Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Exploration de la séparation en langues dans les modèles de traitement de la parole auto-supervisés multilingues préentraînés avec des données écologiques

William N. Havard[1, 2] Shrita Hassamal[3] Muhsina Alleesaib[4] Guilhem Florigny[4]
Guillaume Fon Sing[5] Anne Abeillé[5] Benjamin Lecouteux[2] Emmanuel Schang[1]
(1) LLL, Université d'Orléans, CNRS, F-45000 Orléans, France
(2) LIG, Université Grenoble Alpes, CNRS, Grenoble INP, F-38000 Grenoble, France
(3) Mauritius Institute of Education, Maurice
(4) University of Mauritius, Maurice
(5) LLF, Universite Paris-Cité, CNRS, F-75013 Paris, France
william.havard@univ-orleans.fr

RÉSUMÉ

Les modèles auto-supervisés omnilingues de traitement de la parole sont adaptables mais manquent de plausibilité écologique et cognitive. Entraînés sur des corpus monolingues, ils négligent le multilinguisme réel et le code-switching. De précédents travaux suggèrent que de tels modèles procèdent à des regroupements en langues dans l'espace latent, mais cela pourrait être dû à des biais acoustiques ou paralinguistiques plutôt qu'à de véritables traitements linguistiques. Nous avons entraîné un modèle WAV2VEC2 sur des données multilingues de Maurice, incluant des locuteurs plurilingues et du code-switching, et avons étudié les représentations latentes du modèle. Nos analyses montrent que les facteurs acoustiques et paralinguistiques sont encodés sans apprentissage actif, tandis que le regroupement par langue émerge avec un réel apprentissage. Ces résultats éclairent ainsi sur les véritable capacités linguistiques et paralinguistiques des modèles auto-supervisés de la parole.

ABSTRACT

Emergence of Language Separation in Ecological Multilingual Models of Speech Processing

Self-supervised omnilingual models of speech processing are adaptable but lack ecological and cognitive plausibility. Trained on monolingual corpora, they overlook real-world multilingualism and code-switching phenomena. Previous studies suggest these models perform language grouping in their latent space, but this could be due to acoustic or paralinguistic biases rather than a genuine linguistic processing. We trained a WAV2VEC2 model on multilingual data from Mauritius, including plurilingual speakers and utterances with code-switched sections, and analysed its latent representations. Our analyses show that acoustic and paralinguistic factors are encoded without active learning, while language grouping emerges through an active learning process. These results shed light on the true linguistic and paralinguistic capabilities of self-supervised speech models.

MOTS-CLÉS : analyse des représentations latentes, modèles multilingues, traitement de la parole.

KEYWORDS: analysis of latent representations, multilingual models, speech processing.

ARTICLE : **Soumis à** Conference on Computational Natural Language Learning (CoNLL).

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 390–403.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1 Introduction

Les modèles auto-supervisés multilingues de traitement de la parole constituent la base de nombreux modèles état de l'art, après affinage. Communément désignés sous le nom de *modèles de fondation* en raison de leur large applicabilité (voir Yang *et al.*, 2024), ils font preuve d'une remarquable adaptabilité sur une large gamme de tâches et de langues, allant de la reconnaissance automatique de la parole jusqu'à la détection d'émotions. Ces modèles sont appelés *multilingues* car ils sont exposés à plusieurs langues lors du pré-entraînement : par exemple, WHISPER est entraîné sur 96 langues (Radford *et al.*, 2023), XLSR de 53 (Conneau *et al.*, 2021) à 128 (Babu *et al.*, 2022), MHUBERT sur 147 (Zanon Boito *et al.*, 2024) et MMS sur 1406 (Pratap *et al.*, 2024). Cependant, ce type de multilinguisme n'est ni écologique ni cognitivement plausible. Qu'ils soient vus sous le prisme du *plurilinguisme*, défini comme le "*répertoire des variétés de langues que de nombreux* ***individus*** *utilisent*" ou du *multilinguisme*, défini comme "*la présence dans* ***une zone géographique déterminée*** *[...] de plus d'une « variété de langues »*" (nous soulignons, Conseil européen, 2007, p.8), aucun de ces modèles n'est réaliste, puisqu'aucune "*sociétés, institutions, groupes et individus utilisent [...] dans leurs vies quotidiennes*" (Commission Européenne, 2007) l'ensemble spécifique de langues utilisé pour entraîner ces modèles. De plus, en raison du fait qu'elles soient parlées dans la même zone géographique ou par des locuteurs plurilingues, des langues en contact donnent généralement lieu — si ce n'est toujours selon Gardner-Chloros (2020) — à des phénomènes de *code-switching*, *code-mixing* et *fused-lect* (Auer, 1999, cité par Vaillant & Léglise 2014). Aucun des modèles mentionnés ne prend en compte ces phénomènes, car ils sont entraînés sur des corpus *monolingues*, où chaque corpus ne comporte que des énoncés dans une unique langue. Par conséquent, de tels modèles devraient porter un nom moins linguistiquement connoté, tel que *omnilingue*, car ils sont théoriquement entraînables pour traiter n'importe quelle combinaison de langues, sans considération écologique ou cognitive.

De précédents travaux ont étudié de tels modèles *omnilingues*, et montrent notamment que les langues peuvent être identifiées à partir de leurs représentations latentes : ils distinguent une langue d'une autre, et chaque langue apparaît comme un cluster distinct dans l'espace de représentation du modèle (p. ex. Fan *et al.*, 2021, parmi d'autres). Cependant, il n'est pas clairement identifié si ce phénomène apparaît « naturellement », ou s'il s'agit plutôt un artefact résultant d'un biais. Cela pourrait être un artefact **(a)** des corpus d'*entraînement* et/ou de *test* utilisés (p. ex. utilisation de corpus distincts pour chaque langue où aucun locuteur ne parle dans les deux langues cibles, c'est-à-dire sans locuteurs plurilingues, comme dans les travaux de de Seyssel *et al.* 2022; Abdullah *et al.* 2024), ou parce que **(b)** les échantillons d'entraînement comportent des énoncés entièrement monolingues, sans aucune alternance codique (*code-switching*), ce qui peut explicitement signaler au modèle que les énoncés proviennent de distributions différentes, ou **(c)** parce que les langues étudiées sont très distinctes (p. ex. anglais, finnois et allemand pour de Seyssel & Dupoux 2020). De plus, **(d)** de précédents travaux (p. ex. Fily *et al.*, 2024) ont montré que des informations acoustiques de bas niveau comme le microphone, l'acoustique de la salle ou des informations paralinguistiques comme le genre (Guillaume *et al.*, 2024) ou l'identité du locuteur (van Niekerk *et al.*, 2021) sont capturés ces modèles. Par conséquent, le clustering des énoncés par langue pourrait simplement être un sous-produit d'un clustering basé sur ces propriétés, où les modèles reconnaîtraient plutôt le locuteur, le microphone ou le corpus, plutôt que la langue. Enfin, il n'est pas clairement établi que les informations acoustiques et paralinguistiques mentionnées ci-dessus sont bien des propriétés *apprises* par les modèles, ou si elles pourraient émerger à cause d'un biais structural de l'architecture et des caractéristiques des données. [1]

1. Le signal de parole n'est pas un signal neutre, car il contient beaucoup d'informations sur le locuteur (p. ex. le genre et la $f0$ sont liés) et même des projections non linéaires aléatoires sur de telles données non aléatoires pourraient conserver ces

Par conséquent, bien que de précédents travaux aient étudié les représentations de modèles *omnilingues* de traitement de la parole, aucun n'a examiné de véritables modèles *multilingues* (dans l'acception du Conseil européen), qui sont fondés sur le plan écologique et cognitif. Les biais précédemment mentionnés *n'excluent pas* la possibilité que le clustering en langues soit un artefact des paramètres d'entraînement, ou d'un clustering des locuteurs, ou d'un clustering basé sur d'autres facteurs acoustiques et/ou paralinguistiques. Notre travail cherche à combler ce vide en étudiant les représentations d'un modèle WAV2VEC2 entraîné sur des données écologiquement valides, où il n'est pas *explicitement* demandé au modèle de séparer les langues les unes des autres, et où **(a')** le corpus d'entraînement et de test sont disjoints, comportant tous deux des locuteurs plurilingues parlant plusieurs langues, **(b')** parfois avec certaines sections comportant de l'alternance codique, **(c')** avec deux langues phonétiquement, phonologiquement et lexicalement proches. Notre travail se concentre sur les langues utilisées à l'Île Maurice (créole mauricien, français et anglais), dont la situation sociolinguistique est décrite plus en détail dans la section suivante. Nous cherchons également a **(d')** rigoureusement identifier les propriétés qui sont *apprises* de celles qui peuvent être facilement identifiées sans beaucoup d'apprentissage.

Ainsi, notre travail vise à répondre à l'ensemble de ces questions et à explorer les capacités de modélisation linguistiques et paralinguistiques d'un modèle auto-supervisé de traitement de la parole, WAV2VEC2 (Baevski *et al.*, 2020), entraîné sur des données multilingues écologiques.

2 Situation Linguistique de l'Île Maurice

Maurice est un pays insulaire (dont l'île éponyme est l'île principale) situé dans l'Océan Indien, où, pour des raisons historiques complexes, trois langues sont principalement utilisées au quotidien : le créole mauricien (*morisien*, utilisé par l'ensemble de la population), le français (utilisé par une majorité de la population), et l'anglais (surtout d'usage scolaire et administratif).[2] Le morisien s'est développé en tant que langue distincte du français à travers un processus de créolisation, principalement en raison de l'esclavage. Alors qu'il a emprunté beaucoup de son lexique du français — sa langue lexificatrice — il a subi des changements phonologiques significatifs (p. ex. /ʃ/→/s/, /ʒ/→/z/, /y/→/i/, entre autres). Alors que certains mots sont restés pratiquement inchangés (p. ex. /diʁ/∼/diɣ/, *dire*) certains, bien qu'étant proches de leurs homologues français, sont le résultat de l'agglutination du déterminant français (Henri & Bonami, 2018) au nom suivant (p. ex. /la/ et /bɔ̃te/, donnant /labɔ̃te/, *(une) bonté*), ou d'un changement dans le système phonologique (p. ex. /byʁo/→/biɣo/, *bureau*), ou les deux à la fois (p. ex. /la/ et /plaʒ/→/laplaz/, *(une) plage*). Lorsque Maurice est devenu une colonie de l'Empire Britannique, de 1810 à 1968, l'anglais fut introduit sur l'île, et est devenu, et reste encore aujourd'hui, la langue officielle de l'administration et du gouvernement. Ainsi, il existe de nombreux emprunts anglais en morisien (p. ex. enn *viewpoint*, un point de vue). Il est fréquent de trouver de l'alternance codique (voir Exemple 1), avec des mots français (fr) et anglais (en) insérés dans des énoncés plus larges où le morisien (mfe) est la matrice (exemple 1).

(1)

ensuite	peut-être	mo	check	mo	bann	social	media	
ensuite	peut-être	1SG	checker	1SG.PS	PL	social	media	
fr	fr	mfe	en	mfe	mfe	en	en	

Ensuite, peut-être je checkerai mes réseaux sociaux.

(2)

[sa	se	mɒ	pœʁ]
DEM	c'est	1SG.PS	peur

Ça, c'est ma peur.

informations (voir Chrupała & Alishahi, 2019).

2. La situation est même plus complexe encore, car des langues indiennes (hindi, bhojpuri, entre autres) peuvent également être utilisées, bien qu'elles soient en voie de disparition sur l'île.

Le français et le morisien sont si proches, qu'il est parfois difficile, voire impossible de décider si un énoncé est dans l'une ou l'autre langue comme dans l'exemple 2, où l'énoncé sera considéré comme du français si [mɒ] est perçu /ma/ et [ɐ] plus proche de /ʁ/ par le locuteur ; ou comme du morisien si [mɒ] est perçu /mo/ et [ɐ] plus proche de /ɣ/. Ce phénomène n'est pas spécifique au morisien, mais se produit avec de nombreux créoles à base lexicale française. Cela a conduit (Ledegen, 2012) à développer un système de transcription double (appelée transcription flottante) pour les énoncés qui pouvaient être à la fois en créole réunionnais et en français, plus tard également utilisé dans d'autres contextes créolophones, en Guyane notamment (Vaillant & Léglise, 2014).

Maurice offre ainsi un parfait exemple de la manière dont les langues en contact interagissent et donnent naissance à des phénomènes linguistiques complexes (Baggioni & de Robillard, 1990; Ludwig *et al.*, 2009). Par conséquent, un modèle entraîné sur ce type de données constitue un terrain d'expérience idéal pour analyser les modèles *multilingues* de parole écologiquement fondés.

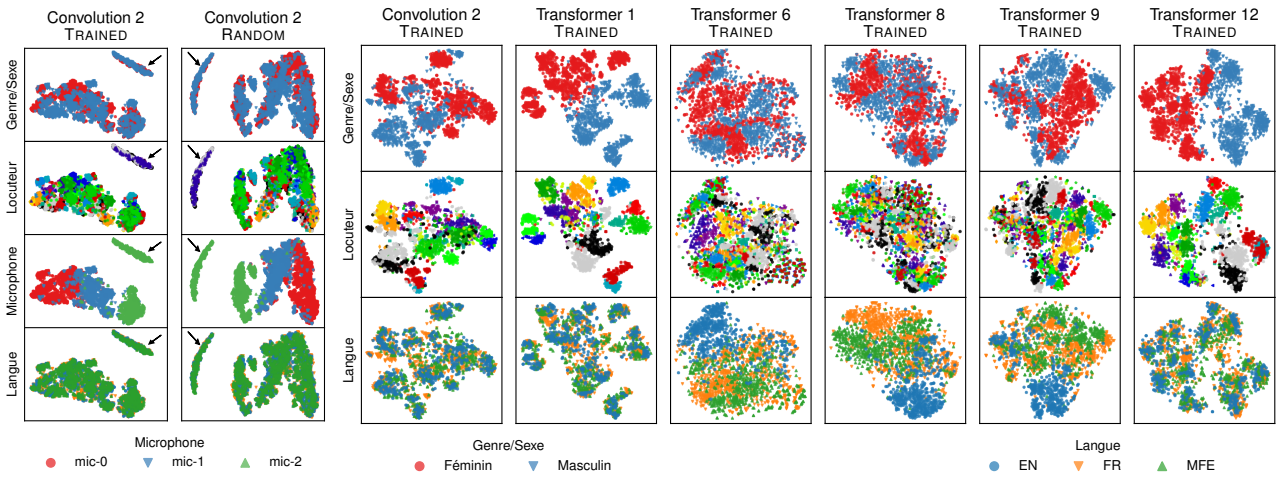
3 Conditions expérimentales

Nous décrivons dans cette section les données que nous utilisons pour entraîner un modèle WAV2VEC2, son protocole d'entraînement, et les données utilisées pour analyser ses représentations latentes.

Données. Les données utilisées dans ces expériences proviennent de deux sources. La première est la *Mauritius Broadcasting Corporation* (MBC), qui diffuse les actualités en morisien tous les jours. Ce type de données est régulièrement utilisé par des linguistes mauriciens et étrangers pour étudier les phénomènes de *code-mixing* (p. ex. Busviah, 2024). Cela nous a donc incités à collecter 485h de données brutes, qui ont abouti à 425h de parole après avoir supprimé le bruit et les sections non-parlées à l'aide d'un modèle de détection de l'activité vocale (Bredin *et al.*, 2020). Les présentateurs, intervieweurs et interviewés parlent principalement en morisien, mais utilisent régulièrement des énoncés monolingues en anglais (p. ex. lorsque le Premier Ministre est interrogé sur les affaires gouvernementales) ou utilisent des sections avec de l'alternance codique lorsqu'ils se réfèrent à des postes ou institutions officielles (p. ex. "*li finn konvoke azordi dan* ***Financial Crimes Commission***", "*elle a été convoquée aujourd'hui à la* ***Commission [d'enquête] des crimes financiers***"). Le français quant à lui est également utilisé pour des énoncés monolingues (p. ex. "*le contrôle [des tickets] doit être fait ici*") ou dans énoncés avec de l'alternance codique (p. ex. "***ce n'est pas la seule chose*** *kot Larabi Saoudit ed nou*", "*ce n'est pas la seule chose pour laquelle l'Arabie Saoudite nous aide*"). [3]

La deuxième source de données se compose d'entretiens que nous avons menés à Maurice. Nous avons interviewé 16 participants, soit en binôme, soit individuellement, et leur avons demandé de parler pendant cinq minutes en morisien, anglais et français. Si les locuteurs n'arrivaient pas à maintenir une conversation pendant cinq minutes, nous leur avons posé des questions sur leurs activités quotidiennes ou leurs projets après l'interview. Cela nous a permis de collecter 2h15 minutes de données brutes, qui ont été réduites à 1h15 minutes après avoir supprimé le bruit, les sections non-parlées et les chevauchements entre locuteurs. Ces données ont ensuite été annotées manuellement pour la langue en utilisant les catégories suivantes : *anglais*, *français* et *morisien* lorsqu'il s'agissait d'un énoncé dans une et une seule de ces langues, ainsi que *français/mauricien* lorsque l'énoncé pouvait être interprété comme appartenant aux deux langues à la fois mais était plus susceptible d'être du français selon le contexte, et *mauricien/français* dans la situation inverse. La distribution des segments est présentée en Table 1. Les exemples 1 & 2 sont des exemples tirés des données que nous avons collectées.

3. https://mbcradio.tv/replay/zournal. Tous les exemples proviennent de "Zournal - février 19, 2025"



(a) (b)

FIGURE 1 – (1a) Visualisation t-SNE des représentations extraites de la deuxième couche de convolutions utilisant le modèle TRAINED (gauche) et le modèle RANDOM (droite) pour tous les microphones, coloriés par microphone, sexe/genre, locuteur et langue, montrant un regroupement par microphone et par condition acoustique (flèches noires). (1b) Visualisation des représentations extraites à différents niveaux du modèle TRAINED, pour le microphone MIC-0 seulement. Un groupement par langue émerge progressivement, atteignant son pic dans les couches du milieu (Transformer 8 et 9) et disparaît ensuite. Pour les Figure 1a et Figure 1b les vecteurs ont été normalisés (cote Z, soit *Z-score*).

Afin de tester si les modèles capturent des informations liées au microphone (Fily *et al.*, 2024), chaque session d'enregistrement avec les participants a été enregistrée avec au moins *deux* et jusqu'à *trois* microphones. [4]

Langue Sexe/Genre	EN	FR	MFE	FR/MFE	MFE/FR	Total
F	304	372	272	37	35	1020
M	226	339	320	27	40	952
Total	530	711	592	64	75	1972

TABLE 1 – Distribution des segments collectés et manuellement annotés, catégorisés par langue (EN : anglais, FR : français, et MFE : morisien, et énoncés bi-langue ambigus) et par sexe/genre (F : féminin, M : masculin).

Modèles TRAINED & RANDOM. Nous pré-entraînons un modèle WAV2VEC2-BASE (Baevski *et al.*, 2020) à l'aide de PyTorch (Paszke *et al.*, 2019) et du toolkit `fairseq` (Ott *et al.*, 2019) sur les données MBC. [5] Le modèle est entraîné à prédire des représentations contextuelles continues de la parole en masquant des segments issus des convolutions et en optimisant une fonction de coût contrastive. Cela encourage les représentations continues des segments masqués cibles à s'aligner étroitement avec les représentations latentes quantifiées de la parole issues de la dernière couche convolutionnelle, tout en distinguant ces représentations de distracteurs échantillonnés à partir de trames voisines.

Les hyperparamètres du modèle sont identiques à ceux de (Evain *et al.*, 2021) avec 7 couches de convolutions de 512 canaux, des chevauchements de $(5, 2, 2, 2, 2, 2, 2)$ et des noyaux de $(10, 3, 3, 3, 3, 2, 2)$, suivies de 12 blocs de Transformers, avec dimension finale de 768, dimension interne de $3,072$, et 12 têtes d'attention. Le module de quantification utilise 2 tables de quantification (*codebooks*) avec un vocabulaire de 320 et une dimension de 128. Au lieu d'entraîner le modèle pour un nombre fixe

4. 1 Zoom H4n et 1 Zoom H2 avec micros intégrés, et 1 Marantz PMD661 MKII avec 2 micro-cravates AKG MPA III.
5. En utilisant respectivement 406h35, 9h06 et 9h06 pour les ensembles d'entraînement, de validation et de test. L'ensemble de test n'est pas utilisé dans les expériences présentées dans ce papier.

d'étapes (par exemple Parcollet *et al.* 2024 entraîne leur modèle 1k-BASE pour 200k étapes), nous avons décidé de stopper manuellement l'entraînement lorsque le modèle avait convergé, défini comme le point où les courbes d'entraînement et de validation se coupent. Nous voulons souligner que notre modèle est pré-entraîné à partir de zéro (c'est-à-dire en partant d'une initialisation aléatoire des poids) plutôt qu'en utilisant un apprentissage par transfert à partir d'un modèle existant (p. ex. en utilisant une approche de pré-entraînement continu). Cela signifie que le modèle n'a été exposé *qu'à* 406 heures de de donnée d'entraînement issues de MBC et à aucun autre contenu. D'un point de vue cognitif, cette approche rend le modèle similaire à un locuteur mauricien qui a principalement été exposé aux trois langues principales utilisées à Maurice. Nous appelons le modèle dont l'architecture et l'entraînement est présenté ci-dessus le modèle TRAINED. Notez que ce modèle n'a pas été affiné pour effectuer une tâche particulière, et a seulement été pré-entraîné sur la tâche standard de WAV2VEC2.

Nous utilisons également dans nos expériences un modèle de contrôle appelé modèle RANDOM. Ses poids sont initialisés aléatoirement et ne sont pas ajustés à travers un processus d'apprentissage : les paramètres du modèle sont fixés une fois pour toutes à des valeurs aléatoires, et il n'a vu aucune des données d'entraînement. Ainsi, l'extraction de vecteurs de caractéristiques (*features*) avec ce modèle ne donnera qu'une projection non linéaire aléatoire des données d'entrée. Nous utilisons ce modèle RANDOM comme référence (*baseline*) pour évaluer l'apprentissage réalisé par le modèle TRAINED.

Données d'évaluation des représentations. L'ensemble d'évaluation utilisé pour analyser les représentations du modèle est uniquement constitué des données que nous avons collectées à Maurice. Par conséquent, l'ensemble de pré-entraînement et l'ensemble d'évaluation sont complètement disjoints, à la fois en termes de contenu et de locuteurs. Chaque énoncé est extrait de l'enregistrement source et est donné aux modèles TRAINED et RANDOM pour extraire des vecteurs. Puisque l'architecture WAV2VEC2 ne produit pas un vecteur unique qui représente l'énoncé entier, nous suivons la méthodologie établie par d'autres avant nous (p. ex. Tjandra *et al.*, 2022; de Seyssel *et al.*, 2022), et utilisons une agrégation par moyenne (*mean pooling*) le long de l'axe temporel pour obtenir un unique vecteur par énoncé. Cette opération est appliquée à la sortie de chaque couche convolutionnelle, de Transformer, ainsi qu'aux vecteurs des tables de quantification, et aux labels des vecteurs des tables de quantification une fois convertis en vecteurs *one-hot*.

4 Analyse

Dans cette section, nous détaillons l'analyse que nous avons menée sur les représentations, en commençant par une exploration visuelle, confirmée ultérieurement à l'aide de classificateurs de sondage/diagnostic (*probing*, voir Hupkes & Zuidema, 2018; Belinkov, 2022).

4.1 Explorations visuelles

Modèles TRAINED et RANDOM. Tout d'abord, nous commençons par examiner les différences entre les vecteurs extraits du modèle TRAINED et du modèle RANDOM (Figure 1a). Pour les deux modèles, les plots t-SNE révèlent un regroupement cohérent des énoncés en fonction du microphone. Non seulement nous observons un regroupement en fonction du microphone mais également des sous-groupes correspondant à différentes conditions acoustiques. Par exemple, le sous-groupe dénoté par une flèche noire correspond à une session d'enregistrement où le gain du microphone a par inadvertance été fixé à une valeur faible. Par conséquent, il semble que les informations acoustiques de bas

niveau concernant le microphone utilisé ou les conditions acoustiques ne sont pas des caractéristiques que le modèle TRAINED *apprend* à identifier. Il semble en effet que l'architecture possède un fort biais inductif qui permet de *préserver* ces informations, sinon, il ne serait pas possible d'observer un schéma de regroupement similaire avec les représentations issues du modèle RANDOM.

Évolution par couches. Nous nous intéressons maintenant à l'évolution des représentations en fonction des couches (Figure 1b) du modèle TRAINED. Nous observons une séparation claire entre les énoncés prononcés par les femmes et les hommes à chaque couche (ligne supérieure), sauf dans les couches du milieu (Transformer 6, 8 et 9). Un schéma similaire est observable en ce qui concerne le regroupement par locuteur (ligne du milieu), avec déjà un groupement très clair dès les premières couches du modèle (Convolution 2). En ce qui concerne le regroupement en fonction de la langue (ligne inférieure) le schéma est à l'opposé des autres variables. Le regroupement par langue apparaît *seulement* dans les couches du milieu, avec une séparation claire entre l'anglais d'une part et le français et le morisien d'autre part pour la couche Transformer 6, et une évolution pour Transformer 8 où le français et le morisien semblent être plus séparés. Ce regroupement par langue semble ensuite se dégrader dans la couche suivante (Transformer 9) et disparaître finalement dans la dernière couche (Transformer 12). Il semble donc qu'il y a une évolution claire dans la nature des représentations du modèle TRAINED, où un regroupement par langue émerge progressivement, culminant dans les couches du milieu et disparaissant ensuite. Cela semble être fait au détriment de la séparation des locuteurs, et de la séparation en fonction du sexe à un moindre degré.

4.2 Sondage des représentations

Il est important de tester les représentations des modèles en utilisant des classificateurs de sondage/diagnostic appropriés, car les regroupements t-SNE pourraient n'être que des artefacts visuels de la projection de représentations de haute dimension vers une dimension plus basse. Nous analysons les représentations du modèle TRAINED et RANDOM en entraînant des classificateurs linéaires pour prédire les propriétés suivantes : microphone, sexe/genre du locuteur, identité du locuteur, et langue. Nous divisons notre jeu d'évaluation à 75/25 pour l'entraînement et le test des classifieurs. Les vecteurs sont normalisés en utilisant la cote Z (*Z-score*, c.-à.-d. à une moyenne de zéro et une variance unitaire), et entraînons des modèles de régression logistique indépendants pour le modèle TRAINED et RANDOM, pour chaque variable et pour chaque couche, avec une régularisation ℓ_1, en utilisant la bibliothèque `scikit-learn` (Pedregosa *et al.*, 2011). Nous entraînons chaque classificateur 10 fois à l'aide de différentes partitions d'entraînement et de test. Pour éviter tout biais, nous équilibrons le nombre d'exemples pour chaque classe. Les résultats sont présentés dans Figure 2a.

Nous rappelons au lecteur que le modèle RANDOM sert de référence, permettant d'évaluer dans quelle mesure les propriétés précédemment mentionnées peuvent être prédites sans ou avec peu apprentissage. Nous nous attendons à ce que plus une propriété est difficile à encoder, plus le modèle RANDOM sera proche de l'exactitude aléatoire théorique, qui représente l'exactitude attendue si les classificateurs choisissaient une étiquette au hasard (soit $\frac{1}{\#\,classes}$). [6]

Enregistreur, Genre & Locuteur. Tout d'abord, nous observons (Figure 2a) que les classificateurs entraînés sur les représentations du modèle TRAINED sont capables de prédire l'information de microphone et de genre avec des exactitudes élevées à toutes les couches (> 0.9). Cependant, nous observons que ces propriétés semblent assez faciles à prédire, car les classificateurs entraînés pour

6. $\frac{1}{3}$ microphone, $\frac{1}{2}$ genres (femme, homme), $\frac{1}{16}$ locuteurs, et $\frac{1}{3}$ langues (morisien, français, anglais).

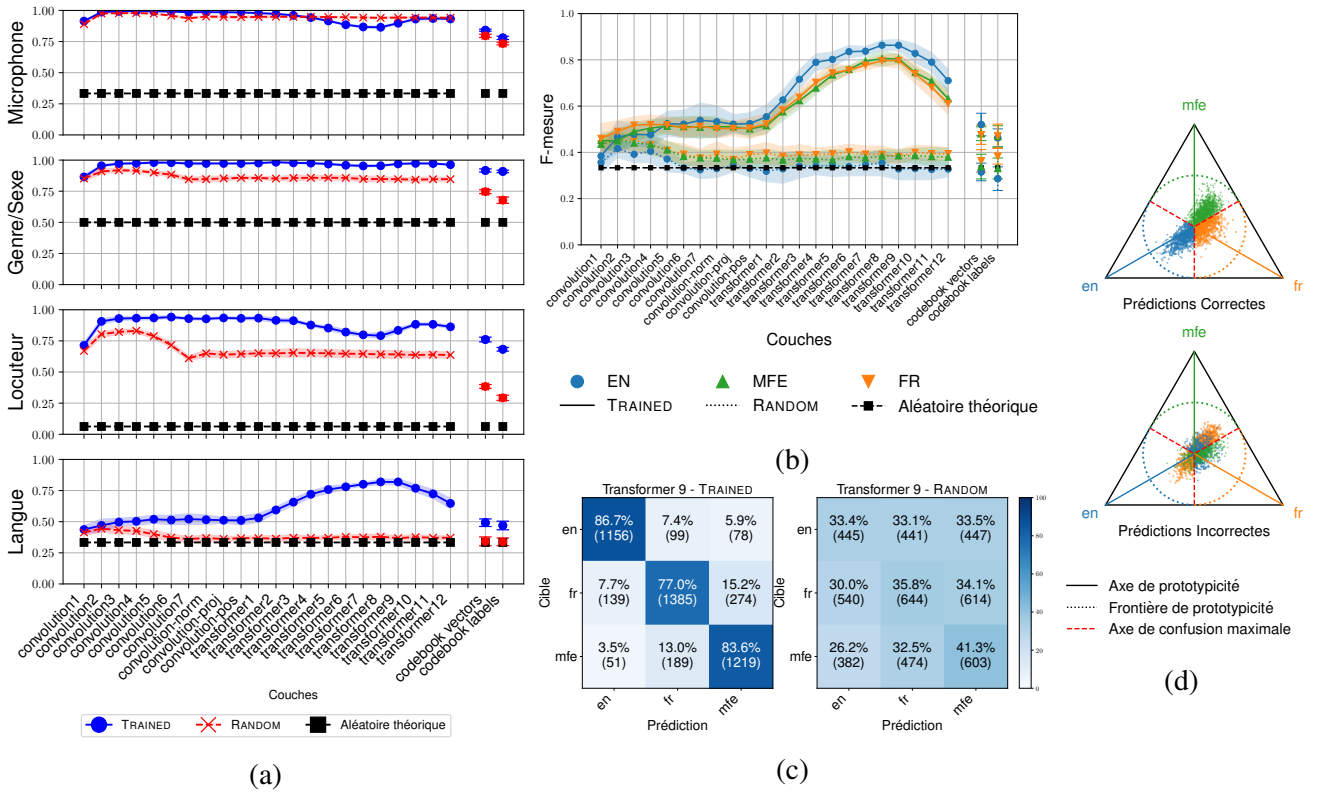


FIGURE 2 – (2a) Exactitude des classificateurs linéaires sur chaque variable pour le modèle **TRAINED**, le modèle **RANDOM** et l'**exactitude aléatoire théorique**, moyennés sur 10 exécutions ($\pm$ déviation standard). *Codebook vectors* et *codebook labels*, correspondant aux vecteurs quantifiés et aux labels des ces vecteurs, ont été tracés séparement, car ils sont différents en nature des autres représentations, qui elles sont continues. (2b) F-mesure par langue pour TRAINED (solide) and RANDOM (pointillés) par couche. (2c) Matrices de confusion agrégées sur 10 exécutions pour la 9^e couche de transformers. (2d) Probabilité d'appartenance d'un énoncé (où chaque point représente un énoncé) à chaque langue (signalée au sommet des simplexes) pour les énoncés correctement (haut) et incorrectement (bas) classifiés. Le centre des simplexes représente l'équiprobabilité d'appartenance à toutes les classes $(0.33, 0.33, 0.33)$, et un point proche d'un sommet indique une probabilité de 1 d'appartenance à cette classe et 0 aux autres.

prédire les mêmes variables en utilisant les représentations du modèle RANDOM obtiennent des exactitudes proches (> 0.8). Nous observons un schéma similaire avec la prédiction de l'identité du locuteur, avec cependant une exactitude plus basse en utilisant les représentations du modèle RANDOM, montrant que cette information est plus difficile à prédire à partir d'une projection aléatoire.

Cependant, pour ces trois variables, l'exactitude aléatoire théorique et l'exactitude empirique observée avec le modèle RANDOM sont très différentes. Il est donc clair que même si nos expériences confirment de précédents résultats concernant le fait que les modèles multilingues encodent l'information de microphone, de genre, ainsi que l'identité du locuteur, [7] nos résultats suggèrent que ces propriétés ne sont pas tant *apprises* par les modèles mais qu'elles émergent plutôt par un biais structurel de l'architecture, autrement nous n'aurions pas pu prédire ces variables à partir d'une projection aléatoire.

7. voir (Fily *et al.*, 2024; Guillaume *et al.*, 2024). Également, nous avons observé que la normalisation cote Z atténue les détails liés au locuteur mais ne concluons pas comme van Niekerk *et al.* (2021) que "la normalisation enlève l'information du locuteur et du genre", car cette information est largement préservée et peux toujours être prédite avec des méthodes linéaires.

Langue. La tendance dessinée par l'exactitude moyenne par couche lors de la prédiction de la langue utilisée par les locuteurs de notre corpus est en contraste avec les variables précédentes. En effet, à l'exception des premières couches de convolutions, l'exactitude aléatoire théorique et l'exactitude aléatoire empirique (modèle RANDOM) sont identiques. À l'inverse, l'exactitude des classificateurs entraînés sur les représentations du modèle TRAINED évoluent à toutes les couches, atteignant son pic aux couches Transformer 8 et 9 (avec une exactitude moyenne de $0,8$). Nous observons ainsi une courbe en cloche typique (Pasad *et al.*, 2021; Abdullah *et al.*, 2024) de la fonction de coût WAV2VEC2 qui rapproche les représentations des couches supérieures aux représentations quantifiées des convolutions. Ces expériences confirment que les modèles auto-supervisés multilingues de parole *apprennent* à différencier les langues. Puisque chaque locuteur dans notre jeu d'évaluation parle dans chacune des langues cibles (morisien, français et anglais), nous pouvons écarter avec certitude la possibilité que l'identification de langue résulte d'un artefact de locuteurs différents utilisant des langues différentes. Nous démontrons ainsi que le modèle réalise un véritable traitement linguistique. C'est la première fois à notre connaissance qu'une telle conclusion a été établie avec un tel protocole.

Enfin, les plots t-SNE (Figure 1b) nous ont permis de faire l'hypothèse que lorsque des groupements en langues apparaissent, les groupements en locuteurs et en genres deviennent moins structurés, ce qui suggère que le regroupement basé sur la langue coïncide avec une perturbation du regroupement pour les locuteurs et, à moindre égard, pour le genre. Cette observation est confirmée par la classification linéaire pour le Transformer 8, car des exactitudes accrues sur la prédiction de la langue coïncident avec la plus basse exactitude moyenne de prédiction pour l'identification du locuteur et du microphone. Cependant, la prédiction du genre n'est pas aussi impactée que la représentation visuelle le suggère. Les représentations du Transformer 9, telles qu'elles apparaissent dans les plots t-SNE et confirmées par les précisions des classifieurs linéaires, semblent maintenir à la fois la séparation en langues, tout en ayant une plus forte séparation en locuteurs que Transformer 8.

Confusions. Nous observons (Figure 2b) que l'anglais est mieux reconnu que le français ou le morisien, et plus tôt dans l'ensemble des couches : la F-mesure pour l'anglais atteint $0,8$ avec les représentations de la 4^e couche, tandis que le français et le morisien n'atteignent un niveau similaire, et leur pic, qu'à la 8^e et 9^e couche Tranformer. Nous interprétons cela comme un signe que l'anglais est plus distinct et, par conséquent, identifiable que les deux autres langues, bien que notre modèle ait été entraîné sur des données principalement en morisien. Les matrices de confusion (Figure 2c) révèlent que lorsque les classifieurs attribuent à tort un segment français à une autre langue, ils le classifient généralement comme du morisien (66% du temps), et inversement, comme français lorsqu'il s'agit de morisien (78% du temps). Le motif pour l'anglais est plus équilibré avec presque la même quantité d'énoncés erronément attribués au français (56%) et au morisien (44%). Nous interprétons ce phénomène comme un signe de la proximité entre le français et le morisien.

Enfin, la représentation des probabilités d'appartenance de chaque énoncé à chaque classe, projetés sur un simplexe (Figure 2d, agrégé sur 10 exécutions),[8] montre la forte entropie des classifieurs, avec beaucoup de points au centre de gravité et le long des médianes (représentant ainsi des axes de confusion), signe d'une grande équiprobabilité d'appartenance d'un énoncé à chaque classe ; une faible entropie aurait été signalée par de nombreux points placés aux sommets. Les énoncés en anglais sont plus alignés avec le sommet correspondant du simplexe, représentant ainsi un axe de prototypicalité, confirmant de nouveau que l'anglais est mieux reconnu par le modèle. Les énoncés en morisien se sont vus affecter une part non négligeable de probabilité d'appartenance au français, s'écartant de leur axe de prototypicalité, signe de la proximité de ces deux langues pour le modèle.

8. La distribution de probabilité sur 3 classes p_1, p_2, p_3 est projetée sur un simplexe $2D$ par $x = p_1 + \frac{p2}{2}, y = \frac{\sqrt{3}}{2} \times p_2$

5 Discussion & Conclusion

Discussion. Nous pouvons donc conclure avec certitude que les modèles *multilingues* auto-supervisés de la parole – tels que WAV2VEC2– comme leurs homologues *omnilingues*, apprennent à identifier les langues à partir de leurs entrées. Bien que qu'une telle conclusion ait également été faite par (Fan *et al.*, 2021; Abdullah *et al.*, 2024), nos résultats montrent que ce phénomène n'est pas à expliquer par **(a')** un artefact des données *d'entraînement*, car cela se produit avec un modèle entraîné sur un ensemble de données multilingues écologiques où des locuteurs plurilingues utilisent plusieurs langues, **(b')** parfois dans le même énoncé avec des sections présentant de l'alternance codique. De plus, en ayant recueilli nous-mêmes un jeu d'évaluation indépendant, où plusieurs locuteurs parlent en trois langues au cours d'un seul entretien, nous pouvons également conclure avec certitude que ce phénomène n'est pas à attribuer aux données de *test*. Enfin, **(c')** nous avons observé un groupement en langues avec des langues qui sont phonétiquement, phonologiquement et lexicalement proches : le morisien et le français. Ce résultat a des implications linguistiques importantes, car la proximité typologique de ces deux langues fait débat, certain soutenant que les créoles sont typologiquement distincts des langues non créoles (Bakker *et al.*, 2011), et d'autres le contraire (Degraff, 2003; Fon Sing, 2017). Nos résultats, bien que limités dans leur portée, ne soutiennent pas fortement la revendication selon laquelle "les créoles se détachent" (Bakker *et al.*, 2011). Si c'était le cas, nous observerions un degré de séparation important entre l'anglais et le français d'un côté, et le morisien de l'autre. Au contraire, le français et le morisien sont plus souvent confondus entre eux. C'est en revanche l'anglais qui se détache dans nos expériences, car il n'est pas seulement reconnu avec une plus grande exactitude, mais il émerge également plus tôt dans les couches du modèle. Malgré leur apparente proximité, le morisien et le français apparaissent comme des langues distinctes dans les représentations du modèle entraîné, ce qui suggère que celui-ci est capable de capturer suffisamment de différences linguistiques entre les deux pour les distinguer. À la différence des grands modèles *omnilingues* entraînés sur des milliers d'heures de données, notre modèle *multilingue* a été entraîné sur seulement 400 heures, ce qui suggère que la séparation en langues est une propriété fondamentale qui émerge même avec un faible volume de données et qu'elle est donc essentielle au modèle pour exécuter sa tâche. Concernant **(d')** les capacités de modélisation acoustique et paralinguistique des modèles auto-supervisés, notre comparaison entre les modèles entraîné et aléatoire suggère que les informations acoustiques et paralinguistiques de bas niveau, telles que le type de microphone, le locuteur ou le genre, ne sont pas activement *apprises* mais plutôt *retenues* par le modèle. Il semble important en effet de garder trace de la source et du canal utilisés pour transmettre un message pour le décoder plus précisément, notamment lors d'une transmission bruitée (effet de *cocktail-party* Cherry 1953).

Conclusion. Dans ce travail, nous avons entraîné un modèle *multilingue* de traitement de la parole, à partir de zéro, basé sur l'architecture WAV2VEC2, sur des données écologiques, pour traiter les trois principales langues utilisées à Maurice : le morisien, le français et l'anglais. Nous avons analysé les représentations apprises par le modèle sur des données que nous avons collectées sur place, à Maurice, afin de comprendre si celui-ci capture des informations linguistiques (langue), acoustiques (microphone) et paralinguistiques (locuteur, genre) ; et comment celles-ci diffèrent d'un modèle aléatoire. Nos résultats révèlent que les informations acoustiques et paralinguistiques sont encodées dans les représentations, mais avec un apprentissage dédié minimal. La séparation en langues apparaît, elle, par un apprentissage, ce qui met en évidence sa nature distincte dans le traitement du modèle.

Les futures directions de recherche incluent notamment l'exploration des distances entre langues dans l'espace de représentation latent pour identifier les points de similarité et de divergence linguistique, en particulier entre une langue créole et sa langue lexificatrice, ici le morisien et le français.

Remerciements

Cette recherche a été financée par l'Agence nationale de la recherche (ANR) au titre du projet ANR-20-CE38-0006 (projet CREAM). Les expériences ont été menées à l'aide de Grid'5000, développé sous INRIA ALADDIN avec l'appui du CNRS, RENATER et diverses universités ; de CaSciModOT au Centre de Calcul Scientifique en région Centre-Val de Loire ; et les ressources HPC de IDRIS fournies par GENCI (allocation 2024-AD011014940). Enfin, la collecte des données a été effectuée dans le cadre des Partenariats Hubert Curien (PHC) Le Réduit (en 2023 *via* le projet SUMAU, et en 2024 *via* le projet DOUCE) avec l'appui du MEAE et du MESR en France ainsi que l'appui du Ministère de l'Éducation, de l'Enseignement supérieur, de la Recherche scientifique et de la Technologie à Maurice.

Références

ABDULLAH B. M., SHAIK M. M. & KLAKOW D. (2024). Wave to interlingua : Analyzing representations of multilingual speech transformers for spoken language translation. In *Interspeech 2024*, p. 362–366. DOI : 10.21437/Interspeech.2024-2109.

AUER P. (1999). From codeswitching via language mixing to fused lects : Toward a dynamic typology of bilingual speech. *International Journal of Bilingualism*, **3**(4), 309–332. DOI : 10.1177/13670069990030040101.

BABU A., WANG C., TJANDRA A., LAKHOTIA K., XU Q., GOYAL N., SINGH K., VON PLATEN P., SARAF Y., PINO J., BAEVSKI A., CONNEAU A. & AULI M. (2022). XLS-R : Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, p. 2278–2282. DOI : 10.21437/Interspeech.2022-143.

BAEVSKI A., ZHOU H., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA : Curran Associates Inc.

BAGGIONI D. & DE ROBILLARD D. (1990). *Île Maurice : une francophonie paradoxale*. Editions L'Harmattan.

BAKKER P., DAVAL-MARKUSSEN A., PARKVALL M. & PLAG I. (2011). Creoles are typologically distinct from non-creoles. *Journal of Pidgin and Creole Languages*, **26**(1), 5–42. DOI : 10.1075/jpcl.26.1.02bak.

BELINKOV Y. (2022). Probing classifiers : Promises, shortcomings, and advances. *Computational Linguistics*, **48**(1), 207–219. DOI : 10.1162/coli_a_00422.

BREDIN H., YIN R., CORIA J. M., GELLY G., KORSHUNOV P., LAVECHIN M., FUSTES D., TITEUX H., BOUAZIZ W. & GILL M.-P. (2020). pyannote.audio : neural building blocks for speaker diarization. In *ICASSP 2020*, Barcelona, Spain.

BUSVIAH N. (2024). Étude qualitative des « mélanges de langues » dans les structures prédicatives du discours mauricien. *SHS Web of Conferences*, **191**, 02006. DOI : 10.1051/shsconf/202419102006.

CHERRY E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, **25**(5), 975–979. DOI : 10.1121/1.1907229.

CHRUPAŁA G. & ALISHAHI A. (2019). Correlating neural and symbolic representations of language. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting*

of the Association for Computational Linguistics, p. 2952–2962, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1283.

COMMISSION EUROPÉENNE (2007). *High level group on multilingualism – Final report*. Publications Office of the European Union.

CONNEAU A., BAEVSKI A., COLLOBERT R., MOHAMED A. & AULI M. (2021). Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*, p. 2426–2430. DOI : 10.21437/Interspeech.2021-329.

CONSEIL EUROPÉEN (2007). From linguistic diversity to plurilingual education : Guide for the development of language education policies in europe (executive version). (visité le 18 mars 2025).

DE SEYSSEL M. & DUPOUX E. (2020). Does bilingual input hurt ? A simulation of language discrimination and clustering using i-vectors. In *CogSci 2020 - 42nd Annual Virtual Meeting of the Cognitive Science Society*, Toronto / Virtual, Canada. HAL : hal-02959451.

DE SEYSSEL M., LAVECHIN M., ADI Y., DUPOUX E. & WISNIEWSKI G. (2022). Probing phoneme, language and speaker information in unsupervised speech representations. In *Interspeech 2022*, p. 1402–1406. DOI : 10.21437/Interspeech.2022-373.

DEGRAFF M. (2003). Against creole exceptionalism. *Language*, **79**(2), 391–410.

EVAIN S., NGUYEN M. H., LE H., ZANON BOITO M., MDHAFFAR S., ALISAMIR S., TONG Z., TOMASHENKO N., DINARELLI M., PARCOLLET T., ALLAUZEN A., ESTÈVE Y., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2021). Task Agnostic and Task Specific Self-Supervised Learning from Speech with LeBenchmark. In *Thirty-fifth Conference on NeurIPS (NeurIPS 2021)*, NeurIPS 2021 Datasets and Benchmarks Track, on-line, United States. HAL : hal-03407172.

FAN Z., LI M., ZHOU S. & XU B. (2021). Exploring wav2vec 2.0 on speaker verification and language identification. In *Interspeech 2021*, p. 1509–1513. DOI : 10.21437/Interspeech.2021-1280.

FILY M., WISNIEWSKI G., GUILLAUME S., ADDA G. & MICHAUD A. (2024). Establishing degrees of closeness between audio recordings along different dimensions using large-scale cross-lingual models. In Y. GRAHAM & M. PURVER, Éds., *Findings of the Association for Computational Linguistics : EACL 2024*, p. 2332–2341, St. Julian's, Malta : Association for Computational Linguistics.

FON SING G. (2017). Creoles are not typologically distinct from non-creoles. *Language Ecology*, **1**(1), 44–74. DOI : 10.1075/le.1.1.04fon.

GARDNER-CHLOROS P. (2020). *Contact and Code-Switching*, In *The Handbook of Language Contact*, chapitre 9, p. 181–199. John Wiley & Sons, Ltd. DOI : https ://doi.org/10.1002/9781119485094.ch9.

GUILLAUME S., FILY M., MICHAUD A. & WISNIEWSKI G. (2024). Gender and language identification in multilingual models of speech : Exploring the genericity and robustness of speech representations. In *Interspeech 2024*, p. 3330–3334. DOI : 10.21437/Interspeech.2024-953.

HENRI F. & BONAMI O. (2018). Prédire l'agglutination de l'article en mauricien. *Faits de langue*, **49**, 113–138.

HUPKES D. & ZUIDEMA W. (2018). Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure (extended abstract). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, p. 5617–5621 : International Joint Conferences on Artificial Intelligence Organization. DOI : 10.24963/ijcai.2018/796.

LEDEGEN G. (2012). Prédicats "flottants" entre le créole acrolectal et le français à La Réunion : exploration d'une zone ambiguë. In C. C. ET GOURY L., Éd., *Changement linguistique et langues*

en contact : approches plurielles du domaine prédicatif, p. 251–270. CNRS Éditions. HAL : hal-00905504.

LUDWIG, RALPH, HENRI, FABIOLA & BRUNEAU-LUDWIG F. (2009). Hybridation linguistique et fonctions sociales : Aspects des contacts entre créole, français et anglais à maurice.

OTT M., EDUNOV S., BAEVSKI A., FAN A., GROSS S., NG N., GRANGIER D. & AULI M. (2019). fairseq : A fast, extensible toolkit for sequence modeling. p. 48–53, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-4009.

PARCOLLET T., NGUYEN H., EVAIN S., ZANON BOITO M., PUPIER A., MDHAFFAR S., LE H., ALISAMIR S., TOMASHENKO N., DINARELLI M., ZHANG S., ALLAUZEN A., COAVOUX M., ESTÈVE Y., ROUVIER M., GOULIAN J., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2024). Lebenchmark 2.0 : A standardized, replicable and enhanced framework for self-supervised representations of french speech. *Computer Speech & Language*, **86**, 101622. DOI : https ://doi.org/10.1016/j.csl.2024.101622.

PASAD A., CHOU J.-C. & LIVESCU K. (2021). Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, p. 914–921. DOI : 10.1109/ASRU51503.2021.9688093.

PASZKE A., GROSS S., MASSA F., LERER A., BRADBURY J., CHANAN G., KILLEEN T., LIN Z., GIMELSHEIN N., ANTIGA L., DESMAISON A., KÖPF A., YANG E., DEVITO Z., RAISON M., TEJANI A., CHILAMKURTHY S., STEINER B., FANG L., BAI J. & CHINTALA S. (2019). *PyTorch : an imperative style, high-performance deep learning library*, In *NeurIPS*. Curran Associates Inc. : Red Hook, NY, USA.

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. (2011). Scikit-learn : Machine learning in python. *J. Mach. Learn. Res.*, **12**(null), 2825–2830.

PRATAP V., TJANDRA A., SHI B., TOMASELLO P., BABU A., KUNDU S., ELKAHKY A., NI Z., VYAS A., FAZEL-ZARANDI M., BAEVSKI A., ADI Y., ZHANG X., HSU W.-N., CONNEAU A. & AULI M. (2024). Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, **25**(97), 1–52.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23 : JMLR.org.

TJANDRA A., CHOUDHURY D. G., ZHANG F., SINGH K., CONNEAU A., BAEVSKI A., SELA A., SARAF Y. & AULI M. (2022). Improved language identification through cross-lingual self-supervised learning. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6877–6881. DOI : 10.1109/ICASSP43922.2022.9747667.

VAILLANT P. & LÉGLISE I. (2014). À la croisée des langues. Annotation et fouille de corpus plurilingues. *Revue des Nouvelles Technologies de l'Information*, **RNTI-SHS-2**, 81–100. HAL : halshs-01063067.

VAN NIEKERK B., NORTJE L., BAAS M. & KAMPER H. (2021). Analyzing speaker information in self-supervised models to improve zero-resource speech processing. In *Interspeech 2021*, p. 1554–1558. DOI : 10.21437/Interspeech.2021-1182.

YANG S.-W., CHANG H.-J., HUANG Z., LIU A. T., LAI C.-I., WU H., SHI J., CHANG X., TSAI H.-S., HUANG W.-C., FENG T.-H., CHI P.-H., LIN Y. Y., CHUANG Y.-S., HUANG T.-H., TSENG W.-C., LAKHOTIA K., LI S.-W., MOHAMED A., WATANABE S. & LEE H.-Y. (2024). A

large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **32**, 2884–2899. DOI : 10.1109/TASLP.2024.3389631.

ZANON BOITO M., IYER V., LAGOS N., BESACIER L. & CALAPODESCU I. (2024). mhubert-147 : A compact multilingual hubert model. In *Interspeech 2024*, p. 3939–3943. DOI : 10.21437/Interspeech.2024-938.