Détection de la nasalité en parole à partir de wav2vec 2.0

Lila Kim* — **Cédric Gendrot***

* *Laboratoire de Phonétique et Phonologie (CNRS U. Sorbonne Nouvelle)*

RÉSUMÉ. *La nasalité s'observe à l'oral sur les consonnes et les voyelles (par exemple, « balle »* vs *« malle » ; « bas »* vs *« banc »). Elle peut s'étudier dans une optique linguistique (e.g. coarticulation) mais aussi pour la caractérisation du locuteur et la détection de pathologies de la parole. Du fait de la difficulté à analyser la nasalité par des mesures acoustiques traditionnelles, nous proposons une mesure basée sur des techniques de* Deep Learning*, que nous évaluons en comparant avec des mesures aérodynamiques prises directement sur le locuteur. Les représentations vectorielles du signal sonore sont extraites à l'aide de deux encodages différents du modèle wav2vec 2.0, XLSR et Lebenchmark, en faisant varier la taille de la séquence extraite ainsi que l'utilisation finale de ces représentations vectorielles. Les résultats obtenus montrent des classifications allant jusqu'à 99 %. L'utilisation de séquences courtes montre une meilleure détection de la nasalité phonétique avec ses variations dues au contexte ou au locuteur ; les séquences longues sont plus performantes pour la détection de la nasalité phonémique.*

MOTS-CLÉS : *parole, modèles neuronaux, nasalité, physiologie.*

TITLE. *Detecting Nasality in Speech Using Neural Models*

ABSTRACT. *Nasality can be observed in languages on consonants (e.g. "balle"* vs *"malle") and on vowels ("bas"* vs *"banc"). It can be studied from a linguistic perspective, but also for speaker characterization or speech pathologies. Given the difficulty of analyzing nasality with traditional acoustic measurements, we propose a measurement based on* Deep Learning *techniques, which we compare with aerodynamic data directly measured from the speaker. Vector representations of the sound signal are extracted using two different encodings of the wav2vec 2.0 model, varying the size of the extraction as well as the final use of these vector representations. The results obtained show classifications of up to 99%. The use of short sequences shows a better detection of phonetic nasality with its variations due to context or speaker; long sequences perform better for the detection of phoneme nasality.*

KEYWORDS: *Speech, Neural language models, Nasality. Physiology.*

1. Introduction

Cette étude vise à caractériser la nasalité dans les productions de parole de locuteurs français à partir de modèles neuronaux utilisés pour la reconnaissance automatique de la parole (e.g. wav2vec 2.0). Les modèles de neurones autosupervisés permettent de fournir une représentation de l'oral essentiellement dans le but d'effectuer des tâches de reconnaissance de la parole. Nous souhaitons montrer que certaines couches de ces modèles neuronaux peuvent également être utilisées pour la détection de traits phonologiques dans la langue ainsi que pour la caractérisation fine de la voix du locuteur. Nous analysons pour ce faire la nasalité présente dans la parole à la fois dans ses traits phonologiques et dans ses caractéristiques phonétiques. Le schéma méthodologique est décrit dans la figure 1.

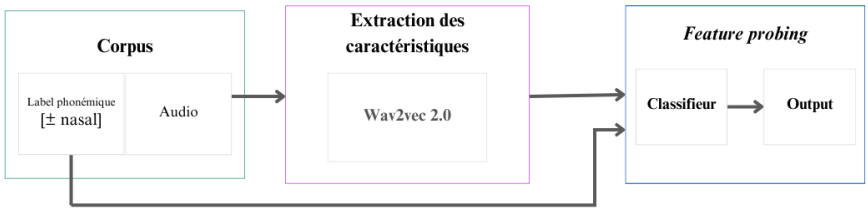


FIGURE 1. *Aperçu de la principale méthodologie expérimentale*

1.1. *État de l'art*

1.1.1. *Nasalité*

Le trait phonologique de nasalité est répandu dans les langues du monde, avec environ 97 % d'entre elles comportant au moins une consonne nasale et 22 % comportant au moins une voyelle nasale (Maddieson et Abramson, 1987 ; Stefanuto et Vallée, 1999). Pour les langues où la nasalité est une caractéristique phonologique, comme en français, en portugais (Wetzels, 1997) ou en sundanae (Robins, 1953), elle assure une identification lexicale (par exemple balle *vs* malle ; bas *vs* banc).

Dans une langue où la nasalité phonologique ne porte pas sur les voyelles, elle peut néanmoins être présente sur celles-ci sous forme de phénomène de coarticulation. Par exemple dans le mot « ban » en anglais, la voyelle sera nasalisée tout ou partiellement par assimilation régressive, et les locuteurs de l'anglais peuvent utiliser cette nasalisation pour prédire la consonne suivante (i.e. coda) et, surtout, le mot lexical (Malécot, 1960 ; Fromkin *et al.*, 1998 ; Cohn, 1993). Zellou (2022) a également montré que cette coarticulation nasale est fortement dépendante du locuteur et met donc en évidence des stratégies individuelles et sociolectales (Zellou, 2022).

Physiologiquement, le trait [$\pm nasal$] est caractérisé par l'abaissement du voile du palais, résultant du couplage de deux cavités, nasale et orale. Plus précisément, l'ouverture du port vélopharyngé permet de distinguer les différents types de sons nasals : consonnes nasales, voyelles nasales, ou une coarticulation nasale (Lagefoged et Maddieson, 1996). L'abaissement du velum entraîne des effets acoustiques sur les sons nasals. Ces effets sont notamment l'introduction de résonances nasales qui induisent des réductions de l'énergie dans le spectre acoustique par des pôles nasals, des changements dans les structures globales des formants et des changements dans l'enveloppe spectrale des voyelles (Delattre, 1954 ; Fant, 1971 ; House et Stevens, 1956 ; Stevens, 2000 ; Maeda, 1982b ; Maeda, 1982a ; Carignan, 2017). En outre, l'abaissement du voile du palais peut modifier la forme de la cavité buccale, ce qui affecte encore les formants (Feng, 1986 ; Fily, 2018).

Ces changements acoustiques et articulatoires compliquent les analyses phonétiques manuelles et automatiques, car elles sont très sensibles à des facteurs tels que l'accentuation, l'articulation et les variations linguistiques. Les méthodes phonétiques attestées dans la littérature (Chen, 1995 ; Chen, 1997 ; Styler, 2017) pour mesurer la nasalité sont notamment :

— *la mesure A1-p0*, pour laquelle une diminution de l'amplitude de A1 est attendue avec une nasale ;

— *la largeur de bande du F1*, une nasalité fait élargir le premier formant (le chevauchement des zéros influence et surface de la cavité nasale augmentée) ;

— *la pente spectrale*, puisqu'une nasalité plus importante accentue la pente nasale.

Ces mesures sont difficiles à obtenir automatiquement et elles sont sujettes à de nombreuses erreurs (Styler, 2017). L'auteur recommande par ailleurs de comparer les nasales aux non nasales strictement pour un même timbre vocalique (/a/ *vs* /ɑ̃/, /e/ *vs* /ɛ̃/, etc.), pour un même locuteur, et dans des contextes comparables. Il est donc nécessaire d'avoir recours à une mesure plus fiable et plus souple. Carignan (2021) a proposé une méthode basée sur des MFCC qui a obtenu des corrélations fluctuant entre 0.85 et 0.92 avec la mesure du débit d'air nasal proportionnel, avec une légère variation observée entre différents locuteurs. Mais celle-ci s'appliquait sur des voyelles tenues. Nous proposons ici une méthode qui s'applique sur de la parole naturelle. Nous analysons également simultanément consonnes et voyelles, sans nous limiter aux voyelles comme cela se fait habituellement.

La nasalité peut également être présente en tant que facteur non phonologique. Dans ce cas, elle est toujours dépendante physiologiquement de l'abaissement du voile du palais, mais elle varie en fonction de contextes positionnels, situationnels, d'habitudes ou de caractéristiques du locuteur. Les fins d'énoncés correspondent généralement aux moments de relâchement du locuteur avec un abaissement du voile du palais (position de respiration) (Berti, 1976). Ainsi les phones suités en fin d'énoncés seront fréquemment réalisés comme nasals. La sociolinguistique nous révèle des cas de nasalité en fonction de contextes situationnels : les femmes Cayuvava nasalisent leur voix comme une forme de politesse lorsqu'elles s'adressent aux hommes

(Laver, 2009). Cette manière de parler est couramment utilisée lorsque l'une des personnes a un statut inférieur à l'autre (Laver, 2009). Il existe également un style de discours en coréen appelé « Aegyo », où les locuteurs nasalisent la fin de la phrase pour paraître charmants et mignons (Puzar et Hong, 2018 ; Crosby, n.d.).

La nasalité est fréquemment considérée comme une composante de la qualité de la voix. Or celle-ci est considérée comme un facteur important pour caractériser un locuteur (Gold et French, 2019). Elle peut être un aspect constant de la voix d'un locuteur en raison de particularités physiologiques (principalement laryngées et supralaryngées), ou de facteurs idiolectaux ou sociolinguistiques. Par exemple, les fins de phrases des parisiens sont fréquemment assez nasales, surtout par l'insertion de « hein » (ou « han ») en fin de phrase. Cependant, la qualité de la voix peut également varier chez un même locuteur, notamment en fonction du style de discours ou de l'état émotionnel (Nolan, 2014). Si la qualité de la voix ne permet pas d'identifier un locuteur, elle permet néanmoins de fournir une caractéristique fiable en plus d'autres caractéristiques telles que la fréquence fondamentale ou l'articulation des voyelles et des consonnes. Elle apporte en ce sens l'explicabilité que les systèmes globaux d'identification ou de vérification du locuteur ne peuvent apporter.

Enfin, la nasalité, en tant que caractéristique vocale, peut également être étudiée dans un contexte pathologique. Dans le cas de la fente de la voûte palatine, le port vélopharyngé reste ouvert plus longtemps que ce qui est normalement prévu, ce qui entraîne un phénomène d'hypernasalisation (Chen, 1997).

Dans le domaine de la reconnaissance automatique des locuteurs, il a été fréquemment observé (Kahn, 2011) que les nasales sont pertinentes pour caractériser les locuteurs. Cela peut s'expliquer par le fait que les caractéristiques de la cavité nasale varient d'un individu à l'autre et restent relativement stables – car la cavité est peu malléable – pendant la production de la parole (Dang *et al.*, 1994 ; Serrurier, 2006), ce qui en fait une cavité de résonance distincte et fiable, propre à chaque locuteur. Pour finir, la nasalité est fréquemment influencée par plusieurs facteurs : il a été constaté que des différences significatives existent dans la pression du débit d'air oral et nasal entre les individus de sexe masculin et féminin (Clarke, 1975). La taille de la cavité nasale a un impact sur la fermeture du port vélopharyngé (Amelot, 2004) et cette force influence le degré de nasalité (Esling *et al.*, 2019). L'opposition phonologique nasale *vs* non nasale est donc physiologiquement réalisée de façon plus graduelle qu'elle n'apparaît au premier abord avec une forte variation entre interlocuteurs qui pourrait fournir des informations sur le contexte ou sur le locuteur.

1.1.2. *Approches neuronales*

Les systèmes de reconnaissance automatique du locuteur peuvent être subdivisés en différentes tâches, notamment la vérification du locuteur et l'identification du locuteur (O'Shaughnessy, 1987 ; Campbell, 1997). Dans le cadre de ces tâches, on cherche à établir l'identité du locuteur à partir de la production de parole. L'extraction des caractéristiques est devenue importante dans le domaine car elle ouvre la voie à une amélioration constante de la qualité des systèmes de reconnaissance du locuteur

(Meuwly, 2001). Cette amélioration peut être cruciale pour garantir la robustesse du modèle au bruit ambiant ou à la réverbération. Pour ce faire, plusieurs approches ont été explorées : systèmes experts, approches statistiques et approches neuronales.

Depuis 2010 et avec l'avènement des approches neuronales, la phonétisation, l'utilisation de connaissances psycho-acoustiques et le traitement du signal deviennent des étapes entièrement neuronales (Graves *et al.*, 2006 ; Hannun *et al.*, 2014 ; Amodei *et al.*, 2016). Les réseaux de neurones tels que les CNN et plus récemment ceux intégrant un mécanisme d'attention, connu sous le nom de *Transformer* (Vaswani *et al.*, 2017), ont connu une évolution significative et ont progressivement remplacé les modèles de mélanges gaussiens dans le domaine de la reconnaissance automatique de la parole (Hinton *et al.*, 2012). Les limites de l'approche du *Deep Learning* sont généralement liées à la nécessité d'une très grande quantité de données annotées manuellement. Pour dépasser le manque de données annotées, d'autres types d'apprentissage « légèrement supervisés » ou « autosupervisés » et ont été entrepris (Lee *et al.*, 2021 ; Radford *et al.*, 2023). Ces modèles sont préalablement entraînés sur des milliers d'heures d'audio non annotées, puis réentraînés sur un ensemble de données annotées de plus petite taille pour effectuer une tâche spécifique (Radford *et al.*, 2018 ; Devlin *et al.*, 2018 ; Baevski *et al.*, 2020). Comparativement à l'entraînement non supervisé, l'entraînement autosupervisé est bénéfique pour les tâches liées à la parole ou pour les langues qui manquent de ressources linguistiques car il permet d'accomplir des tâches avec des performances améliorées malgré un nombre de données étiquetées limité (Guillaume *et al.*, 2023). Les vecteurs obtenus par ces modèles contiennent une quantité importante d'informations sur la parole qu'il est nécessaire d'appréhender, tant pour l'utiliser dans d'autres tâches plus spécifiques comme c'est le cas de la présente étude, mais également pour des raisons d'éthique et de protection de la vie privée.

Parmi les exemples connus de modèles autosupervisés préentraînés, notre travail repose sur le modèle wav2vec 2.0, un modèle d'apprentissage automatique capable d'encoder les données audio brutes en représentations vectorielles (Baevski *et al.*, 2020). Le modèle se compose de trois éléments principaux : un encodeur, un réseau contextuel basé sur les *Transformers*, et un module de quantification. L'encodeur convolutionnel est chargé de traiter le signal audio brut, afin d'extraire des représentations de la parole. Ces représentations latentes sont ensuite discrétisées par le module de quantification. Les *Transformers* jouent un rôle essentiel en obtenant des vecteurs contextuels qui englobent un large éventail d'informations sur les caractéristiques acoustiques, couvrant à la fois le début, le milieu et la fin des phonèmes. Ils parviennent à réaliser cela en capturant des informations à l'échelle de l'ensemble de la séquence audio, tout en modélisant les interactions complexes entre les différentes représentations latentes (Baevski *et al.*, 2020).

Il existe plusieurs variantes du modèle comme EN préentraîné sur 53 000 heures de parole en anglais (Baevski *et al.*, 2020), XLSR (*Cross-Lingual Speech Representation*) préentraîné sur un ensemble de 53 langues (Conneau *et al.*, 2020), LeBenchmark préentraîné sur environ 3 000 heures de parole en français (Parcollet *et al.*, 2023). Le

modèle wav2vec 2.0 est souvent employé dans le cadre du *downstream task*, où l'on affine (*fine tune*) le modèle sur des données annotées de quelques minutes jusqu'à plusieurs centaines d'heures afin d'effectuer une tâche spécifique comme la reconnaissance automatique d'une nouvelle langue par exemple. Dans le cadre de cette étude, nous utilisons le modèle wav2vec 2.0 comme un extracteur de caractéristiques dans une méthode appelée *feature probing* (Triantafyllopoulos *et al.*, 2022 ; Shah *et al.*, 2021 ; Ma *et al.*, 2020) qui s'appuie sur le fait que l'information linguistique est présente dans les représentations intermédiaires issues du modèle wav2vec 2.0 (Triantafyllopoulos *et al.*, 2022). Notre objectif est d'utiliser certaines caractéristiques des couches intermédiaires de wav2vec 2.0 afin d'entraîner un modèle à accomplir une tâche spécifique sans *fine-tuning*[1] (Triantafyllopoulos *et al.*, 2022). Une hypothèse fréquemment mise en avant est que chaque couche du *Transformer* contient un type d'information différent, tel qu'une information linguistique, acoustique, phonétique ou articulatoire (Adi *et al.*, 2016 ; Conneau *et al.*, 2018 ; Ma *et al.*, 2020 ; Shah *et al.*, 2021 ; Triantafyllopoulos *et al.*, 2022 ; Yang *et al.*, 2023).

Notre étude a pour but d'identifier automatiquement la nasalité dans toutes les productions de parole, qu'il s'agisse de voyelles ou de consonnes, et ce travail peut être découpé en trois sous-objectifs :

— d'abord, nous cherchons à évaluer la capacité de l'encodeur wav2vec 2.0 à détecter la nasalité dans la parole ;

— nous confrontons cette classification à des mesures physiologiques prises directement à partir du locuteur ;

— enfin, nous visons à démontrer que notre approche est capable de mettre en évidence la variabilité entre les locuteurs et la nasalité propre à chaque locuteur.

Nous tentons donc de démontrer ici que des modèles initialement conçus pour la reconnaissance automatique de la parole peuvent être détournés de leur tâche première sans *fine-tuning*, tout en conservant une bonne interprétabilité des résultats.

2. Ressources

Les données d'entraînement proviennent de quatre corpus du français, ESTER, NCCFr, PTSVOX et BREF, sur lesquels les représentations vectorielles obtenues par wav2vec 2.0 sont entraînées avec un *multilayer perceptron*, et nous avons testé le modèle ainsi obtenu sur des données acoustiques pour lesquelles une mesure physiologique a été effectuée en guise de référence. Nous présentons ci-dessous les corpus et l'extraction des sons ayant servi à l'entraînement. Dans un deuxième temps, nous abordons les algorithmes utilisés pour obtenir les représentations vectorielles. Pour finir, les données de test sont détaillées.

1. Le *fine-tuning* est rendu impossible dans notre cas d'étude par la petite taille des fenêtres d'analyse que nous utilisons, i.e. la longueur du phonème

2.1. *Corpus et extraction des sons pour l'entraînement*

Les phonèmes ciblés sont les trois voyelles nasales /ɑ̃,ɛ̃,ɔ̃/ ainsi que leurs homologues orales /a,ɛ,o/ en français ainsi que des consonnes nasales et orales /m,n,b,d,v,l/. Notre choix pour ces phonèmes s'est principalement porté sur des sons voisés avec le trait nasal ou non nasal et des variations dans les lieux d'articulation. Ces phonèmes nasals et leur correspondant oral ont une articulation proche et se distinguent par la hauteur du velum.

L'extraction des réalisations de ces phonèmes a été effectuée à l'aide d'un script Praat utilisant une fenêtre rectangulaire pour isoler uniquement le phonème à ses frontières. Dans un deuxième temps, une fenêtre d'une seconde avant et après les frontières du phonème sont également extraites afin de mieux correspondre à l'entraînement du modèle wav2vec 2.0 qui s'appuie sur des séquences de quelques secondes. Les représentations vectorielles sont ainsi extraites *a posteriori* à l'endroit exact où se trouve notre cible d'intérêt. Notre hypothèse est que le contexte temporel aidera le modèle à prendre en compte les contrastes phonémiques et sera plus performant dans la détection de la nasalité phonologique. Afin d'évaluer la précision de nos modèles, nous nous sommes appuyés sur la nasalité phonologique de la langue comme référence. Cependant, il est important de noter que ces caractéristiques phonologiques ne garantissent pas que les sons sont réellement produits avec une nasalité, car cette réalisation est influencée par divers facteurs, mentionnés dans l'introduction.

Pour l'entraînement et la validation, nous avons extrait les différents types de phonèmes depuis quatre corpus différents, chacun représentant un type de parole distinct. Les corpus de données utilisés dans cette étude comprennent :

— NCCFr (*Nijmegen Corpus of Casual French*) : il s'agit d'un corpus contenant 36 heures de parole continue, principalement sous la forme de conversations amicales impliquant 46 locuteurs français (Torreira *et al.*, 2010) ;

— ESTER (Évaluation de systèmes de transcription enrichie d'émissions radiophoniques) : ce corpus a été créé pour évaluer des systèmes de transcription automatique pour le français. Il comprend des conversations radiophoniques en français, totalisant 100 heures de parole préparée et lue (Gravier *et al.*, 2004 ; Galliano *et al.*, 2006). Seule une partie de 30 heures a été retenue pour cet entraînement ;

— PTSVOX : ce corpus a été développé pour mesurer les variations intra- et interlocuteurs dans le contexte de la comparaison de voix à des fins judiciaires. Il comprend des enregistrements de parole d'environ 90 heures, impliquant 369 locuteurs de français (Chanclu *et al.*, 2020). Nous n'avons retenu qu'une petite partie de ce corpus avec des alignements vérifiés, pour les productions de seulement 24 locuteurs ;

— BREF : il s'agit d'un corpus conçu pour le développement et l'évaluation des systèmes de reconnaissance de la parole, ainsi que pour étudier les variations phonologiques. Les données proviennent d'articles du journal Le Monde et ont été lues par 120 locuteurs du français de Paris, totalisant 100 heures de parole continue (Lamel *et al.*, 1991). Là encore, tous les alignements en phonèmes ne nous ayant pas été communiqués, seule la moitié du corpus BREF a été utilisée pour les entraînements.

Cette diversité de sources vise à renforcer la robustesse du modèle face aux variations de données et au bruit. Au total, 75 000 voyelles et consonnes nasales et 75 000 voyelles et consonnes orales ont été extraites de manière aléatoire sans aucune sélection de contexte phonétique ou prosodique. Sur la totalité des données extraites, 80 % ont été dédiées pour l'apprentissage des modèles tandis que les 20 % restants ont été utilisés pour la validation. Nous présentons la liste des phonèmes extraits utilisés pour l'entraînement des modèles dans les tableaux 1 et 2, accompagnée du nombre d'occurrences de chaque phonème (voir section 2.3).

2.2. *Obtention des représentations vectorielles et initialisation d'un classifieur*

Dans notre étude, nous visons à évaluer la capacité des réseaux de neurones à détecter la nasalité sur tous les phonèmes confondus, en nous appuyant sur un modèle de parole autosupervisé wav2vec 2.0 (Baevski *et al.*, 2020). Nous avons opté pour celui-ci après des premières tentatives infructueuses à partir de MFCC, qui ne fournissaient pas de performances satisfaisantes. En effet, le taux d'exactitude global était de 79,55 % avec les mêmes structures et paramètres d'apprentissage présentés ci-après, et des erreurs très fréquentes ont été observées en particulier sur les consonnes, avec moins de 25 % de précision pour des phones tels que [b]. D'autres essais impliquant un apprentissage de la nasalité exclusivement à partir de réseaux de neurones convolutifs ont permis d'aboutir à des résultats atteignant 88 % d'exactitude, mais ceux-ci semblaient pouvoir être améliorés en utilisant la puissance des modèles préentraînés plus récents. Parmi les différentes variantes du modèle disponibles, notre choix s'est porté sur le modèle wav2vec 2.0-large-xlsr-53 et le modèle wav2vec 2.0-FR-3K-large-LeBenchmark. Le modèle XLSR, qui a été préentraîné sur 53 langues, génère un vecteur de représentation de la parole qui surmonte les frontières linguistiques (Conneau *et al.*, 2020). En ce qui concerne le modèle LeBenchmark, il a été préentraîné sur 2 900 heures de parole en français, avec une conception axée sur l'optimisation de ses performances dans des tâches en français (Parcollet *et al.*, 2023).

Notre objectif consiste à comparer les performances de ces modèles dans le cadre de la détection de la nasalité, et pour ce faire, nous avons opté pour l'approche de *feature probing* expliquée en section 1, figure 1. Il a été constaté que l'information acoustique prédomine dans les premières couches de *Transformer* du modèle wav2vec 2.0 (Pasad *et al.*, 2021 ; Pasad *et al.*, 2022). Les représentations de la parole extraites ont alors été utilisées comme entrée d'un classifieur dans le but d'obtenir un score de probabilité pour chaque stimulus.

L'approche d'extraction des représentations vectorielles est basée sur la méthodologie présentée par Guillaume *et al.* (2023), qui se concentre sur une analyse linguistique d'une langue à partir de la parole dans un extrait audio de 5 secondes. Elle consiste à extraire des représentations vectorielles à partir des séquences audio, puis à utiliser la stratégie de *max pooling* pour agréger ces représentations en un seul vecteur représentatif de l'ensemble du signal. Les valeurs de ces représentations sont ensuite sauvegardées dans un fichier *pickle*, un format binaire, pour permettre une utilisation

future et une importation rapide des données. Ces représentations ont ensuite été utilisées pour alimenter l'un des deux classifieurs, soit un *multilayer perceptron* (MLP), soit la régression logistique. L'intérêt particulier que revêt le MLP ici est de pouvoir récupérer les *embeddings* et ainsi spécialiser les *features* issus de wav2vec 2.0 dans une tâche de détection de nasalité interprétable.

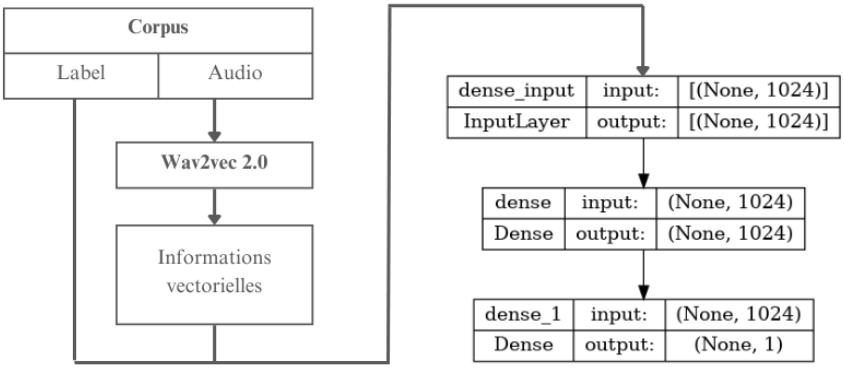


FIGURE 2. *Démarche de l'architecture d'apprentissage complète du modèle MLP*

Le MLP est composé de deux couches denses : la première couche contient 1024 neurones pour traiter et comprendre les représentations vectorielles issues de chaque modèle W2V2 en utilisant la fonction d'activation ReLu. La seconde couche est dédiée à déterminer si le son est prononcé avec nasalité (=1) ou sans nasalité (=0). Dans cette couche, la fonction d'activation utilisée est *sigmoïd* et la fonction de perte appliquée est *binary cross entropy*. Nous avons implémenté notre classifieur à l'aide de la bibliothèque de réseaux neuronaux Keras en Python (Chollet *et al.*, 2015). L'apprentissage a été réalisé en utilisant l'optimiseur Adam, le taux d'apprentissage a été ajusté au cours des entraînements pour arriver à la valeur de 0,000125. La taille du lot (*batch size*) a été fixée à 256 et nous avons effectué 150 époques d'entraînement avec une stratégie d'arrêt précoce (*early stopping*) pour éviter le surapprentissage. L'architecture complète du MLP est illustrée dans la figure 2.

En ce qui concerne la régression logistique, nous avons utilisé la bibliothèque d'apprentissage automatique en Python *scikit-learn* avec les paramètres par défaut (Pedregosa *et al.*, 2011).

2.3. *Données acoustiques pour l'évaluation du modèle*

Les données de test ont été recueillies auprès de six locuteurs masculins, tous natifs du français, âgés en moyenne de 36 ans. Les enregistrements ont été effectués dans une chambre sourde pour éviter tout bruit indésirable. Les stimuli ont été générés dans le

cadre de structures VCV ou VNV, où C représente [p,b,t,d,v,s,z], N représente [m,n], et V représente [i,a,y,u,o,e,ɑ̃,ɛ̃,ɔ̃]. Ces séquences de stimuli ont été intégrées dans une phrase de cadre : « Non tu n'as pas dit XXX quatre fois, mais tu as dit YYY et ZZZ quatre fois ». Il est important de noter que les mots XXX, YYY et ZZZ correspondent à des structures VCV ou VNV et ne sont pas porteurs de sens (i.e. logatomes).

La collecte des données aérodynamiques et acoustiques a été effectuée simultanément à l'aide d'un masque pneumotachographique (appelé *Aero mask*) conçu au Laboratoire de Phonétique et Phonologie (LPP) (Elmerich *et al.*, 2020 ; Elmerich *et al.*, 2023a ; Elmerich *et al.*, 2023b ; Kim *et al.*, 2023). Ce masque permet l'enregistrement distinct du débit d'air à travers la bouche et le nez sans introduire de distorsions acoustiques. Par conséquent, un total de 269 sons de chaque classe ont été extraits et découpés à leurs frontières respectives. Ici encore, dans un deuxième temps, des séquences plus longues incluant une seconde autour du phone ont été extraites afin d'être comparées aux séquences correspondant aux seuls phones. Les mesures aérodynamiques des sons ont été enregistrées dans un fichier au format *csv* pour comparer avec les résultats du réseau de neurones profonds. Les données utilisées pour l'entraînement, la validation et le test se résument dans les tableaux 1 et 2.

Catégorie	Son [+ nasal]						
Phonème	ɑ̃	ɛ̃	ɔ̃	m	n	ɲ	total
Entraînement	14 827	7 538	9 893	15 173	12 459	110	60 000
Validation	3 734	1 941	2 462	3 834	2 999	30	15 000
Test	66	66	66	36	35	0	269

TABLEAU 1. *Nombre d'occurrences des segments [+ nasal] utilisés*

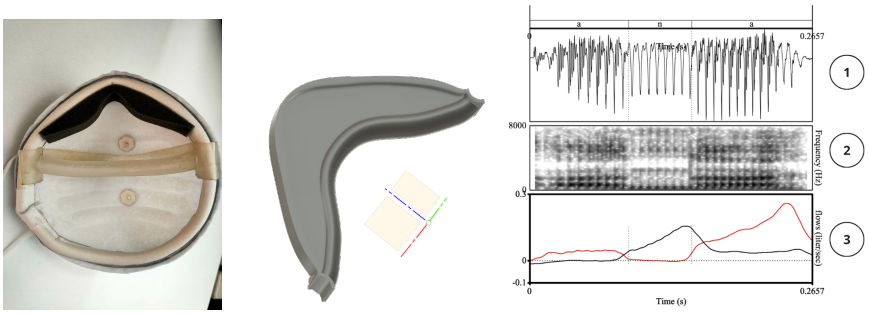
Catégorie	Son [- nasal]									
Phonème	a	e	ɛ	o	ɔ	b	d	l	v	total
Entraînement	15 670	7 308	5 392	2 486	1 319	2 824	9 649	11 335	4 017	60 000
Validation	3 854	1 793	1 276	598	316	679	2 524	2 923	1 037	15 000
Test	66	39	27	66	0	25	29	0	17	269

TABLEAU 2. *Nombre d'occurrences des segments [- nasal] utilisés*

2.4. *Mesures physiologiques pour évaluer la corrélation avec la probabilité de nasalité prédite*

L'abaissement du voile du palais ne suffit pas à lui seul à déterminer la présence ou l'absence de nasalité. Par exemple, bien que le voile du palais soit abaissé de manière plus significative pour les voyelles nasales, une élévation similaire du voile du palais peut être observée lors de la production de voyelles orales, de consonnes nasales, voire parfois de consonnes orales (Rossato *et al.*, 2003). Dans ce contexte, les mesures aérodynamiques sont intéressantes pour vérifier si les phones oraux sont produits avec un flux d'air nasal, ce qui indique l'ouverture vélopharyngée permettant à l'air de circuler dans la cavité nasale. Dans le cadre de cette étude, nous avons recours à 3 mesures

aérodynamiques en guise de référence et de comparaison : le débit d'air nasal (DAN), le débit d'air buccal (DAB) et le débit d'air nasal proportionnel.



(a) (b) (c)

FIGURE 3. *Enregistrement avec le masque (a) masque en papier fibre avec plaque et 2 adaptateurs reliés aux capteurs de pression ; (b) séparation flexible intégrée au masque pour séparer les flux d'air nasal et oral ; (c) exemple d'enregistrements acoustiques et de débit d'air de [ana]. De haut en bas, (1) signal audio capturé avec un microphone, (2) spectrogramme, (3) débit d'air nasal (DAN en noir) et débit d'air oral (DAB en rouge)*

Les débits d'air nasal et oral ont été mesurés à l'aide d'un masque en tissu développé au Laboratoire de Phonétique et Phonologie (LPP). Ce masque illustré dans la figure 3 permet de réaliser des enregistrements simultanés de données acoustiques sans distorsion et de données aérodynamiques, ce qui facilite l'exploitation de ces données acoustiques. Ainsi, il permet de faire le lien entre les aspects aérodynamiques, acoustiques, voire perceptuels de la parole. À l'intérieur du masque se trouvent deux parties distinctes : la partie buccale et la partie nasale, conçues pour mesurer les débits d'air nasal et oral de manière simultanée mais indépendante. Les capteurs de pression intégrés dans chaque section convertissent les valeurs de débit d'air en litres. Ainsi, une calibration distincte des capteurs de pression nasal et oral est réalisée pour chaque locuteur (Elmerich *et al.*, 2020 ; Elmerich *et al.*, 2023a).

Les débits sont mesurés sur la totalité de la phrase, permettant d'obtenir une courbe de débit d'air nasal et buccal synchronisée avec le signal acoustique. Ensuite, nous avons calculé les moyennes des débits d'air nasal et buccal (DAN et DAB) en litres par seconde sur la base de segments. Cette mesure a été réalisée sur l'ensemble de la production vocale, incluant voyelles et consonnes. Puisque les débits d'air nasal et buccal sont des valeurs absolues, nous avons eu recours également à une mesure de débit d'air nasal proportionnel, définie comme le rapport entre les débits d'air nasal et oral $(DAN/(DAN + DAB)) \times 100$ (en %), conformément à des travaux antérieurs (Delvaux, 2000).

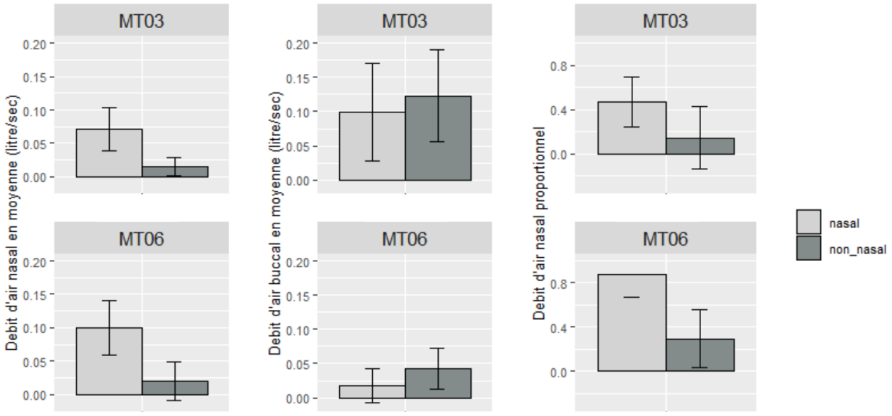


FIGURE 4. *(a) Débit d'air nasal en moyenne (litre/sec) ; (b) Débit d'air oral en moyenne (litre/sec) ; (c) Débit d'air nasal proportionnel (%) selon les locuteurs MT03 et MT06 et les catégories de sons*

La figure 4 illustre les distributions des débits d'air nasal, buccal et du débit d'air nasal proportionnel, moyennés par locuteur et par classe phonémique. Quatre tendances se dessinent :

— le débit d'air nasal est présent pour les phones de classe nasale, mais il peut également être présent pour les phones de classe orale ;

— le débit d'air buccal est présent dans les deux classes, à la fois pour les phones de classe orale et nasale ;

— le débit d'air nasal proportionnel (nasalance) est plus élevé pour la catégorie nasale que pour la catégorie orale ;

— dans toutes les mesures, une variation entre les locuteurs est présente.

Une analyse de variance a été réalisée pour déterminer si les valeurs de débit d'air nasal varient entre les locuteurs. Les résultats ont révélé un effet statistiquement significatif du débit d'air nasal sur les locuteurs ($p = 0.000152$). D'après les schémas présentés en figure 4, on observe clairement une variation entre les locuteurs. Par exemple, le débit d'air buccal pour le locuteur MT06 est le plus bas parmi tous les locuteurs. Les phones du locuteur MT06, prononcés avec un débit d'air nasal élevé et un minimum de débit d'air buccal, présentent ainsi le niveau du débit d'air nasal proportionnel le plus élevé. Sur le même principe (non illustré), les locuteurs MT05 et MT07 montrent un débit d'air nasal plus élevé pour les phones de catégorie nasale que les autres locuteurs, suggérant une voix plus nasale.

3. Expérience : détection de nasalité au moyen des réseaux de neurones profonds

Dans cette section, nous décrivons les expériences que nous avons menées pour développer un réseau de neurones profonds capable de détecter la nasalité. Nous avons extrait des caractéristiques de chaque couche intermédiaire du modèle wav2vec 2.0 (LeBenchmark et XLSR) pour alimenter un classifieur. L'analyse de ces caractéristiques aide à identifier les traits les plus appropriés et pertinents pour la nasalité. Cette expérience se déroule en plusieurs étapes. Tout d'abord, la comparaison des deux modèles, Lebenchmark et XLSR, est effectuée afin de déterminer lequel est optimal pour détecter la nasalité dans les données en français. Ensuite, nous procédons à la comparaison des classifieurs de type MLP et régression logistique dans le but d'établir si l'un des deux est nécessaire pour obtenir une meilleure performance. En effet, le MLP est reconnu pour sa capacité à mieux traiter les données non linéaires. En troisième lieu, la longueur d'extrait audio est examinée en comparant les séquences courtes (correspondant à un phone) et les séquences plus longues (une seconde avant et après le phone). Enfin, les corrélations entre la probabilité de nasalité et le débit d'air nasal sont mesurées afin de confirmer et de détailler nos résultats par locuteur.

3.1. Probing task *et évolution des représentations selon les couches*

3.1.1. *Comparaison des modèles Lebenchmark et XLSR*

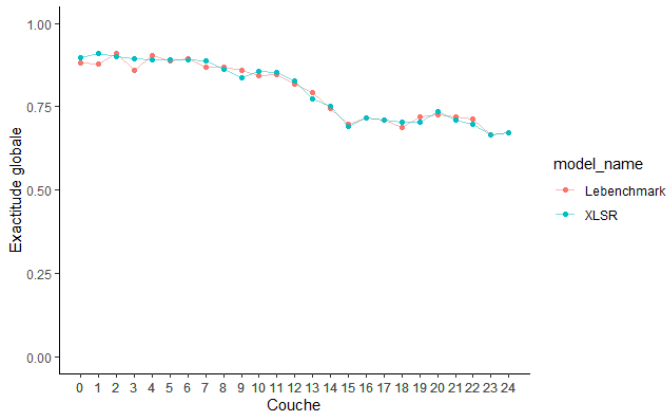


FIGURE 5. *Distribution de l'exactitude globale en fonction des couches du wav2vec 2.0 selon deux modèles Lebenchmark et XLSR*

Dans le but de déterminer la couche optimale à exploiter pour l'entraînement d'un classifieur en vue d'obtenir de meilleures performances, nous avons évalué comment les représentations évoluent à travers différentes couches, allant de la sortie de l'encodeur CNN jusqu'à la dernière couche de *Transformer*, dans le contexte de la détection

de la nasalité. La figure 5 illustre la distribution de l'exactitude globale en fonction de différentes couches de deux versions du modèle wav2vec 2.0 : LeBenchmark et XLSR. Cela souligne que l'information sur la nasalité est plus particulièrement présente dans la sortie de l'encodeur CNN et dans les premières couches du *Transformer* pour les deux modèles.

En termes de performance, le modèle Lebenchmark est équivalent à celui du modèle XLSR, avec des taux d'exactitude globale allant jusqu'à 90.52 %. Nous poursuivons toutefois la présentation de notre travail en utilisant uniquement le modèle Lebenchmark, puisque ce modèle est spécifiquement préentraîné sur le français, mais également pour des raisons de place. Nous avons donc utilisé ce modèle pour extraire des représentations vectorielles, sur lesquelles nous avons ensuite appliqué deux classifieurs : un *multilayer perceptron* et une régression logistique.

3.1.2. *Comparaison de deux classifieurs : MLP et régression logistique*

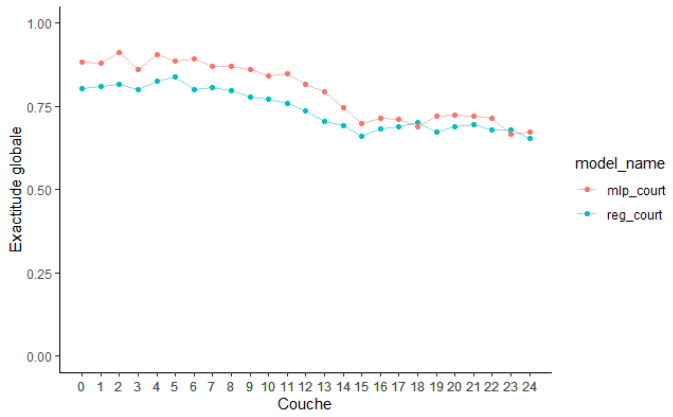


FIGURE 6. *Distribution de l'exactitude globale en fonction des couches du wav2vec 2.0 selon deux manières de* feature probing

Il existe différents types de classifieurs pour effectuer une tâche de *probing*, notamment le MLP (English *et al.*, 2022) et la régression logistique (Guillaume *et al.*, 2023), ou la régression linéaire (Lenglet *et al.*, 2023). Nous émettons l'hypothèse que le MLP s'adapte mieux à la nasalité puisque capable d'intégrer des dimensions non linéaires. L'atout de ce dernier est qu'il permet également de récupérer les *embeddings* afin de constituer de nouveaux vecteurs d'analyse.

La figure 6 présente les taux d'exactitude globale pour la caractéristique $\pm nasal$, en fonction des classifieurs utilisés. Cette analyse comparative de la performance des classifieurs a été réalisée en observant l'évolution dans différentes couches du modèle Lebenchmark. Les résultats révèlent que la performance du modèle de régression logistique est légèrement inférieure à celle du modèle MLP. Les caractéristiques obtenues des 13 premières couches se révèlent plus particulièrement bénéfiques

lorsqu'elles sont introduites dans un modèle de MLP, améliorant ainsi sa capacité à classifier la nasalité.

3.1.3. *Comparaison sur la longueur des extraits audio*

Deux approches ont été explorées pour extraire des représentations vectorielles à partir des données audio. Dans la première approche, inspirée de l'approche phonétique, les phones sont découpés à leurs frontières et des représentations vectorielles sont extraites sur ces séquences courtes. La seconde approche implique l'utilisation de séquences plus longues, avec l'ajout d'une seconde au début et à la fin d'un phone, suivie de la récupération des vecteurs centraux (correspondant au phone) dans un deuxième temps.

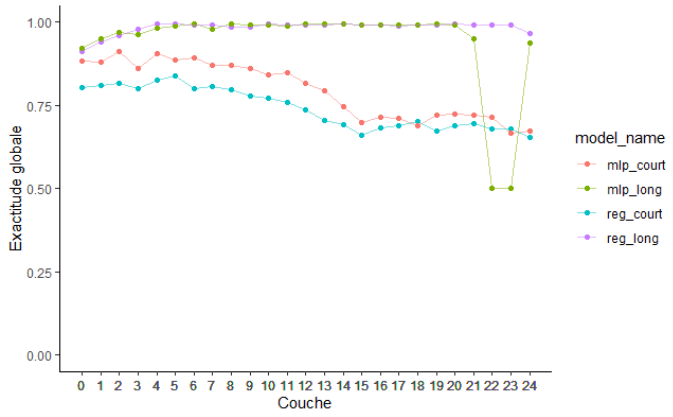


FIGURE 7. *Distribution de l'exactitude globale selon la longueur de l'extrait sonore et de la méthode de* feature probing

L'analyse de la performance des classifieurs en fonction de la durée de l'extrait audio est illustrée dans la figure 7. Cette comparaison entre les séquences longues et courtes montre une meilleure performance des modèles entraînés sur des séquences longues. Contrairement aux séquences courtes, qui se révèlent plus avantageuses dans les couches initiales, les séquences longues présentent une autre tendance : les informations relatives à la nasalité sont détectables au-delà de 95 % dans la plupart des couches, à l'exception du modèle MLP entraîné avec des caractéristiques des couches 22 et 23.

3.1.4. *Corrélations entre les différents modèles*

Dans cette étude, le coefficient de corrélation de Pearson est employé pour analyser la relation linéaire entre la probabilité de nasalité et le débit d'air nasal. Les corrélations sont évaluées avec le débit d'air nasal tel qu'obtenu par l'*aeromask*.

La figure 8 met en lumière les corrélations observées dans différentes couches du wav2vec 2.0. À l'exception de la sortie de l'encodeur CNN, les séquences longues ne

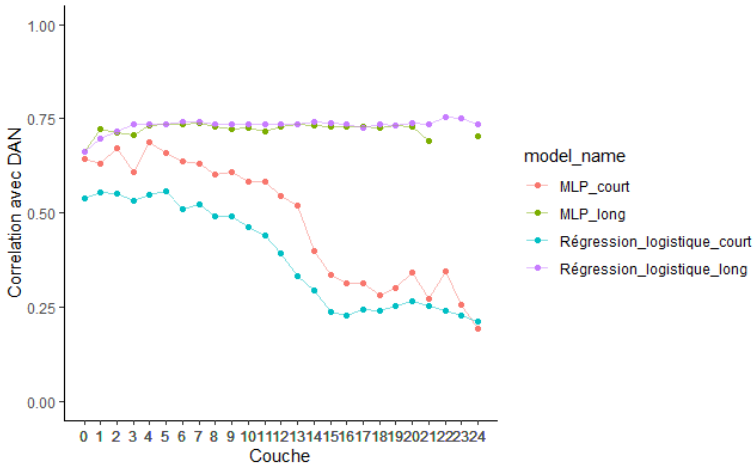


FIGURE 8. *Étude de la relation entre la probabilité de nasalité et le débit d'air nasal selon les couches du modèle wav2vec 2.0, en prenant en compte la longueur de l'extrait et la méthode de* feature probing

révèlent pas de tendance claire dans l'évolution au sein des couches du *Transformer*, mais elles présentent toutes une corrélation significative avec le débit d'air nasal, avec une valeur de corrélation r approchant de 0,75. En revanche, l'évolution au fil des couches est plus discernable avec les séquences courtes. La corrélation est significative dans les couches initiales et s'affaiblit progressivement. La couche n° 4, montrant la corrélation la plus forte avec les mesures aérodynamiques, se rapproche également des corrélations obtenues avec les séquences longues. Ainsi, nous avons choisi d'étudier cette couche pour comparer les séquences longues et courtes dans la section 4.

4. Interprétation des résultats et discussion

Dans notre étude, nous avons examiné la mesure de la nasalité en utilisant l'apprentissage autosupervisé, wav2vec 2.0. Nous avons montré que :

— les performances de deux encodeurs (LeBenchmark et XLSR) sont très semblables ;

— un classifieur de type MLP est préférable avec de meilleurs taux de classification dans une partie des expériences et des taux équivalents dans l'autre partie ;

— l'extraction de séquences longues autour du phonème permet d'obtenir de meilleures classifications sur nos classes phonémiques.

Dans cette section, nous évaluons les questions subséquentes à ces résultats. Puisque wav2vec 2.0 est un modèle spécifiquement entraîné pour la reconnaissance automatique de la parole, en quoi la classification diffère d'un apprentissage des pho-

nèmes de la langue ? Peut-on interpréter ces résultats à la lumière des mesures aérodynamiques prises en parallèle du signal acoustique ? Peut-on déterminer à l'aide de ces modèles la variabilité inter- et intralocuteur que l'on observe sur les données aérodynamiques ? Toutes ces questions abordent une question centrale : la nasalité phonémique et la nasalité phonétique peuvent-elles êtres distinguées par nos modèles ?

4.1. *Détection de la nasalité phonétique ou phonémique*

Notre objectif initial était d'explorer une nouvelle méthode de mesure de la nasalité en encodant les données audio brutes à l'aide de deux variantes du modèle d'apprentissage automatique autosupervisé : W2V2-LeBenchmark et W2V2-XLSR.

Dans l'ensemble, nos classifieurs ont montré des performances élevées dans la classification de la nasalité, jusqu'à 91 % d'exactitude globale pour les séquences courtes, et 99 % pour les séquences longues.

Lorsque nous avons analysé l'évolution des représentations vectorielles à travers les couches de wav2vec 2.0, nous avons remarqué deux comportements distincts :

– pour les séquences courtes, il est possible pour nos deux modèles de classifier avec une grande précision les stimuli en utilisant les représentations issues des premières couches du modèle wav2vec 2.0. Cependant, à mesure que nous avançons dans les couches du modèle, leur performance diminue ;

– pour les séquences longues, l'utilisation des vecteurs contextuels extraits de toutes les couches du modèle wav2vec 2.0 améliorent les performances par rapport aux séquences courtes. Cependant, aucune évolution de la performance n'est observée au fil des couches, que ce soit une amélioration ou une détérioration.

Ces observations concordent avec les recherches antérieures sur l'évolution des couches de wav2vec 2.0 (Pasad *et al.*, 2021 ; Pasad *et al.*, 2022), qui ont montré que les couches du *Transformer* suivent la hiérarchie acoustique-linguistique, à savoir que les traits acoustiques sont fortement associés aux couches initiales du *Transformer*, suivis par l'identité phonétique vers la couche 10, et pour finir l'identité lexicale puis sémantique. Pour les séquences courtes, c'est l'information acoustique qui est majoritairement pertinente pour notre classifieur. Quant aux séquences longues, l'information pertinente est située entre la couche dédiée aux traits acoustiques et celle liée à l'identité phonétique, ce qu'on pourrait interpréter comme étant le trait (ou la classe) phonologique de nasalité.

La figure 9 présente une visualisation des représentations contextuelles utilisant la méthode t-SNE, qui permet de réduire la dimensionnalité des données de haute dimension pour les visualiser (Van der Maaten et Hinton, 2008). À gauche, les caractéristiques extraites de la quatrième couche du *Transformer* (T4) sur les séquences longues sont représentées, et à droite, ce sont les *embeddings* issus de l'apprentissage, c'est-à-dire de la couche dense du modèle MLP (MLP-T4). Les différentes couleurs indiquent les traits de nasalité des phones. Nous observons une distinction moins nette

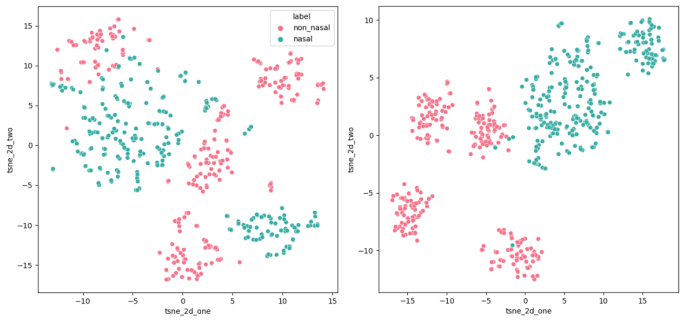


FIGURE 9. *Diagramme t-SNE : représentations de la quatrième couche de* Transformer *(gauche) et de la couche dense du MLP-T4 (droite)*

entre les segments des deux catégories ainsi qu'un éparpillement sur l'espace à deux dimensions, suggérant que les *embeddings* du wav2vec 2.0 (T4) ne distinguent pas en premier lieu la nasalité des phones. En revanche, les *embeddings* du MLP-T4 permettent une distinction plus évidente car les phones de chaque classe se distinguent selon leur nasalité. Ces résultats mettent en évidence l'importance de l'utilisation du MLP pour spécialiser le modèle dans la détection de la nasalité. En effet, wav2vec 2.0 a été entraîné dans l'optique d'une tâche de reconnaissance automatique de la parole et il est attendu que les phonèmes soient appris au cours de cette tâche. Le MLP utilisé à la fin de notre expérience permet de rediriger cet apprentissage.

Ces résultats ont renforcé notre conviction qu'il était possible de faire une distinction entre les caractéristiques acoustiques de la nasalité et la nasalité en tant que classe phonémique. La section suivante vise à explorer les mesures physiologiques dans le but d'interpréter les décisions de wav2vec 2.0 dans ces termes.

4.2. *Explicabilité par des mesures aérodynamiques*

Nous avons effectué précédemment (section 3.1.4) une étude aérodynamique en utilisant la mesure du débit d'air nasal pour déterminer si les classifications entrent dans une relation de corrélation avec les mesures physiologiques. Dans cette section, nous avons mené une étude comparative entre les mauvaises attributions et les différents débits d'air afin d'établir si elles résultaient de la réalisation phonétique des phones de classe nasale avec plus d'oralité, ou des phonèmes de classe orale prononcés avec nasalité. Les phénomènes de coarticulation très présents dans la parole sont particulièrement marqués pour la nasalité (le voile du palais étant un organe plus lent), ils influencent donc la réalisation phonétique de nasalité d'un phonème quelle que soit sa classe. Un phonème oral identifié comme nasal par notre classifieur pourrait *in fine* être dû à une réalisation nasale bien que ce phonème reste oral phonologiquement.

Nous avons observé deux tendances communes à notre modèle MLP-T4 :

– lorsque le débit d'air nasal est en dehors de la moyenne pour les phones de classe nasale, les classifieurs les identifient comme des phones oraux ;

– lorsque le débit d'air nasal n'est pas compris dans la plage des moyennes ou qu'il est négatif pour les phones de classe orale, les classifieurs les reconnaissent comme des phones nasals.

Ces deux schémas sont associés à des valeurs atypiques pour chaque catégorie. Par exemple, dans l'ensemble, les voyelles /ɛ̃/ présentent des valeurs de débit d'air nasal (DAN) plus faibles car cette voyelle a le plus faible flux nasal des voyelles nasales du français (Amelot, 2004). Un autre exemple concret pour la classe orale concerne la voyelle /o/ incorrectement identifiée. Le DAN de cette voyelle présente une plage de valeurs particulièrement étendue, pouvant être au-dessus de la moyenne dans certains cas, mais en dessous de zéro dans d'autres. Pour les valeurs négatives, cette voyelle correspond vraisemblablement à l'un des schémas décrits par Ohala *et al.* (1975), où l'explosion de la consonne est anticipée au début de la voyelle. Afin de vérifier cette hypothèse, une analyse de l'environnement phonémique, qui influence le débit d'air nasal, serait nécessaire dans une étude ultérieure.

4.3. *Variabilité inter- et intralocuteur*

Nous avons également exploré la capacité de nos modèles d'apprentissage automatique à capturer la nasalité spécifique à chaque locuteur, ainsi que la variabilité intralocuteur et interlocuteur.

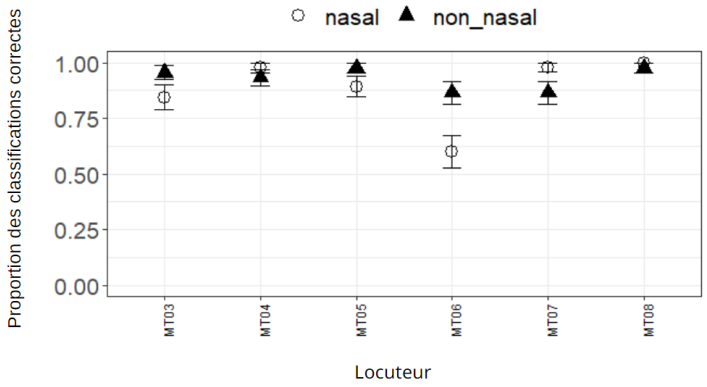


FIGURE 10. *Taux de classification correcte par MLP-T4 pour les phones de catégorie nasale et orale par locuteur (séquences courtes)*

Nous avons étudié la distribution des bonnes attributions en fonction des locuteurs avec le modèle MLP, avec la quatrième couche de *Transformer* (MLP-T4 pour les

séquences courtes[2]). Il est apparu dans la figure 10 que les productions nasales des locuteurs MT03 et MT06 sont plus souvent identifiées comme orales et classées de manière incorrecte. Les erreurs sur ces locuteurs ne peuvent pas souvent être justifiées uniquement par le débit d'air nasal. En revanche, pour le locuteur MT07, les phones de classe orale ont été plus précisément identifiés que ceux de classe nasale.

En se basant sur ces constatations, nous pouvons envisager que ces locuteurs ont une voix très distinctive par rapport aux autres, ce qui est reflété à la fois par les erreurs de classification et par les mesures physiologiques. Par exemple, ces résultats suggèrent des types de voix différents selon les locuteurs :

– les locuteurs MT03 et MT06 présentent une caractéristique vocale distinctive. Pour MT03, les mesures de débit d'air nasal (propre et proportionnel) sont minimales et les mesures de débit d'air buccal sont plus élevées pour la classe nasale. Les productions des phones [+nasal] de MT06 sont caractérisées par un débit d'air buccal minimal, avec un débit d'air nasal se situant au milieu de ceux des autres locuteurs, entraînant ainsi un débit d'air nasal proportionnellement plus élevé ;

– le locuteur M07 a une voix plus nasalisée, avec un débit d'air nasal le plus élevé parmi les locuteurs ;

– les autres locuteurs présentent une caractéristique vocale avec une bonne distinction entre la production orale et nasale de la voix. Les mauvaises attributions pour ces locuteurs s'expliquent par le débit d'air nasal.

Les mauvaises attributions de classe en fonction des phonèmes et des locuteurs avec leur débit d'air nasal ont également été étudiées. Elles permettent d'apporter une explication sur nos analyses sur les caractéristiques vocales :

– pour les locuteurs MT06 et MT03, les erreurs de classification ne peuvent pas être expliquées uniquement par le débit d'air nasal, mais plutôt par le débit d'air buccal. Pour MT06, une augmentation est observée pour les phones de classe nasale, tandis que pour MT03, elle est observée pour les phones [- nasal] ;

– le locuteur MT07 présente des erreurs de classification sur les phones [+ nasal] lorsque ces derniers sont prononcés avec un faible débit d'air nasal ;

– les erreurs des autres locuteurs sont liées au débit d'air nasal : les phones [+ nasal] avec un faible débit d'air nasal sont identifiés comme [- nasal], tandis que ceux [- nasal] avec un débit d'air nasal élevé sont identifiés comme [+ nasal].

4.4. *Limites et futures études*

Dans le cadre de cette étude, nous avons effectué une comparaison entre les probabilités attribuées par le classifieur et une mesure aérodynamique afin de confirmer

2. Les séquences longues ne permettent pas une distinction nette entre les locuteurs, car seules dix erreurs de classification sont observées au total, et celles-ci sont bien réparties parmi les locuteurs.

que notre modèle est capable de détecter la nasalité phonétique comme phonémique. Pour cette validation, nous avons utilisé le niveau de différents débits d'air comme point de référence, bien que certaines limites aient été identifiées. Il est possible – bien que peu probable – que les valeurs de débit d'air nasal proportionnel soient erronées, notamment lorsqu'elles ont une valeur négative. Cependant, dans certaines situations, la nasalité reste perceptible dans le segment, même lorsque le débit d'air nasal est aussi réduit que celui d'un phone oral. Un exemple caractéristique est observé avec les deuxièmes voyelles nasales dans un logatome, tel que /ɔ̃tɔ̃/. Dans ce cas, la deuxième voyelle nasale présente systématiquement un débit d'air nasal inférieur à celui de la première voyelle nasale du logatome. Pourtant, lors de l'écoute de l'extrait de la deuxième voyelle, la nasalité demeure perceptible. Si ces variations spécifiques peuvent être articulatoirement expliquées (Ohala *et al.*, 1975) sans remettre en question les résultats de nos modèles, il n'en reste pas moins qu'une étude perceptive permettant de valider la nasalité identifiée par des auditeurs naïfs sera pertinente afin de mieux interpréter les résultats mais également pour une meilleure caractérisation de la voix du locuteur.

5. Conclusion

L'étude que nous avons proposée visait à élucider la manière dont la nasalité des sons en français est catégorisée par deux variantes de modèles neuronaux wav2vec 2.0. Ces deux modèles, qu'ils soient préentraînés exclusivement sur le français ou non, présentent une exactitude globale élevée dans la classification de la nasalité pour les consonnes et les voyelles. L'application d'un MLP après récupération des *embeddings* a permis de spécialiser notre modèle vers la détection de la nasalité, alors que ces modèles neuronaux sont initialement entraînés pour l'apprentissage des phonèmes. Nous avons examiné deux longueurs audio pour extraire des représentations contextuelles. Les séquences courtes ont permis de saisir globalement la nasalité phonétique ou acoustique, ce qui explique en partie les erreurs de classification. En effet, il n'est pas rare que des phonèmes de classe orale soient nasalisés par le contexte et/ou par le locuteur (et inversement), et ces phénomènes peuvent être confirmés par les mesures aérodynamiques. En revanche, les séquences longues ont mieux capturé la nasalité phonémique et les contrastes phonologiques entre les phonèmes avec des performances proches de 100 %.

6. Bibliographie

Adi Y., Kermany E., Belinkov Y., Lavi O., Goldberg Y., « Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks », 8, 2016.

Amelot A., « Étude aérodynamique, fibroscopique, acoustique et perceptive des voyelles nasales du français », 2004.

Amodei D., Ananthanarayanan S., Anubhai R., Bai J., Battenberg E., Case C., Casper J., Catanzaro B., Cheng Q., Chen G. *et al.*, « Deep speech 2 : End-to-end speech recognition in

english and mandarin », *International conference on machine learning*, PMLR, p. 173-182, 2016.

Baevski A., Zhou H., Mohamed A., Auli M., « wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations », 6, 2020.

Berti F. B., « An electromyographic study of velopharyngeal function in speech », *Journal of Speech and Hearing Research*, vol. 19, p. 225-240, 1976.

Campbell J. P., « Speaker recognition : A tutorial », *Proceedings of the IEEE*, vol. 85, nº 9, p. 1437-1462, 1997.

Carignan C., « Covariation of nasalization, tongue height, and breathiness in the realization of F1 of Southern French nasal vowels », *Journal of Phonetics*, vol. 63, p. 87-105, 7, 2017.

Carignan C., « A practical method of estimating the time-varying degree of vowel nasalization from acoustic features », *The Journal of the Acoustical Society of America*, vol. 149, p. 911-922, 2, 2021.

Chanclu A., Georgeton L., Fredouille C., Bonastre J.-F., « PTSVOX : une base de données pour la comparaison de voix dans le cadre judiciaire (PTSVOX : a Speech Database for Forensic Voice Comparison) », *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition)*, p. 73-81, 2020.

Chen M. Y., « Acoustic parameters of nasalized vowels in hearing-impaired and normal-hearing speakers », *The Journal of the Acoustical Society of America*, vol. 98, nº 5, p. 2443-2453, 1995.

Chen M. Y., « Acoustic correlates of English and French nasalized vowels », *The Journal of the Acoustical Society of America*, vol. 102, nº 4, p. 2360-2370, 1997.

Chollet F. *et al.*, « Keras », , https ://keras.io, 2015.

Clarke W. M., « The measurement of the oral and nasal sound pressure levels of speech », *Journal of Phonetics*, vol. 3, nº 4, p. 257-262, 1975.

Cohn A. C., « Nasalisation in English : Phonology or phonetits », *Phonology*, vol. 10, p. 43-81, 1993.

Conneau A., Baevski A., Collobert R., Mohamed A., Auli M., « Unsupervised Cross-lingual Representation Learning for Speech Recognition », 6, 2020.

Conneau A., Kruszewski G., Lample G., Barrault L., Baroni M., « What you can cram into a single vector : Probing sentence embeddings for linguistic properties », 5, 2018.

Crosby D. M., « OPPA-NG GAMSAHAMNITA-NG i The Phonetics of Nasal Cutenes final », n.d.

Dang J., Honda K., Suzuki H., « Morphological and acoustical analysis of the nasal and the paranasal cavities », *The Journal of the Acoustical Society of America*, vol. 96, p. 2088-2100, 1994.

Delattre P., « Les Attributs Acoustiques De La Na-Salité Vocalique Et Consonantique », *Studia linguistica*, vol. 8, nº 1-2, p. 103-109, 1954.

Delvaux V., « Étude aérodynamique de la nasalité en français », *Actes des XXIIIe JEP*, 2000.

Devlin J., Chang M.-W., Lee K., Toutanova K., « Bert : Pre-training of deep bidirectional transformers for language understanding », *arXiv preprint arXiv :1810.04805*, 2018.

Elmerich A., Amelot A., Maeda S., Laprie Y., Papon J. F., Crevier-Buchman L., « F1 and F2 measurements for French oral vowel with a new pneumotachograph mask », *ISSP 2020-12th International Seminar on Speech Production*, 2020.

Elmerich A., Gao J., Amelot A., Crevier-Buchman L., Maeda S., « Combining acoustic and aerodynamic data collection : A perceptual evaluation of acoustic distortions », *Interspeech 2023*, 2023a.

Elmerich A., Kim L., Gendrot C., Amelot A., Crevier-Buchman L., Maeda S., « Nasality detection from acoustic data with a convolutional neural network and comparison with aerodynamic data », 2023b.

English P. C., Kelleher J., Carson-Berndsen J., « Domain-informed probing of wav2vec 2.0 embeddings for phonetic features », *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, p. 83-91, 2022.

Esling J. H., Moisik S. R., Benner A., Crevier-Buchman L., « voice quality », 2019.

Fant G., *Acoustic theory of speech production : with calculations based on X-ray studies of Russian articulations*, n° 2, Walter de Gruyter, 1971.

Feng G., Modélisation acoustique et traitement du signal de parole : le cas des voyelles nasales, PhD thesis, Institut National Polytechnique de Grenolde, 1986.

Fily M.*, « Caractérisation de la nasalité en contexte de parole : séparation du signal oral et nasal pour la recherche des corrélats de la nasalité dans le signal oral. Application au français et au mandarin »*, Master's thesis, May, 2018.

Fromkin V., Rodman R., Hyams V., *An Introduction to Language 6e*, Orlando, FL : Hartcourt Brace College Publishers, 1998.

Galliano S., Geoffrois E., Gravier G., Bonastre J.-F., Mostefa D., Choukri K., « Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. », *LREC*, Citeseer, p. 139-142, 2006.

Gold E., French P., « International practices in forensic speaker comparisons : second survey », *International Journal of Speech, Language and the Law*, vol. 26, n° 1, p. 1-20, 2019.

Graves A., Fernández S., Gomez F., Schmidhuber J., « Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks », *Proceedings of the 23rd international conference on Machine learning*, p. 369-376, 2006.

Gravier G., Bonastre J.-F., Geoffrois E., Galliano S., McTait K., Choukri K., « The ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. », *LREC*, 2004.

Guillaume S., Wisniewski G., Michaud A., « From'snippet-lects' to doculects and dialects : Leveraging neural representations of speech for placing audio signals in a language landscape », 2023.

Hannun A., Case C., Casper J., Catanzaro B., Diamos G., Elsen E., Prenger R., Satheesh S., Sengupta S., Coates A., Ng A. Y., « Deep Speech : Scaling up end-to-end speech recognition », 12, 2014.

Hinton G., Deng L., Yu D., Dahl G. E., Mohamed A.-r., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T. N. *et al.*, « Deep neural networks for acoustic modeling in speech recognition : The shared views of four research groups », *IEEE Signal processing magazine*, vol. 29, n° 6, p. 82-97, 2012.

House A. S., Stevens K. N., « Analog studies of the nasalization of vowels. », *The Journal of speech and hearing disorders*, vol. 21, p. 218-232, 1956.

Kahn J., « Parole de locuteur : performance et confiance en identification biométrique vocale », Avignon, 2011.

Kim L., Gendrot C., Elmerich A., Amelot A., Maeda S., « Détection de la nasalité du locuteur à partir de réseaux de neurones convolutifs et validation par des données aérodynamiques », *18e Conférence en Recherche d'Information et Applications\\16e Rencontres Jeunes Chercheurs en RI\\30e Conférence sur le Traitement Automatique des Langues Naturelles\\25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, ATALA, p. 101-108, 2023.

Lagefoged P., Maddieson I., « The sounds of the world's languages », 1996.

Lamel L. F., Gauvain J.-L., Eskénazi M. *et al.*, « Bref, a large vocabulary spoken corpus for french1 », *training*, vol. 22, n° 28, p. 50, 1991.

Laver J., « The Description of Voice Quality in General Phonetic », *Cambridge : CUP*, 2009.

Lee A., Gong H., Duquenne P.-A., Schwenk H., Chen P.-J., Wang C., Popuri S., Adi Y., Pino J., Gu J., Hsu W.-N., « Textless Speech-to-Speech Translation on Real Data », 12, 2021.

Lenglet M., Perrotin O., Bailly G., « A closer look at latent representations of end-to-end TTS models », *Journée commune AFIA-TLH/AFCP–"Extraction de connaissances interprétables pour l'étude de la communication parlée"*, 2023.

Ma D., Ryant N., Liberman M., « Probing Acoustic Representations for Phonetic Properties », 10, 2020.

Maddieson I., Abramson A. S., « Patterns of Sounds by Ian Maddieson », *The Journal of the Acoustical Society of America*, vol. 82, n° 2, p. 720-721, 08, 1987.

Maeda S., « Acoustic cues of vowel nasalization : a simulation study. », *Recherches/Acoustique*, 1982a.

Maeda S., « The role of the sinus cavities in the production of nasal vowels », *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7, IEEE, p. 911-914, 1982b.

Malécot A., « Vowel nasality as a distinctive feature in American English », *Language*p. 222-229, 1960.

Nolan F., « Forensic Speaker Identification and the Phonetic », *A Figure of Speech : A Festschrift for John Laver*p. 385, 2014.

Ohala J. J. *et al.*, « Phonetic explanations for nasal sound patterns », *Nasálfest : Papers from a symposium on nasals and nasalization*, Stanford University Language Universals Project Palo Alto, CA, p. 289-316, 1975.

O'Shaughnessy D., *speech communication human and machine*, Institute of Electrical and Electronics Engineers, 1987.

Parcollet T., Nguyen H., Evain S., Boito M. Z., Pupier A., Mdhaffar S., Le H., Alisamir S., Tomashenko N., Dinarelli M., Zhang S., Allauzen A., Coavoux M., Esteve Y., Rouvier M., Goulian J., Lecouteux B., Portet F., Rossato S., Ringeval F., Schwab D., Besacier L., « LeBenchmark 2.0 : a Standardized, Replicable and Enhanced Framework for Self-supervised Representations of French Speech », 9, 2023.

Pasad A., Chou J.-C., Livescu K., « Layer-wise Analysis of a Self-supervised Speech Representation Model », 7, 2021.

Pasad A., Shi B., Livescu K., « Comparative layer-wise analysis of self-supervised speech models », 11, 2022.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M.,

Perrot M., Duchesnay E., « Scikit-learn : Machine Learning in Python », *Journal of Machine Learning Research*, vol. 12, p. 2825-2830, 2011.

Puzar A., Hong Y., « Korean Cuties : Understanding Performed Winsomeness (Aegyo) in South Korea », *Asia Pacific Journal of Anthropology*, vol. 19, p. 333-349, 8, 2018.

Radford A., Kim J. W., Xu T., Brockman G., McLeavey C., Sutskever I., « Robust speech recognition via large-scale weak supervision », *International Conference on Machine Learning*, PMLR, p. 28492-28518, 2023.

Radford A., Narasimhan K., Salimans T., Sutskever I. *et al.*, « Improving language understanding by generative pre-training », 2018.

Robins R. H., « The phonology of the nasalized verbal forms in Sundanese », *Bulletin of the School of Oriental and African Studies*, vol. 15, n° 1, p. 138-145, 1953.

Rossato S., Badin P., Bouaouni F., « Velar movements in French : an articulatory and acoustical analysis of coarticulation », *Proceedings of the 15th international congress of phonetic sciences*, Barcelona, Spain, p. 3141-3144, 2003.

Serrurier A., 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., PhD thesis, Institut National Polytechnique de Grenoble-INPG, 2006.

Shah J., Singla Y. K., Chen C., Shah R. R., « What all do audio transformer models hear ? Probing Acoustic Representations for Language Delivery and its Structure », 1, 2021.

Stefanuto M., Vallée N., « Consonant systems : From universal trends to ontogenesis », *Proceedings of the XIVth International Congress of Phonetic Sciences*, vol. 3, p. 1973-76, 1999.

Stevens K. N., *Acoustic phonetics*, vol. 30, 2000.

Styler W., « On the acoustical features of vowel nasality in English and French », *The Journal of the Acoustical Society of America*, vol. 142, p. 2469-2482, 10, 2017.

Torreira F., Adda-Decker M., Ernestus M., « The Nijmegen corpus of casual French », *Speech Communication*, vol. 52, n° 3, p. 201-212, 2010.

Triantafyllopoulos A., Wagner J., Wierstorf H., Schmitt M., Reichel U., Eyben F., Burkhardt F., Schuller B. W., « Probing Speech Emotion Recognition Transformers for Linguistic Knowledge », 4, 2022.

Van der Maaten L., Hinton G., « Visualizing data using t-SNE. », *Journal of machine learning research*, 2008.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I., « Attention is all you need », *Advances in neural information processing systems*, 2017.

Wetzels W. L., « The lexical representation of nasality in Brazilian Portuguese », 1997.

Yang M., Shekar R. C. M. C., Kang O., Hansen J. H. L., « What Can an Accent Identifier Learn ? Probing Phonetic and Prosodic Information in a Wav2vec2-based Accent Identification Model », 6, 2023.

Zellou G., *Coarticulation in Phonology*, Cambridge University Press, 8, 2022.