Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Utilisation de wav2vec 2.0 pour des tâches de classifications phonétiques : aspects méthodologiques

Lila Kim[1] Cédric Gendroit[1]
(1) **Laboratoire de Phonétique et Phonologie (CNRS & U. Sorbonne Nouvelle)**, 4 rue des Irlandais, 75005 Paris, France
lila.kim@sorbonne-nouvelle.fr, cedric.gendrot@sorbonne-nouvelle.fr

RÉSUMÉ

L'apprentissage auto-supervisé, particulièrement dans le contexte de la parole, a démontré son efficacité dans diverses tâches telles que la reconnaissance du locuteur et la reconnaissance de la parole. Notre question de recherche se concentre sur l'efficacité des représentations vectorielles - extraites de phonèmes - plus courtes par rapport à des séquences plus longues dans la détection de la nasalité. Deux approches distinctes ont été étudiées : extraire des vecteurs sur la durée du phonème et prendre des séquences plus longues avec une seconde ajoutée de chaque côté du phonème, puis récupérer la partie centrale a posteriori. Les résultats révèlent que les modèles réagissent différemment selon les phones et les locuteurs, avec une variabilité observée à ces niveaux. Le modèle à séquences longues surpasse le modèle à séquences courtes en assurant une corrélation plus robuste avec le débit d'air nasal.

ABSTRACT

Using wav2vec 2.0 for phonetic classification tasks : methodological aspects

Self-supervised learning, particularly in the context of speech, has been shown to be effective in a variety of tasks such as speaker recognition and automatic speech recognition. Our research question focuses on the effectiveness of vector representations extracted from shorter versus longer phoneme sequences in detecting nasality. Two distinct approaches were studied : extracting vectors over the duration of the phoneme and taking longer sequences with a second added on each side of the phoneme, then recovering the central part a posteriori. The results show that the models react differently depending on the phone and the speaker, with variability observed at both levels. The long sequence model outperformed the short sequence model by correlating more robustly with nasal airflow.

MOTS-CLÉS : parole, wav2vec 2.0, nasalité, physiologie.

KEYWORDS: speech, wav2vec 2.0, nasality, physiology.

1 Introduction

Depuis l'utilisation récurrente de l'apprentissage auto-supervisé dans les tâches de reconnaissance automatique de la parole, plusieurs études ont appliqué ces modèles de Transformers à des domaines tels que la reconnaissance du locuteur, la détection de code-switching ou d'émotions, etc. (Fan et al., 2021; Pepino et al., 2021; Tseng et al., 2021; Cormac English et al., 2022). De plus, Pasad et al. ont montré notamment que les informations diffèrent dans les représentations vectorielles selon les

couches de Transformers (Pasad et al., 2022, 2023). Il est à noter qu'un modèle de Transformers tel que wav2vec 2.0 prend en compte les informations contextuelles dans une séquence et travaille généralement sur des séquences de plusieurs secondes (Baevski et al., 2020). C'est dans ce contexte que s'inscrit notre question de recherche : les représentations vectorielles extraites sur une séquence de phonèmes permettraient-elles une meilleure performance dans la détection de la nasalité par rapport à une séquence plus longue ? Notre travail consiste donc à explorer la longueur de la séquence pour la prise de vecteurs : la première consiste à prendre des vecteurs sur la durée du phonème, tandis que la seconde consiste à prendre une séquence plus longue en ajoutant une seconde dans les deux côtés du phonème et à récupérer la partie centrale a posteriori.

En premier lieu, nous décrirons les ressources utilisées lors de l'entraînement et du test, ainsi que les méthodes d'extraction utilisées avec le modèle auto-supervisé wav2vec 2.0. En second lieu, nous nous pencherons sur deux approches différentes pour l'extraction des vecteurs et la détection de la nasalité à l'aide d'une régression logistique. Nous évaluerons ensuite le modèle entraîné sur des données acoustiques, en comparant les résultats avec des données physiologiques obtenues simultanément avec l'acoustique, servant ainsi de référence.

2 État de l'art

2.1 Modélisation acoustique

La parole est un phénomène complexe influencé par divers éléments tels que l'articulation, l'origine géographique ou sociale du locuteur, son état émotionnel et des aspects pragmatiques comme l'auditeur. Pour la transcription automatique de la parole et la modélisation du signal acoustique, des approches ont évolué des systèmes experts vers des approches neuronales. Les premières approches se concentraient sur l'aspect linguistique, alors que les méthodes probabilistes, notamment les modèles de Markov cachés, ont commencé à dominer à partir des années 90. (Juang and Rabiner, 1991; Patel and Srinivas Rao, 2010).

Avec l'avènement des modèles connexionnistes, la phonétisation, l'utilisation de connaissances psycho-acoustiques et le traitement du signal se sont transformés vers des méthodes entièrement neuronales, combinant les représentations spectrales telles que les MFCC avec des perceptrons multicouches. Pour surmonter les défis du Deep Learning tels que le besoin de grandes quantités de données et le manque d'annotations manuelles, des approches d'apprentissage légèrement supervisées ou auto-supervisées ont été entreprises. Ces modèles sont préalablement entraînés sur de grands nombres d'heures d'audio non annoté, puis ajustés sur des ensembles de données annotées de plus moindre taille pour des tâches spécifiques (Baevski et al., 2020). En utilisant la méthode du "probing", Pasad et al. analysent les informations contenues dans les différentes couches de ces modèles en cherchant à mieux comprendre la nature des données à ces niveaux (Pasad et al., 2023, 2022).

2.2 Nasalité

La nasalité, fréquemment considérée comme une composante de la qualité de la voix, est une caractéristique omniprésente dans les langues du monde, se produit lorsque le voile du palais s'abaisse. Ce phénomène crée des effets acoustiques distincts sur les sons nasals (Maeda, 1982). Elle est

essentielle dans la production de la parole pour distinguer phonologiquement les sons nasals des sons oraux, que ce soit dans le cas des voyelles (comme /a/ et /ã/, par exemple) ou des consonnes (comme /b/ et /m/, par exemple). La nasalité d'un son peut être propagée à son voisin oral en raison de réalisations articulatoires, telles qu'un abaissement prématuré, un relèvement tardif du velum (Amelot et al., 2008; Brkan, 2018). Cette coarticulation nasale, influencée par le contexte phonémique, peut se produire dans des langues où la nasalité est un trait phonologique distinctif (comme le français, où /a/ dans "maman" est nasalisé), mais aussi dans des systèmes de langue où cette distinction n'est pas présente (comme en anglais, par exemple dans "can't").

La qualité de voix a de grandes implications dans la caractérisation du locuteur (Gold and French, 2019). Elle peut être un élément permanent de la voix d'un locuteur due à des facteurs physiologiques, mais aussi sujette à la variabilité intra-locuteur, notamment dans le style de discours ou l'émotion (Nolan, 2014). Les nasales offrent une caractéristique fiable pour la reconnaissance des locuteurs (Kahn, 2011) en raison de la morphologie de la cavité nasale stable et variable entre locuteurs (Dang et al., 1994; Serrurier, 2006). Cependant, l'analyse acoustique de la nasalité est complexe car le couplage de deux cavités provoquent des modifications acoustiques en engendrant des pôles et zéros sur le spectre acoustique. Bien que les méthodes d'analyse aient été entreprises pour la nasalité (Chen, 1997; Styler, 2017), elles sont très influencées par les caractéristiques articulatoires propres à chaque son, et à chaque locuteur.

3 Protocole expérimental

3.1 Données pour l'entraînement

Pour l'entraînement et la validation, nous avons extrait les différents types de phones à partir de quatre corpus distincts, chacun représentant un type de parole spécifique. Les corpus de données utilisés dans cette étude comprennent :

1. NCCFr (The Nijmegen Corpus of Casual French) : conversations amicales, impliquant 46 locuteurs français. (Torreira et al., 2010) ;
2. ESTER (Evaluation de Systèmes de Transcription enrichie d'Emissions Radiophoniques) : conversations radiophoniques en français, parole préparée et lue (Gravier et al., 2004; Galliano et al., 2006). Seule une partie de 30 heures a été retenue pour cet entraînement.
3. PTSVOX : créé pour évaluer les variations intra- et inter-locuteurs. (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 ;
4. BREF : développé dans le but du développement et de l'évaluation des systèmes de reconnaissance de la parole, 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.

Dans le cadre de ce travail, nous avons décidé d'extraire 8 voyelles et 7 consonnes nasales et orales confondues. Les voyelles sujettes à l'extraction sont 3 paires de voyelles /a,ɛ,o,ɑ̃,ɛ̃,ɔ̃/ qui peuvent se distinguer par le trait de nasalité [± nasal]. Nous sommes conscients de la distinction articulatoire entre une voyelle orale et sa contrepartie nasale, cependant, dans le contexte de cette étude, nous avons choisi de concentrer notre attention sur la nasalité en particulier. Deux voyelles /e,ɔ/ ont été ajoutées car la phonétisation des voyelles moyennes en français n'est pas toujours systématique. En

ce qui concerne les consonnes, nous avons retenu quatre consonnes orales et trois consonnes nasales : /b,d,v,l,m,n,ɲ/. Elles présentent différentes manières et lieux d'articulation (bilabiale, labio-dentale, dentale, alvéolaire, et occlusive ou fricative). La même liste de phonèmes a été utilisée pour les deux approches mentionnées ultérieurement dans la section 3.3.3, sans prendre en compte le contexte phonétique des phonèmes examinés.

3.2 Données acoustiques et physiologiques pour l'évaluation du modèle

Les données de test se composent de deux parties : acoustique et physiologique. Elles ont été recueillies simultanément à l'aide d'un masque "Aeromask" développé au Laboratoire de Phonétique et Phonologie (Elmerich et al., 2023). Ce masque enregistre la voix ainsi que les débits d'air nasal et buccal sans perturber la propagation sonore, ce qui permet d'utiliser l'acoustique pour évaluer le réseau de neurones et les débits d'air pour vérifier la présence de nasalité dans les phones évalués. Les enregistrements des phrases ont été réalisés avec six locuteurs masculins, tous natifs du français. Les stimuli étaient insérés dans des mots sans signification littérale (i.e., logatomes) sous forme de 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]. (Elmerich et al., 2020, 2023). 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 ». Ainsi, les mots XXX, YYY et ZZZ correspondent à des structures VCV ou VNV. La segmentation manuelle a été effectuée pour ces données. À partir de ces listes de phones, nous avons sélectionné les mêmes phones que pour l'entraînement, à l'exception de /l/ qui n'est pas présent dans la liste, ce qui donne /a,ɛ,o,ɑ̃,ɛ̃,ɔ̃,b,d,v,m,n/. En résumé, 269 sons de chaque classe ont été extraits au total. Les mesures aérodynamiques des phones ont été consignées dans un fichier au format CSV en vue d'une comparaison ultérieure avec les résultats du réseau de neurones profonds. Les données utilisées pour l'entraînement, la validation et le test sont récapitulées dans le tableau 1. Les phonèmes des données de test ont été répartis selon le nombre suivant :ɑ̃ (66), ɛ̃ (66), ɔ̃ (66), m (36), n (35), a (66), E (66), o (66), b (25), d (29), v (17).

Jeu de données	**Phone [+ nasal]**	**Phone [- nasal]**
Entraînement	60 000	60 000
Validation	15 000	15 000
Test	269	269

TABLE 1 – statistiques des données utilisées pour l'entraînement, la validation et le test

Les données aérodynamiques que nous utiliserons comme référence consistent en trois valeurs : le débit d'air nasal (DAN), le débit d'air buccal (DAB) et le débit d'air nasal proportionnel. Le calcul de ces valeurs est expliqué dans (Kim et al., 2023).

3.3 Méthodologie

3.3.1 Wav2vec 2.0

Notre recherche s'inscrit dans le cadre de l'exploration de la manière dont le modèle wav2vec 2.0 encode l'information de nasalité dans ses représentations vectorielles. Nous nous concentrons

particulièrement sur le modèle "wav2vec 2.0-FR-3K-large-LeBenchmark", pré-entraîné sur 2 900 heures de divers types de discours en français (spontané, lu et diffusé). Ce modèle a été spécifiquement conçu pour optimiser ses performances dans des tâches liées au français (Parcollet et al., 2023).

Le fonctionnement du modèle wav2vec 2.0 consiste à prendre le signal brut comme données d'entrée, traitées par l'encodeur convolutionnel. Toutes les 25 millisecondes d'audio sont transformées en une séquence de vecteurs, avec un chevauchement de 5 millisecondes entre chaque paire d'échantillons. Ces séquences subissent une normalisation et une fonction d'activation GELU avant d'être acheminées vers les transformers. Pendant la phase de pré-entraînement, le module de quantification est utilisé pour discrétiser les valeurs de sortie de l'encodeur. Les représentations latentes obtenues de l'encodeur subissent ensuite une analyse et une contextualisation par les couches de Transformers, qui capturent l'information sur l'ensemble de la séquence. Le modèle large, en particulier, comporte 24 couches de transformation, chacune produisant un vecteur de 1 024 dimensions en représentations latentes. Les dimensions de la feed-forward sont de 4 096, avec 16 mécanismes d'attention (Baevski et al., 2020).

3.3.2 Génération des représentations vectorielles

L'approche d'extraction des embeddings s'appuie sur la méthodologie présentée par (Guillaume et al., 2023), dont l'étude se focalise sur une analyse linguistique d'une langue à partir de la parole dans un extrait audio de 5 secondes où la stratégie de max pooling a été utilisée pour agréger les différentes représentations latentes d'un enregistrement en un seul vecteur, qui représente l'ensemble du signal. Avec cette méthode, deux longueurs d'extrait audio ont été étudiées pour obtenir les représentations vectorielles de nos données. La première, inspirée de l'approche phonétique, implique l'extraction des représentations vectorielles directement sur les phonèmes découpées à leurs frontières. Dans notre cas d'étude, la petite taille des fenêtres d'analyse que nous utilisons rend l'affinage (fine-tuning) impossible. La deuxième approche consiste à utiliser des séquences plus longues, en ajoutant une seconde au début et à la fin du phonème. Une fois que le wav2vec 2.0 prend des caractéristiques contextuelles sur toute la séquence, nous récupérons le vecteur du milieu a posteriori. Par exemple, si la voyelle dure 200 ms, nous avons extrait une séquence de 2,2 secondes, puis effectué un max pooling sur les 200 ms du milieu lors de la récupération des caractéristiques vectorielles. De cette manière, l'information contextuelle sur l'ensemble de la séquence de 2,2 secondes peut être capturée par les blocs de transformations et être présente dans le vecteur du milieu qui représenterait le phonème en question. La récupération du milieu a été effectuée en retirant les secondes ajoutées. Nous reconnaissons que le vecteur du milieu ne correspondrait pas parfaitement à l'ensemble du phonème en question, rendant ainsi la comparaison imparfaite. Les représentations vectorielles ainsi obtenues ont été labellisées à l'aide des labels phonologiques des sons [+ nasal] et [- nasal].

3.3.3 Feature probing

Un modèle de régression logistique a été mis en place pour déterminer si le phone prononcé est réalisé avec nasalité (=1) ou sans nasalité (=0). Pour cela, la bibliothèque d'apprentissage automatique en python "scikit-learn" a été utilisée avec les hyperparamètres définis par défaut. La probabilité d'appartenance à la classe nasale est considérée comme une probabilité de nasalité dans les analyses (voir 4.2). La procédure de notre méthodologie est décrite dans la figure 1.

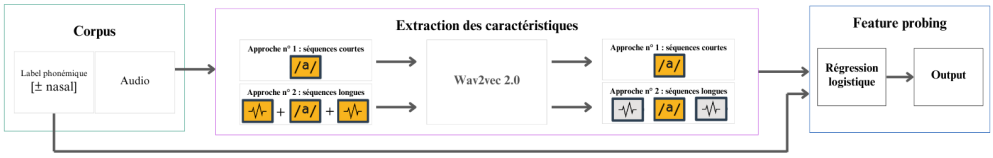

FIGURE 1 – Aperçu de la méthodologie expérimentale comprenant l'architecture du modèle de régression logistique

4 Résultats

L'analyse des résultats obtenus avec les réseaux de neurones profonds par la mesure physiologique aide à vérifier l'indice de nasalité dans la réalisation des phones. Dans la section 4.1, notre objectif est d'établir si la nasalité est détectable lorsqu'un classifieur est basé sur les caractéristiques extraites par le modèle auto-supervisé wav2vec 2.0, et si les erreurs produites par les réseaux peuvent être expliquées par les débits d'air nasal et buccal. Enfin, dans la section 4.2, nous chercherons à déterminer si les classifieurs ont appris à séparer les phonèmes plutôt qu'à détecter la nasalité, en utilisant les mêmes représentations vectorielles.

4.1 Performance du système selon l'approche d'extraction

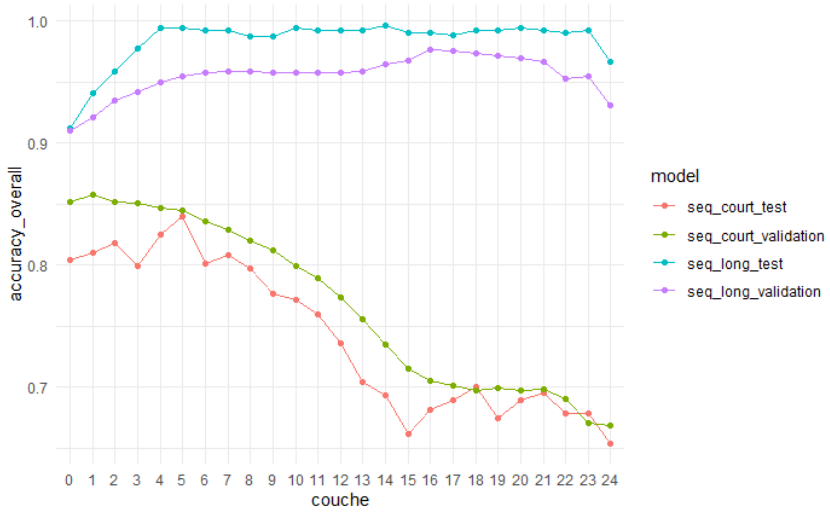

FIGURE 2 – Distribution de l'exactitude globale en fonction des couches du wav2vec 2.0 selon la longueur d'extrait audio

Les taux d'exactitude globale pour la caractéristique de nasalité [± nasal] à travers les différentes couches de Transformers sont illustrés dans la figure 2. Il convient de noter que le seuil sur les probabilités de sortie est fixé à 0,5 pour le choix d'une classe. Afin de déterminer la couche optimale à exploiter, nous avons examiné l'évolution des performances des différentes couches en ce qui concerne la nasalité, allant de l'encodeur CNN à la dernière couche de Transformers. La figure met en évidence la présence d'informations liées à la nasalité dans pratiquement toutes les couches lorsque l'extrait audio est long. Sur les séquences courtes, la nasalité est particulièrement marquée dans la

sortie de l'encodeur CNN et dans les premières couches de Transformers.

Selon Pasad et al., les premières couches du modèle wav2vec 2.0, y compris l'encodeur CNN, sont associées à l'identité acoustique et aux caractéristiques du spectrogramme (Pasad et al., 2022, 2023). À la lumière de ces observations, nous avons décidé de nous focaliser sur la première couche de Transformer afin d'améliorer l'identification de la nasalité en utilisant les caractéristiques acoustiques plutôt que phonémiques. Ainsi, dans le cadre de la classification de la nasalité avec les caractéristiques extraites de la première couche, les performances ont été meilleures pour les séquences longues, avec une exactitude globale de 94.05%, par rapport à 81.04% pour les séquences courtes.

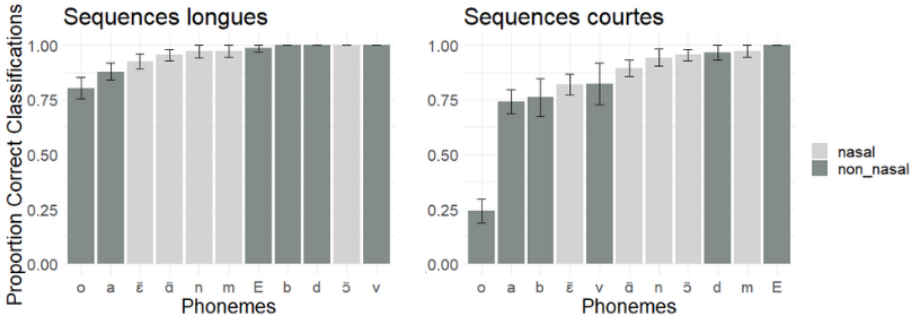

FIGURE 3 – Taux de classification correcte pour chaque phonème (séquences longues à gauche et séquences courte à droite)

Dans la figure 3, la proportion de bonnes attributions de classe pour chaque phonème est représentée. Il convient de noter que dans cette visualisation, /E/ représente /e,ɛ/. Les performances varient selon les phonèmes et les modèles. Par exemple, les phonèmes /ɔ̃,E,m,n,d/ présentent un taux élevé de bonnes classifications, tandis que les voyelles orales /o,a/ sont moins bien classées par les deux modèles. De plus, les voyelles nasales présentent des niveaux de difficulté similaires : /ɛ̃/ est considérée comme la plus difficile à détecter en termes de nasalité, tandis que /ɔ̃/ est identifiée comme la plus facile. En ce qui concerne le modèle à séquences longues, la consonne la plus difficile à prédire est la nasale /n,m/, tandis que pour le modèle à séquences courtes, c'est la consonne orale /b/.

4.2 Comparaison des résultats de classifieurs avec les données physiologiques

Dans cette étude, le coefficient de corrélation de Pearson est utilisé pour examiner la relation linéaire entre la probabilité d'appartenance à la catégorie nasale et le débit d'air nasal. Les corrélations ont été mesurées de trois façons distinctes : (i) nous avons utilisé le débit d'air nasal tel qu'obtenu par l'aeromask (ii) le débit d'air nasal pour chaque paire minimale de phones nasal et oral, par exemple /a/-/ɑ̃/. (iii) la normalisation a été réalisée pour chaque paire nasal-oral et pour chaque locuteur. Pour la consonne /v/ sans correspondant nasal, la valeur a été normalisée par rapport à son ensemble.

Que ce soit avec le débit d'air brut ou normalisé, le modèle à séquences longues présente une corrélation plus forte que le modèle à séquences courtes. Nous remarquons deux observations en commun pour les deux modèles. Dans l'ensemble, les probabilités de nasalité sont les plus fortement corrélées avec les valeurs normalisées par phonème et par locuteur. Ceci montre que le débit d'air nasal est propre aux phonèmes et aux locuteurs. Ensuite, la corrélation est la plus forte pour le locuteur MT04 et cette observation est commune dans les deux modèles. Cependant, le locuteur ayant la corrélation la plus faible diffère selon la longueur d'extrait audio et les mesures de débit d'air nasal.

Débit d'air nasal	Locuteur	MT03	MT04	MT05	MT06	MT07	MT08	Tous
moyenne	Séquences longues	0,75	0,73	0,68	0,70	0,76	0,77	0,70
phonèmes		0,66	0,76	0,68	0,68	0,72	0,72	0,68
phonèmes+locuteurs		0,70	0,79	0,69	0,68	0,73	0,68	0,71
moyenne	Séquences courtes	0,61	0,65	0,59	0,59	0,46	0,69	0,55
phonèmes		0,52	0,69	0,58	0,51	0,45	0,64	0,53
phonèmes+locuteurs		0,55	0,70	0,61	0,52	0,48	0,60	0,57

TABLE 2 – Comparaison des résultats obtenus avec les modèles de classification avec le débit d'air nasal (DAN) à l'aide du coefficient de corrélation de Pearson.

5 Discussion et conclusion

Notre objectif était d'étudier la longueur des séquences pour l'extraction de vecteurs afin de faciliter une tâche de classification phonétique, en particulier celle de la nasalité. Deux longueurs ont été examinées : une séquence d'un phonème et une séquence plus étendue avec une seconde ajoutée de chaque côté du phonème. Ces deux approches ont donné des performances satisfaisantes dans la tâche proposée. Les séquences plus longues ont atteint une exactitude globale de 94,05 %, tandis que les séquences plus courtes ont obtenu 81,04 %.

Nos deux modèles ont réussi à se spécialiser à la nasalité dans la parole, mais avec un comportement variant selon les phonèmes et locuteurs. Dans la section 4.1, il a été démontré que le comportement des modèles diffère selon les phonèmes, ce phénomène peut s'expliquer par la variation des positions des articulateurs pendant la réalisation d'un phone et par le fait que le voile du palais se positionne différemment selon les voyelles (Delvaux and Metens, 2002; Amelot et al., 2008). Par exemple, dans le cas de /ɑ̃/, qui est le plus correctement identifié parmi les voyelles nasales, le voile du palais s'abaisse davantage et la position de la langue devient plus postérieure jusqu'à ce qu'elle atteigne le velum. Comme cette voyelle induit une ouverture de la bouche ainsi qu'une ouverture du port vélopharyngé, l'air peut circuler dans la cavité nasale (Delvaux and Metens, 2002; Amelot et al., 2008).

La comparaison entre les probabilités de nasalité et les données physiologiques révèle une corrélation entre le débit d'air nasal et les probabilités obtenues avec nos modèles. Cette relation de corrélation varie en fonction des phonèmes et des locuteurs. Par exemple, la corrélation est plus forte lorsque le débit d'air nasal est normalisé par phonème et par locuteur que pour le DAN brut. Le locuteur MT04 présente une corrélation particulièrement forte. Ce locuteur peut être considéré comme ayant une bonne distinction entre la production orale et nasale de la voix.

En conclusion, notre étude a mis en évidence l'utilisation de deux longueurs de séquences pour extraire des informations vectorielles dans le cadre d'une tâche spécifique liée à la nasalité. En comparant nos classifieurs avec des mesures aérodynamiques, une corrélation significative a été observée entre les débits d'air nasal et les probabilités de nasalité. Les résultats révèlent le comportement différencié des modèles selon les phonèmes et les locuteurs, avec une variabilité interlocuteur constatée. Les performances restent constantes chez les locuteurs ayant une bonne distinction entre la production orale et nasale dans la voix, ainsi que chez ceux possédant une voix distinctive. Cependant, il convient de noter que la spécification de l'entraînement sur la nasalité permet une bonne corrélation avec le débit d'air nasal, facilitant ainsi la mise en évidence de phénomènes tels que la quantité de nasalité.

Références

Angélique Amelot, Patricia Basset, Shinji Maeda, Kiyoshi Honda, and Lise Crevier-Buchman. Etude simultanée des mouvements du voile du palais et de l'ouverture du port vélopharyngé. *XXVIIè JEP*, pages 65–68, 2008.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations, October 2020. URL http://arxiv.org/abs/2006.11477. arXiv :2006.11477 [cs, eess].

Altijana Brkan. *Etude comparative des phénomènes de coarticulation nasale en anglais américain, bosnien, français, norvégien et ourdou*. PhD thesis, Université Sorbonne Paris Cité, 2018.

Anaïs Chanclu, Laurianne Georgeton, and Corinne Fredouille. PTSVOX : une base de données pour la comparaison de voix dans le cadre judiciaire. 2020.

Marilyn Y. Chen. Acoustic correlates of English and French nasalized vowels. *The Journal of the Acoustical Society of America*, 102(4) :2360–2370, October 1997. ISSN 0001-4966, 1520-8524. DOI : 10.1121/1.419620. URL https://pubs.aip.org/jasa/article/102/4/2360/562446/Acoustic-correlates-of-English-and-French.

Patrick Cormac English, John D. Kelleher, and Julie Carson-Berndsen. Domain-Informed Probing of wav2vec 2.0 Embeddings for Phonetic Features. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 83–91, Seattle, Washington, 2022. Association for Computational Linguistics. DOI : 10.18653/v1/2022.sigmorphon-1.9. URL https://aclanthology.org/2022.sigmorphon-1.9.

Jianwu Dang, Kiyoshi Honda, and Hisayoshi Suzuki. Morphological and acoustical analysis of the nasal and the paranasal cavities. *The Journal of the Acoustical Society of America*, 96(4) : 2088–2100, 1994. Publisher : Acoustical Society of America.

Véronique Delvaux and Thierry Metens. Propriétés acoustiques et articulatoires des voyelles nasales du français. 2002.

Amélie Elmerich, Angelique Amelot, Shinji Maeda, Yves Laprie, Jean Francois Papon, and Lise Crevier-Buchman. F1 and f2 measurements for french oral vowel with a new pneumotachograph mask. In *ISSP 2020-12th International Seminar on Speech Production*, 2020.

Amélie Elmerich, Jiayin Gao, Angelique Amelot, Lise Crevier-Buchman, and Shinji Maeda. Combining acoustic and aerodynamic data collection : A perceptual evaluation of acoustic distortions. In *INTERSPEECH 2023*, pages 3078–3082. ISCA, August 2023. DOI : 10.21437/Interspeech.2023-1918. URL https://www.isca-archive.org/interspeech_2023/elmerich23_interspeech.html.

Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu. Exploring wav2vec 2.0 on speaker verification and language identification, January 2021. URL http://arxiv.org/abs/2012.06185. arXiv :2012.06185 [cs, eess].

Sylvain Galliano, Edouard Geoffrois, Guillaume Gravier, Jean-François Bonastre, Djamel Mostefa, and Khalid Choukri. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *LREC*, pages 139–142. Citeseer, 2006.

Erica Gold and Peter French. International practices in forensic speaker comparisons : second survey. *International Journal of Speech, Language and the Law*, 26(1) :1–20, 2019.

G Gravier, J-F Bonastre, E Geoffrois, S Galliano, K Mc Tait, and K Choukri. The ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. 2004.

Séverine Guillaume, Guillaume Wisniewski, and Alexis Michaud. Fromsnippet-lects' to doculects and dialects : Leveraging neural representations of speech for placing audio signals in a language landscape. *arXiv preprint arXiv :2305.18602*, 2023.

B H Juang and L R Rabiner. Hidden Markov Models for Speech Recognition. 33(3), 1991.

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

Lila Kim, Cedric Gendrot, Amélie Elmerich, Angelique Amelot, and Shinji Maeda. Détection de la nasalité du locuteur à partir de réseaux de neurones convolutifs et validation par des données aérodynamiques. 2023.

Lori F Lamel, Jean-Luc Gauvain, Mazcine Eskénazi, et al. Bref, a large vocabulary spoken corpus for french1. *training*, 22(28) :50, 1991.

Shinji Maeda. Acoustic cues for vowel nasalization : A simulation study. *The Journal of the Acoustical Society of America*, 72(S1) :S102–S102, November 1982. ISSN 0001-4966, 1520-8524. DOI : 10.1121/1.2019690. URL https://pubs.aip.org/jasa/article/72/S1/S102/733010/Acoustic-cues-for-vowel-nasalization-A-simulation.

Francis Nolan. Forensic Speaker Identification and the Phonetic. *A Figure of Speech : A Festschrift for John Laver*, page 385, 2014. Publisher : Routledge.

Titouan Parcollet, Ha Nguyen, Solene Evain, Marcely Zanon Boito, Adrien Pupier, Salima Mdhaffar, Hang Le, Sina Alisamir, Natalia Tomashenko, Marco Dinarelli, Shucong Zhang, Alexandre Allauzen, Maximin Coavoux, Yannick Esteve, Mickael Rouvier, Jerome Goulian, Benjamin Lecouteux, Francois Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. LeBenchmark 2.0 : a Standardized, Replicable and Enhanced Framework for Self-supervised Representations of French Speech, September 2023. URL http://arxiv.org/abs/2309.05472. arXiv :2309.05472 [cs, eess].

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise Analysis of a Self-supervised Speech Representation Model, December 2022. URL http://arxiv.org/abs/2107.04734. arXiv :2107.04734 [cs, eess].

Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models, March 2023. URL http://arxiv.org/abs/2211.03929. arXiv :2211.03929 [cs, eess].

Ibrahim Patel and Y Srinivas Rao. Speech Recognition Using HMM with MFCC-An Analysis Using Frequency Specral Decomposion Technique. *Signal & Image Processing : An International Journal*, 1(2) :101–110, December 2010. ISSN 22293922. DOI : 10.5121/sipij.2010.1209. URL http://www.aircconline.com/sipij/V1N2/1210sipij09.pdf.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings, April 2021. URL http://arxiv.org/abs/2104.03502. arXiv :2104.03502 [cs, eess].

Antoine Serrurier. 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. 2006.

Will Styler. On the acoustical features of vowel nasality in English and French. *The Journal of the Acoustical Society of America*, 142(4) :2469–2482, October 2017. ISSN 0001-4966, 1520-8524. DOI : 10.1121/1.5008854. URL https://pubs.aip.org/jasa/article/142/4/2469/853233/On-the-acoustical-features-of-vowel-nasality-in.

Francisco Torreira, Martine Adda-Decker, and Mirjam Ernestus. The Nijmegen Corpus of Casual French. *Speech Communication*, 52(3) :201–212, March 2010. ISSN 01676393. DOI :

10.1016/j.specom.2009.10.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0167639309001629.

Liang-Hsuan Tseng, Yu-Kuan Fu, Heng-Jui Chang, and Hung-yi Lee. Mandarin-english code-switching speech recognition with self-supervised speech representation models. *arXiv preprint arXiv :2110.03504*, 2021.