joué de rôle important dans la classification des voyelles nasales car une voyelle ayant une fréquence fondamentale élevée a été bien identifiée par le système.

En ce qui concerne les voyelles 2 produites par le locuteur 29_11_07_nb1_2_16, la voyelle correcte est plus longue (250 ms) en terme de durée et plus élevée (275 Hz) en terme de fréquence fondamentale comparée à la voyelle incorrecte (170 ms et 252 Hz respectivement). La voyelle correcte obtient une intensité maximum de 72 dB et la voyelle incorrecte de 75 dB et la probabilité se situe entre 99% et 35% en fonction du statut de la voyelle (correcte et incorrecte respectivement). Aucun élément de gauche ou de droit n’est observé en commun pour ces deux voyelles. Dans ce cadre, nous ne saurons pas dans quelle voyelle l’intensité ou la fréquence fondamentale a une valeur élevée ou basse.

La différence observée en commun dans les voyelles réalisées par deux locuteurs distincts est la durée : les voyelles correctes 1 et 2 sont plus longues que les voyelles incorrectes 1 et 2. Nous pouvons alors supposer qu'une courte durée pour une voyelle nasale peut provoquer une difficulté au réseau au cours d'une classification.

5.6. Discussion

Deux objectifs principaux ont été introduits lors des expériences abordées dans les diverses sections précédentes : créer une mesure de coarticulation nasale selon le locuteur et vérifier la généralisation du système de réseaux de neurones convolutifs. Et une analyse perceptive permettant de comprendre les erreurs de classification vient à répondre l'hypothèse selon laquelle la perception humaine retombe sur les résultats des réseaux de neurones convolutifs.

5.6.1. Interprétation des résultats pour la caractérisation des locuteurs

Les voyelles orales détectées comme étant nasales et les voyelles nasales identifiées correctement sont de meilleurs paramètres pour dresser une fiche de nasalité par locuteur car le nombre total de ces dernières peut nous amener à identifier quels locuteurs obtiennent plus de nasalité dans la parole que d'autrui. Par ailleurs, le fait de définir deux types de voyelles orales détectées comme nasales permet une analyse plus fine sur les résultats obtenus pour l'expérience en item "*locuteur*" (section 5.1.3) car le phénomène de coarticulation peut ne pas être pris en compte dans l'analyse.

	Faux négatifs dans un contexte oral	Faux négatifs dans un contexte nasal	Vrais négatifs
28_11_07_nb2_1_16	7	0	53
29_11_07_nb2_2_16	6	1	53
05_12_07_nb1_1_16	8	1	53
14_11_07_nb2_1_16	6	1	53
03_12_07_nb1_1_16	7	0	53
04_12_07_nb2_2_16	6	2	52
28_11_07_nb1_1_16	0	0	47
29_11_07_nb1_2_16	1	1	39
14_11_07_nb1_1_16	0	0	41

Table 43: Nombre de faux négatifs et de vrais positifs pour les occurrences de voyelles /a/-/ã/ selon neuf locuteurs

Le tableau présenté ci-dessus est un récapitulatif des résultats obtenus pour neuf locuteurs ayant le plus de voyelles orales et nasales confondues qui sont détectées comme nasales. Lorsque les faux négatifs dans un contexte nasal sont exclus dans le comptage, trois locuteurs (05_12_07_nb1_1_16, 28_11_07_nb2_1_16, et 03_12_07_nb1_1_16) peuvent être repérés parmi l'ensemble des locuteurs. Ils peuvent être considérés comme les locuteurs dont la voix est plus nasale comparés aux autres locuteurs car la nasalité est sur-représentée dans la classification de leurs voyelles, soient soixante-et-un occurrences de voyelles détectées comme nasales pour le locuteur 05_12_07_nb1_1_16, soixante pour les locuteurs 28_11_07_nb2_1_16 et 03_12_07_nb1_1_16 sans compter les voyelles orales identifiées comme nasales dans un contexte nasal. En outre, dans cette expérience, nous avons également pu repérer les locuteurs les moins identifiés comme nasals (28_11_07_nb1_1_16, 29_11_07_nb1_2_16 et 14_11_07_nb1_1_16), ils obtiennent le nombre de vrais positifs et de faux négatifs le plus bas parmi l'ensemble des locuteurs. Autrement dit, le nombre de voyelles orales et nasales détectées comme orales le plus élevé est observé pour ces locuteurs.

Concernant l'item "tronçon" avec lequel nous avons souhaité étudier le phénomène de coarticulation, parmi les trois parties du son, le deuxième tiers correspondant au milieu de la voyelle présente dans l'ensemble le taux de bonnes classifications le plus élevé et stable variant de 76% à 89% selon les modèles. Le dernier tiers obtient les scores analogues que ceux du deuxième tiers. Le premier tiers représente le taux d'exactitude le plus bas dans tous les modèles abordés situé entre 73% et 83%. Ces résultats nous permettent de supposer qu'il s'agirait d'une coarticulation progressive que nous observons. Si le deuxième et le troisième partagent les scores très proches et élevés, on peut s'attendre à l'existence de la nasalité dans ces deux tiers qui est un bon indicateur pour caractériser les locuteurs plutôt que dans le premier tiers. Ce phénomène peut s'expliquer par le mouvement du voile du palais qui peut être un peu lent selon les locuteurs pour fermer le port pharyngé. Pourtant, ces résultats peuvent être simplement dus aux imprécisions dans la segmentation automatique qui a laissé des bouts de consonnes et qui provoque des erreurs dans la classification du système.

Cependant, certaines conditions du son telles que le contexte, le sexe, et les mesures acoustiques (la durée, la fréquence fondamentale et l'intensité) peuvent avoir un impact sur la gestion de

classification du système de réseaux de neurones convolutifs permettant de caractériser les locuteurs.

L'analyse en contexte s'est réalisée à l'aide de deux voyelles /a/ et /ɑ̃/. Quant à la voyelle orale /a/, le système tend à prédire la voyelle de manière incorrecte lorsqu'elle se situe devant une pause ou une consonne dorsale. Les contextes dorsal_pause (20.9% sur la totalité de voyelles dans ce contexte), dorsal_dorsal (13.15%) et labial_dorsal (9.65%) rendent la voyelle plus difficile à être identifiée correctement par le système. Nous pouvons émettre l'hypothèse que certains locuteurs abaissent le voile du palais à la fin d'une phrase par le relâchement ou le repos, ce qui peut donner un timbre nasalisé, et expliquer que le réseau les classifie comme nasal. Pour la voyelle nasale /ɑ̃/, le contexte coronal_dorsal influence le plus dans la gestion de classification du réseau de neurones convolutifs. Lorsque la voyelle est dans un contexte coronal_dorsal, une mauvaise prédiction est observée sur dix occurrences de voyelles par le système, ce qui correspond à une proportion de 10.75% sur l'ensemble de voyelles dans ce contexte. Nous supposons que le contexte coronal_dorsal peut provoquer une difficulté dans la classification du système de réseaux de neurones convolutifs en introduisant la non-nasalité au cours d'une voyelle nasale.

L'influence du contexte a également été observée selon le sexe du locuteur en section 5.2.2. Pour les femmes, le contexte dorsal_pause représente une proportion de 17.5% sur l'ensemble de voyelles orales dans ce contexte, mais deux contextes présentés ci-dessus (dorsal_dorsal et labial_dorsal) pour la voyelle orale /a/ n'obtiennent pas de proportion supérieure à 10%, mais 5.3% et 6.5% respectivement. Pour les hommes, les trois contextes obtiennent une proportion plus élevée que pour les femmes : dorsal_pause à 22.8% sur la totalité des voyelles dans le même contexte, dorsal_dorsal à 21.1% et labial_dorsal à 12.5%. Nous pouvons remarquer la différence importante entre la proportion des erreurs de classification sur l'ensemble de voyelles orales /a/ chez les hommes et chez les femmes. Une proportion égale à 14.1% sur le nombre total des voyelles /a/ est observée pour les hommes et celle égale à 7.8% pour les femmes. C'est-à-dire que le système identifie les voyelles orales /a/ comme nasales deux fois plus fréquemment chez les hommes que chez les femmes. Concernant la voyelle nasale /ɑ̃/, une proportion de voyelles nasales mal identifiées est égale à 11.8% pour les femmes et 10% pour les hommes sur l'ensemble de voyelles nasales dans le contexte "coronal_dorsal". Une différence sur la

proportion de mauvaises classifications sur la totalité des voyelles nasales est également observée entre les hommes et les femmes. Contrairement à la voyelle /a/, la voyelle /ɑ̃/ est classée incorrectement plus chez les femmes que chez les hommes avec une différence moins importante : 8.7% de voyelles nasales pour les femmes et 6.5% pour les hommes. Dans l'ensemble des voyelles orales et nasales confondues, 7.76% des voyelles sont mal identifiées pour les hommes et 5.88% pour les femmes. Cela signifie que le sexe peut avoir une influence sur la classification et que le système tend à détecter les voyelles de manière incorrecte chez les hommes plus souvent que chez les femmes.

Nous pouvons supposer que les hommes produisent la nasalité lors de la réalisation des voyelles orales, surtout pour les voyelles suivies d'une pause ou d'une consonne dorsale. Cette hypothèse peut s'adjoindre à l'hypothèse que nous avons émise dans le cadre de l'analyse en contexte, l'abaissement du voile du palais vers la fin d'une phrase par le relâchement s'effectuera plutôt par les hommes que par les femmes. Le mouvement du voile du palais pour la fermeture du port pharyngé serait plus rapide pour les femmes étant identifiées comme moins nasales que les hommes.

Trois variantes de modèle du réseau de neurones convolutifs ont été initialisées dans le cadre de l'analyse sur la durée (section 5.2.3), de la fréquence fondamentale (section 5.2.4) et de l'intensité (section 5.2.5). Ils permettent d'observer une tendance des réseaux de neurones convolutifs à produire des erreurs sur les voyelles étudiées. Rappelons que le modèle 1 est entraîné et testé sur deux voyelles /a/ et /ɑ̃/, le modèle 2 entraîné avec deux voyelles /a/, /ɑ̃/ et testé sur quatre voyelles /a/, /ɑ̃/, /ɛ/ et /ɛ̃/ et le modèle 3 entraîné et testé sur ces dernières.

La durée se voit donner une influence dans la gestion de classification du système de réseau de neurones convolutifs. Lorsque les voyelles orales telles que /a/ et /ɛ/ ont une durée inférieure à la moyenne, elles tendent à être identifiées comme nasales par les modèles. Le modèle 1 l'observe chez trente-deux locuteurs pour la voyelle /a/, et le modèle 2 chez vingt-quatre locuteurs pour la même voyelle et quarante-cinq locuteurs pour la voyelle /ɛ/. Chez trente-et-un locuteur, la voyelle /ɛ/ ne connaît pas d'erreurs selon le modèle 3 mais la tendance "courte" reste observée pour cette voyelle chez neuf locuteurs. La même tendance est observée pour la voyelle /a/ chez vingt locuteurs selon le modèle 3. Contrairement aux voyelles orales, les voyelles nasales /ɑ̃/ et /ɛ̃/ obtiennent majoritairement la tendance dite "long". Le système a une forte chance de produire

des erreurs face aux voyelles nasales dont la durée est plus longue que la durée moyenne. Cette tendance est observée chez dix-neuf locuteurs pour la voyelle /ɑ̃/ selon le modèle 1, et chez vingt-trois locuteurs pour la même voyelle et chez quarante-et-un locuteurs pour la voyelle /ɛ̃/ selon le modèle 2. Lorsque deux voyelles /ɛ/ et /ɛ̃/ ont été insérées dans l'entraînement comme nous pouvons l'observer dans le modèle 3, deux tendances peuvent être observées selon les voyelles : être identifiée correctement pour la voyelle /ɑ̃/ (chez trente-et-un locuteurs) et être détectée comme orale pour la voyelle /ɛ̃/ (chez vingt-neuf locuteurs).

Il est cependant à noter qu'il est normal que l'on observe moins de voyelles dont la durée est égale à la moyenne que des voyelles courtes ou longues car la tendance "égal" n'est pas calculée dans une fourchette des valeurs, mais la durée doit être parfaitement égale à la moyenne pour faire partie de cette tendance. Nous essayerons donc de ne comparer que deux tendances "court" et "long" dans cette partie. Les erreurs tendent à se produire par le réseau de neurones convolutifs lorsqu'une voyelle orale a une durée courte (ou inférieure à la moyenne) et qu'une voyelle nasale a une durée longue (ou supérieure à la moyenne). En outre, les résultats obtenus pour les images découpées nous amènent à valider la tendance "long" pour les voyelles nasales. trente-six locuteurs pour la voyelle /ɑ̃/ sont observés dans une tendance "long" dans les résultats obtenu avec le modèle 1. La même tendance est observée aussi bien pour le modèle 2 (trente-six pour la voyelle /ɑ̃/ et quarante-deux pour la voyelle /ɛ̃/) que pour le modèle 3 (trente-quatre pour la voyelle /ɑ̃/, vingt-quatre pour la voyelle /ɑ̃/).

Il est possible que les durées intrinsèques des voyelles orales vs. nasales (environ 95 ms en moyenne pour la voyelle /a/ et 77 ms en moyenne pour la voyelle /ɑ̃/ dans notre étude) aient un impact sur ces résultats. Comme le nombre des voyelles orales est supérieur au nombre des voyelles nasales, toutes les voyelles nasales ont été extraites à condition que leur durée n'excède pas 250 ms. Nous pouvons supposer que les voyelles nasales extraites ont une durée très courte ou relativement courte aux voyelles orales comme nous pouvons le remarquer avec les résultats pour la durée en moyenne ci-dessus et que les réseaux les ont utilisées pour faire l'entraînement et la prédiction. Il paraît normal que les réseaux aient une grande chance de se tromper sur les voyelles nasales longues lorsqu'ils sont entraînés avec les nasales courtes.

On peut aussi dire que les voyelles orales courtes (notamment dans un contexte dorsal ou coronal) sont fréquemment réalisées comme des voyelles centrales à cause de la coarticulation.

On ne sait pas pour les voyelles nasales longues, mais il est possible que leurs propriétés articulatoires soient également altérées.

La fréquence fondamentale au milieu de la voyelle peut être un paramètre permettant de gérer la classification du réseau de neurones convolutifs. Pour les images non découpées, le réseau tend à commettre des erreurs pour prédire la classe d'une voyelle lorsque celle-ci obtient une fréquence fondamentale élevée (ou supérieure à la moyenne). Cette tendance "élevé" est observée pour toutes les voyelles de trois variantes de modèles, c'est-à-dire qu'il est fréquent que les voyelles soient incorrectement classées lorsque la fréquence fondamentale au milieu de la voyelle est élevée. La voyelle nasale /ɑ̃/ du modèle 1 obtient une tendance d'erreurs dite "élevé" chez vingt-quatre locuteurs tandis que la voyelle orale /a/ du même modèle obtient deux tendances "basse" et "élevé" chez dix-huit locuteurs. La tendance "élevé" est observée pour les voyelles /a/ (chez dix-neuf locuteurs), /ɑ̃/ (chez dix-neuf locuteurs), /ɛ/ (chez vingt locuteurs) et /ɛ̃/ (chez trente-six locuteurs). Autre que la tendance "null" pour les voyelles /ɑ̃/ et /ɛ/ sur lesquelles le modèle 3 ne fait aucune erreur chez vingt-et-un locuteurs et trente-et-un locuteurs respectivement, les erreurs tendent à se produire face aux voyelles /a/ et /ɛ̃/ lorsqu'elles ont une fréquence fondamentale élevée. Pour les images découpées en trois tronçons, la fréquence fondamentale élevée qui est calculée au milieu de la voyelle implique plus fortement la gestion de classification sur les voyelles orales et nasales confondues. Le modèle 1 fait de mauvaises classifications sur les voyelles /a/ (chez vingt-neuf locuteurs) et /ɑ̃/ (chez trente-huit locuteurs) lorsqu'elles ont une fréquence fondamentale élevée qui est calculée au milieu de la voyelle. Nous pouvons observer la même tendance aussi bien pour les voyelles du modèle 2 (vingt-neuf pour /a/, trente-trois pour /ɑ̃/, trente-deux pour /ɛ/ et trente-huit pour /ɛ̃/) que pour celles du modèle 3 (vingt-cinq pour /a/ et /ɑ̃/, vingt-huit pour /ɛ/ et trente pour /ɛ̃/).

Comparée à la fréquence fondamentale au milieu de la voyelle, la f0 moyenne n'apporte pas un impact aussi significatif aux réseaux dans la phase de classification. Les résultats obtenus pour les images non découpées ne mettent pas en évidence d'une seule tendance pour chacune des voyelles étudiées. La tendance d'erreurs se fluctue entre "bas" et "élevé" selon les modèles initialisés. Cependant, les images découpées en trois tronçons montrent une tendance plus stable que les images non découpées, les voyelles autres que la voyelle /a/ tendent à être incorrectement identifiées par les réseaux lorsque celles-ci ont une intensité moyenne élevée. Le modèle 1

l'observe pour la voyelle /ɑ̃/ chez trente-cinq locuteurs, le modèle 2 pour la même voyelle chez vingt-sept locuteurs, pour la voyelle /ɛ/ chez vingt-quatre locuteurs et pour la voyelle /ɛ̃/ chez trente-six locuteurs. Idem pour le modèle 3, ces trois voyelles /ɑ̃/, /ɛ/ et /ɛ̃/ sont souvent concernés avec la mauvaise classification lorsqu'elles ont une intensité moyenne élevée, ce qui est observé chez vingt-trois, vingt-six et vingt-huit locuteurs respectivement. Contrairement aux trois voyelles mentionnées ci-dessus, la voyelle /a/ se trouve dans deux tendances "bas" et "élevé" chez une vingtaine de locuteurs à chaque modèle.

En ce qui concerne l'intensité, une tendance forte est observée pour les voyelles nasales selon deux mesures (au milieu de la voyelle et moyenne). Aussi bien pour les images non découpées que pour les images découpées, les voyelles nasales sont fréquemment identifiées de manière incorrecte lorsque ces dernières obtiennent une intensité élevée, que ce soit au milieu de la voyelle ou moyenne. Au contraire, les voyelles orales observent deux tendances distinctes. Pour la voyelle /a/, les erreurs tendent à se produire lorsqu'elle a une intensité basse (au milieu de voyelle ou moyenne) tandis que pour la voyelle /ɛ/, la tendance se situe entre "bas" et "élevé".

Parmi les mesures acoustiques telles que la durée, la fréquence fondamentale et l'intensité, seule la durée peut permettre de gérer la classification selon la nasalité pour les images non découpées. Puisque les erreurs se trouvent fréquemment dans les voyelles orales courtes et dans les voyelles nasales longues, les voyelles orales dont la durée est supérieure à la moyenne et les voyelles nasales dont la durée est inférieure à la moyenne peuvent être considérées comme de bons sujets d'extraction. Par ailleurs, entre la fréquence fondamentale au milieu de la voyelle et celle moyenne, la première peut être considérée comme meilleure car il est fréquent que les erreurs se produisent dans la plupart des voyelles étudiées dont la fréquence fondamentale au milieu de voyelle élevée. Quant à l'intensité au milieu de la voyelle et moyenne, les voyelles /a/ produites avec une intensité basse sont difficiles à identifier selon les réseaux tandis que les trois autres voyelles sont considérées comme difficiles lorsqu'elles ont une intensité forte.

5.6.2. Validation acoustique et perceptive

Les mesures acoustiques H1c et H1A1c nous ont permis de vérifier la nasalité dans les voyelles orales et nasales. Les valeurs obtenues pour les voyelles classées comme orales et comme

nasales se montrent dans la même tendance que pour les voyelles /a/ et /ã/ respectivement. De plus, les voyelles orales détectées comme nasales ont une valeur qui se rapprochent de la valeur de la voyelle /ã/ et les voyelles nasales identifiées comme orales obtiennent une valeur semblable à celle de la voyelle /a/. Ces faits nous permettent de dire que les réseaux ont extrait des paramètres pertinents pour classer les voyelles selon leur nasalité.

Ces mesures ont été utilisées pour vérifier la nasalité dans les six locuteurs : trois identifiés comme plus nasals que d'autres ; trois identifiés comme moins nasals que d'autres.

Pour les trois premiers locuteurs (identifiés comme nasals), les valeurs de H1c pour les voyelles /ã/ étaient très élevées que la moyenne des locuteurs à l'exception du locuteur 03_12_07_nb1_1_16 pour lequel des mesures très aléatoires sont obtenues. Pour la voyelle /a/, la voyelle orale identifiée comme nasale a une valeur élevée ou équivalente à la voyelle orale correctement identifiée, ce qui n'a pas été attendu pour cette voyelle. En ce qui concerne la mesure de H1A1c pour ces locuteurs, les valeurs de toutes les voyelles sont élevées. Pour cette mesure, nous avons émis une hypothèse selon laquelle le son oral identifié comme nasal a une augmentation de valeur. Avec cette hypothèse, nous pouvons également supposer que la nasalité est présente dans les voyelles de ces trois locuteurs car elles ont une valeur très élevée.

Pour les trois locuteurs identifiés comme le moins nasals que d'autres, les voyelles orales correctement identifiées et les voyelles nasales détectées comme orales ont les valeurs de H1c globalement élevées. Cette augmentation affectant à la fois les voyelles orales correctement identifiées et les voyelles nasales identifiées comme orales peut introduire la non-nasalité dans les voyelles des locuteurs identifiés comme moins nasals que d'autres. Quant à la mesure de H1A1c, les valeurs sont autour de la moyenne des locuteurs, et les diminutions en valeur sont bien observées pour les voyelles nasales identifiées comme orales de trois locuteurs.

En comparant l'analyse perceptive avec celle acoustique, la perception humaine a pu retomber sur les résultats des réseaux de neurones convolutifs sur les voyelles correctes, mais pas sur les voyelles incorrectes. Les voyelles orales incorrectement détectées par le système ont été toutes identifiées comme orales et comme une voyelle /a/ par les participants et il en était de même pour les voyelles nasales incorrectement identifiées. Par l'analyse qualitative, nous avons repéré le contexte "pause", la voix soufflée et une fréquence fondamentale basse comme provoquant une difficulté dans le cadre de la classification du système. En outre, les humains ont pu

correctement identifier les voyelles nasales bien classées par le système, mais entendent parfois une variation des voyelles nasales. Quant aux voyelles nasales /ɑ̃/ incorrectement détectées par le réseau, les voyelles nasales ont été perçues comme nasales et comme une voyelle /ɑ̃/ selon les participants. Ces voyelles ayant une durée relativement courte ont été identifiées comme non nasales selon le réseau, on peut s'attendre à ce que le système se trompe sur les voyelles nasales courtes.

	min	max	moyenne
8 V correctes	99%	100%	99%
Toutes V correctes	51%	100%	98%

Table 44: Valeurs de probabilités pour les huit voyelles correctes du test de perception et toutes les voyelles

Nous souhaitons regarder sur les valeurs des probabilités d'appartenances pour les huit voyelles correctement identifiées utilisées dans le test de perception avec celles pour toutes les voyelles correctes existant dans notre corpus. La probabilité d'appartenances pour les huit voyelles correctement identifiées se situe entre 99% et 100% avec une moyenne de 99% tandis que celle pour l'ensemble des voyelles correctes se fluctue entre 51% et 100% avec une moyenne de 98%. En ce qui concerne les huit voyelles, il est probable qu'aucune ou très peu de caractéristiques qui introduisent des difficultés dans la classification existent dans la production de ces sons car nous voyons que le système a réussi à bien classer les voyelles. Au contraire, dans le cas de l'ensemble des voyelles, nous supposons que pour les voyelles avec une probabilité de 51%, les réseaux ont beaucoup de difficultés pour les détecter et qu'il peut y avoir des caractéristiques acoustiques tels que la durée, l'intensité ou la fréquence fondamentale qui provoquent souvent les erreurs au cours d'une classification. Cette erreur n'est pas importante dans l'ensemble comme nous pouvons voir avec la probabilité en moyenne qui atteint 98% de bonne classification.

5.6.3. Tentatives de généralisation et améliorations futures

Nous avons initialisé quatre variantes de modèle pour chacun des trois jeux de données dans le but d'établir si les modèles entraînés avec les données de quarante locuteurs sont généralisables sur celles de cinq locuteurs non vues par le système.

	Jeu de donnée 1	Jeu de donnée 1 (images découpées)	Jeu de donnée 2	Jeu de données 2 (images découpées)	Jeu de donnée 3	Jeu de données 3 (images découpées)
Accuracy du modèle original	93%	86%	80%	75%	89%	82%
Accuracy moyen des modèles de test	92.5%	84%	86.25%	81%	87.25%	80.75%

Table 45: Récapitulatif des résultats obtenus pour l'expérience de généralisation

Le tableau ci-dessus est un récapitulatif des résultats obtenus pour chaque jeu de données dans le cadre de l'expérience de généralisation des modèles. Pour le jeu de données noté 1 dans le tableau et constitué avec deux voyelles /a/ et /ɑ̃/ et celui noté 3 et constitué avec /a/, /ɑ̃/, /ɛ/ et /ɛ̃/, les résultats obtenus pour les modèles de test sont stables (92.5% pour le jeu de données 1, 84% pour le jeu de données 1 avec des images découpées, 87.25% pour jeu de données 3 et 80.75% pour le jeu de données 3 avec des images découpées). C'est-à-dire que les modèles entraînés avec les données de quarante locuteurs se généralise bien sur celles de cinq locuteurs non vus lorsque les voyelles entraînées et testées sont identiques. Avec le jeu de données 2, les modèles de test ont été entraînés avec deux voyelles /a/ et /ɑ̃/ et testés sur quatre voyelles /a/, /ɑ̃/, /ɛ/ et /ɛ̃/. Les modèles de test obtiennent de meilleurs résultats que le modèle original, ce qui permet de différencier le jeu de données 2 et les autres. Ils sont exactes à 86.25% pour les images non découpées et 81% pour les images découpées tandis que le modèle obtient 80% d'exactitude pour les images non découpées et 75% pour les images découpées. Les modèles entraînés avec deux voyelles /a/ et /ɑ̃/ de quarante locuteurs réussissent à généraliser les voyelles /a/, /ɑ̃/, /ɛ/ et /ɛ̃/ de cinq locuteurs non vus à quel point que les résultats obtenus pour les modèles de test sont meilleurs que ceux obtenus avec le modèle original.

Certaines analyses peuvent être améliorées. L'une des améliorations futures concerne les mesures acoustiques pouvant influencer la gestion de classification. Nous avons initialisé trois

types de tendance “courte/bas”, “égal” et “long/élevé” afin de mesurer dans quelle condition l’erreur tend à se produire lors de la prédiction. Pour deux tendances “courte/bas” et “long/élevé”, la valeur des mesures acoustiques a été comparée avec la moyenne tandis que pour la tendance “égal”, la valeur a dû être parfaitement égale à la moyenne. Il était donc normal que le nombre de locuteurs concernés par la tendance “égal” soit toujours inférieur aux autres tendances. Pour équilibrer les tendances, soit la tendance “égal” devrait être calculée dans une fourchette, soit la valeur en pourcentage devrait être sujet de comparaison pour désigner une tendance à un locuteur plutôt que de comparer le nombre net des voyelles incorrectement identifiées selon chaque tendance.

L’outil pour créer un questionnaire peut être également amélioré. Le logiciel “Google Form” a été mis en œuvre pour créer le test de perception car il s’agissait d’un seul logiciel de questionnaire que nous connaissions et que, par manque de temps, nous n’avons pas pu découvrir d’autres logiciels. “Google Form” ne permet pas d’insérer un fichier audio dans une question et de l’écouter sans charger, le fichier son sauvegardé dans “Google Drive” a donc été accessible via un lien donné et ouvert dans un nouvel onglet une fois le lien cliqué. Cette façon ne permettait pas un bon déroulement de test car chaque question demandait un temps pour le chargement du fichier audio. Au lieu de “Google Form”, le logiciel “Ibex Farm” dont nous n’avions pas connaissance peut être plus utile pour créer une expérience en ligne.

Enfin, les voyelles incorrectement identifiées par les réseaux de neurones convolutifs étaient souvent réalisées comme autres que ce qui a été attendu. Par exemple, parmi les voyelles orales /a/ identifiées comme nasales étaient fréquemment réalisées comme une autre voyelle orale /œ/ ou /ɔ/ et les voyelles nasales /ɑ̃/ détectées comme orales étaient souvent réalisées comme une autre voyelle nasale /ɛ̃/ ou /ɔ̃/. Cela explique l’une des limites de notre travail : le système cherche à reconnaître une voyelle /a/ plutôt qu’une voyelle orale ou une voyelle /ɑ̃/ plutôt qu’une voyelle nasale. Il en est de même pour les autres voyelles, la plupart des voyelles incorrectement identifiées sont produites comme autre que la voyelle en question. Nous supposons que les variations de production d’une voyelle n’est pas prises en compte dans l’entraînement. Pour améliorer la généralisation de l’algorithme de décision, il sera donc nécessaire de sélectionner les voyelles représentatives ou d’appliquer les méthodes sur un plus large ensemble des données, par exemple, toutes les productions de parole d’un locuteur.

6. Conclusion

L'utilisation des réseaux de neurones convolutifs a montré leur capacité d'extraction des paramètres pour la nasalité en réussissant à repérer les locuteurs ayant une voix identifiée comme plus nasale que d'autres et ceci a été validé au moyen des mesures acoustiques H1c et H1A1c. L'étude sur le test de perception a révélé que les locuteurs percevaient en majorité correctement en montrant leur capacité de reconnaissance des voyelles entendues. Un court essai a été effectué pour étudier le phénomène de la coarticulation nasale à travers les images découpées en tronçons, nous avons remarqué que les premières parties de voyelles sont fréquemment moins bien identifiées. Cependant, le sujet de la détection de la nasalité ne doit pas être conclu aussi rapidement tant que des mesures physiologiques n'ont pas été effectuées.

En procédant aux différents tests pour trouver dans quelle condition le système tend à produire des erreurs face aux voyelles, nous avons essayé d'établir la liste des éléments provoquant une difficulté au cours de la classification des réseaux de neurones convolutifs. Pour les contextes phonémiques, deux ("dorsal_pause", "dorsal_dorsal") ont été identifiés pour la voyelle /a/ et un ("coronal_dorsal") pour la voyelle /ɑ̃/. En ce qui concerne le sexe, les erreurs ont été plus fréquemment observées globalement chez les hommes, pour les voyelles orales chez les hommes et pour les voyelles nasales chez les femmes.

Nos différents algorithmes ont atteint pour une classification des voyelles orales et nasales jusqu'à 97% de bonne classification, et pour une généralisation des modèles sur les données non entraînées au préalable à un bon taux de classification (jusqu'à 94% d'exactitude) à tel point que certains modèles entraînés avec 40 locuteurs et testés sur 5 locuteurs ont parfois abouti à des meilleurs scores que les modèles entraînés et testés sur 45 locuteurs.

L'étude de la détection de nasalité présentée dans ce mémoire n'est qu'une simple approche ; les méthodes fiables de détection de la nasalité sur l'ensemble des productions de parole avec une validation au moyen des mesures physiologiques peuvent faire l'objet d'une thèse et être envisagées dans le domaine de la reconnaissance du locuteur ou dans le cadre du traitement automatique de la parole pathologique.

Annexes

Le lien GitHub permet d'accéder à l'ensemble de scripts et de fichiers contenant les résultats obtenus utilisé dans les expériences et les analyses du mémoire :

https://github.com/LilaKIM/memoireM2

```
if compteur_a < minimum
    if (txt$ = "a")
        for i from 1 to length
            phoneme$ = mid$ (var$, i)
            if txtAvant$ = phoneme$
                txtAvantFinale$ = dict$["'phoneme$'"]
                check1_a = 1
            endif
            if txtApres$ = phoneme$
                txtApresFinale$ = dict$["'phoneme$'"]
                check2_a = 1
            endif
        endfor
        if check1_a = 1 and check2_a = 1
            # pause 'txtAvant$', 'txtAvantFinale$', "A", 'txtApres$', 'txtApresFinale$', 'iInterval'
            tDeb = Get start point: iChamp, iInterval
            tFin = Get end point: iChamp, iInterval
            duree = tFin - tDeb
            if (duree >= duree_a_mean) and (duree <= dureeMax)
                compteur_a += 1
                selectObject: son
                Extract part: tDeb, tFin, "rectangular", 1, "no"
                if compteur_a <= train
                    folder$ = generatedTrainImagesNonNasalFolder$
                    folderSounds$ = generatedTrainSoundsNonNasalFolder$
                else
                    folder$ = generatedTestImagesNonNasalFolder$
                    folderSounds$ = generatedTestSoundsNonNasalFolder$
                endif
                Save as WAV file: folderSounds$ + "/" + nomFichierBase$ + "_" + txt$ + "_" + string$(iInterval) + "_" + txtAvantFinale$ + "_" + txtApre

                selectObject: spectrogramme
                zero_padding_factor = 5 / (dureeMax/duree)
                Select outer viewport: 0, zero_padding_factor, 0, 3
                Paint: tDeb, tFin, 0, 0, 100, "yes", 50, 6, 0, "no"
                Select outer viewport: 0, 5, 0, 3
                Save as 300-dpi PNG file: folder$ + "/" + nomFichierBase$ + "_" + txt$ + "_" + string$(iInterval) + "_" + txtAvantFinale$ + "_" + txtAp
                Erase all
            endif
        endif
    endif
endif
```

Figure 38: Exemple du zéro-padding sur une partie du script (l'ensemble est disponible sur le GitHub)

Bibliographie

Abercrombie, D. (1967). *Elements of general phonetics.* Edinburgh U.P.

Amelot, A. (2004). *Etude aérodynamique, fibroscopique, acoustique et perceptive des voyelles nasales du français*.

Amelot, A., Crevier-Buchman, L., Maeda, S., Amelot, A., & Shinji, M. (2003). Observations of velopharyngeal closure mechanism in horizontal and lateral direction from fiberscopic data Observations of the Velopharyngeal Closure Mechanism in Horizontal and Lateral Directions from Fiberscopic Data. In *Phonetic Sciences*. https://halshs.archives-ouvertes.fr/halshs-00138570

Amino, K., & Arai, T. (2009a). Effects of linguistic contents on perceptual speaker identification: Comparison of familiar and unknown speaker identifications. *Acoustical Science and Technology*, *30*(2), 89–99. https://doi.org/10.1250/ast.30.89

Amino, K., & Arai, T. (2009b). Speaker-dependent characteristics of the nasals. *Forensic Science International*, *185*(1–3), 21–28.

Amino, K., Makinae, H., & Kitamura, T. (2014). *Nasality in Speech and Its Contribution to Speaker Individuality*.

Amino, K., & Osanai, T. (2013). Speaker Identication Using Japanese Monosyllables and Contributions of Nasal Consonants and Vowels to Identication Accuracy. In 法科学技術, (Vol. 18, Issue 1).

Amino, K., Sugawara, T., & Arai, T. (2006). Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties. *Acoustical Science and Technology*, *27*(4), 233–235. https://doi.org/10.1250/ast.27.233

Boersma, P., & Weenink, D. (n.d.). *Praat: Doing phonetics by computer (version 6.1)*. http://www.praat.org/

Carignan, C. (2021). A practical method of estimating the time-varying degree of vowel nasalization from acoustic features. *The Journal of the Acoustical Society of America*, *149*(2), 911–922. https://doi.org/10.1121/10.0002925

Chanclu, A., Amor, I. Ben, Gendrot, C., Ferragne, E., & Bonastre, J. (n.d.). *Automatic classification of phonation types in spontaneous speech : towards a new workflow for the characterization of speakers ' voice quality*.

Dang, J., Honda, K., & Suzuki, H. (1994). Morphological and acoustical analysis of the nasal and the paranasal cavities. *The Journal of the Acoustical Society of America*, *96*(4), 2088–2100.

Dang, J., Wei, J., Honda, K., & Nakai, T. (2016). A study on transvelar coupling for non-nasalized sounds. *The Journal of the Acoustical Society of America*, *139*(1), 441–454. https://doi.org/10.1121/1.4939964

Denes, P., & Pinson, E. (1993). *The Speech Chain*.

Ferragne, E., Gendrot, C., & Pellegrini, T. (n.d.). *TOWARDS PHONETIC INTERPRETABILITY IN DEEP LEARNING APPLIED TO VOICE COMPARISON*. http://http//www.afcp-

Gelly, G. (2017). *Reseaux de neurones recurrents pour le traitement automatique de la parole*. https://tel.archives-ouvertes.fr/tel-01615475

Havel, M., Kornes, T., Weitzberg, E., Jon, O., & Sundberg, L. & J. (2016). *Eliminating paranasal sinus resonance and itseffects on acoustic properties of the nasal tract*.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). *Deep Neural Networks for Acoustic Modeling in Speech Recognition*.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). *Improving neural networks by preventing co-adaptation of feature detectors*.

House, S., Stevens, K. N., & Fant, ; G. (1956). Acoustic Theory of Speech Production. In *J. Speech Hear. Disord* (Vol. 21). http://asadl.org/terms

Kahn, J. (2011). *Parole de locuteur : performance et confiance enidentification biométrique vocale*. https://tel.archives-ouvertes.fr/tel-00995071v2

Laaridh, I. (2017). *Évaluation de la parole dysarthrique : Apport du traitement automatique de la parole face à l'expertisehumaine*.

Ladefoged, P., & Disner, S. F. (2012). *Vowels and consonants*.

LAVER, J. (2009). The Description of Voice Quality in General Phonetic. *Cambridge: CUP*.

Lozano-Diez, A., Plchot, O., Matějka, P., & Gonzalez-Rodriguez, J. (2018). *DNN BASED*

EMBEDDINGS FOR LANGUAGE RECOGNITION.

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). *Rectifier Nonlinearities Improve Neural Network Acoustic Models*.

Maddieson, I., & Abramson, A. S. (1987). Patterns of Sounds by Ian Maddieson . *The Journal of the Acoustical Society of America*, *82*(2), 720–721. https://doi.org/10.1121/1.395386

Maeda, S. (1982). ACOUSTIC CUES OF VOWEL NASALIZATION: A SIMULATION STUDY. *Recherches/Acoustique*, *7*. https://doi.org/10.1121/1.2019690

Nolan, F. (2007). *VOICE QUALITY AND FORENSIC SPEAKER IDENTIFICATION* (Issue 2).

O'Shea, K., & Nash, R. (2015). *An Introduction to Convolutional Neural Networks*. http://arxiv.org/abs/1511.08458

Pruthi, T., & Espy-Wilson, C. Y. (2007). *Acoustic Parameters for the Automatic Detection of Vowel Nasalization*.

Rose, P. (2000). *Forensic Speaker Identification*.

Serrurier, A. (2006). *Modélisation tridimensionnelle des organes de la parole à partir d'images IRM pour la production de nasales-Caractérisation articulatori-acoustique des mouvements du voile du palais*. https://tel.archives-ouvertes.fr/tel-00156977

Serrurier, A., & Badin, P. (2005). Towards a 3D articulatory model of velum based on MRI and CT images. In *A. Serrurier & P. Badin ZAS Papers in Linguistics* (Vol. 40).

Serrurier, A., & Badin, P. (2008). A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data. *The Journal of the Acoustical Society of America*, *123*(4), 2335–2355. https://doi.org/10.1121/1.2875111

Shosted, R. (n.d.). *A descriptive approach to the measurement of nasalization*.

Srivastava, N. (2013). *Improving Neural Networks with Dropout*.

Stevens, K. N. (1998). *Acoustic Phonetics*. The MIT Press.

Styler, W. (2017). On the acoustical features of vowel nasality in English and French. *The Journal of the Acoustical Society of America*, *142*(4), 2469–2482. https://doi.org/10.1121/1.5008854

Styler, W., & Scarborough, R. (n.d.). *Surveying the nasal peak: A1 and P0 in nasal and nasalized vowels*.

Suzuki, H., Nakai, T., Dang, J., & Lu, C. (1990). *SPEECH PRODUCTION MODEL INVOLVING SUBGLOTTAL STRUCTUREE AND ORAL-NASAL COUPLING THROUGH CLOSED VELUM*.

Torreira, F., Adda-Decker, M., & Ernestus, M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, *52*(3), 201. https://doi.org/10.1016/j.specom.2009.10.004ï

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016). *ADIEU FEATURES? END-TO-END SPEECH EMOTION RECOGNITION USING A DEEP CONVOLUTIONAL RECURRENT NETWORK*.

Vaissière, J. (1995). *Nasalité et phonétique*. https://halshs.archives-ouvertes.fr/halshs-00185541

Vaissière, J. (2011). *La phonétique*. Presses Universitaires de France. https://doi.org/10.3917/puf.vaiss.2011.01

Wieser, I., Barros, P., Heinrich, S., & Wermter, S. (2020). Understanding auditory representations of emotional expressions with neural networks. *Neural Computing and Applications*, *32*(4), 1007–1022. https://doi.org/10.1007/s00521-018-3869-3

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. In *Insights into Imaging* (Vol. 9, Issue 4, pp. 611–629). Springer Verlag. https://doi.org/10.1007/s13244-018-0639-9

Yuan, J., Lin, H., & Liu, Y. (1974). *NASAL COARTICULATION IN L1 AND L2 ENGLISH SPEECH: A LARGE-SCALE STUDY*.