Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

L'accent comme structure géométrique persistante dans les représentations de la parole

Noureddine Khaous[1] Guillaume Wisniewski[1]
(1) LLF, CNRS, Université Paris-Cité, F-75013 Paris, France
noureddine.khaous@u-pariscite.fr, guillaume.wisniewski@u-paris.fr

RÉSUMÉ

Les systèmes modernes de reconnaissance automatique de la parole atteignent aujourd'hui une précision proche du niveau humain pour la parole native, mais demeurent nettement moins robustes face aux accents non natifs. Dans cet article, nous examinons si cette limitation reflète une invariance linguistique incomplète dans les représentations de la parole apprises par les modèles récents. À partir du corpus Speech Accent Archive, nous réalisons une analyse géométrique couche par couche des représentations produites par `wav2vec2` et `whisper`. Nous montrons que la parole non native reste systématiquement séparée de la parole native dans l'espace de représentation tout au long du réseau. De plus, ce déplacement géométrique prédit fortement la dégradation des performances de reconnaissance. Ces résultats suggèrent que les représentations apprises restent fortement influencées par l'accent et n'atteignent pas une invariance complète vis-à-vis des variations phonologiques.

ABSTRACT

Non-Native Accents Persist in Speech Model Representations

Modern ASR systems now achieve near-human performance on native speech, yet remain substantially less robust to non-native accents. In this paper, we ask whether this limitation reflects incomplete linguistic invariance in the representations learned by recent speech models. Using the Speech Accent Archive, we perform a layer-wise geometric analysis of the representations produced by `wav2vec2` and `whisper`. We show that non-native speech remains systematically separated from native speech across layers. Moreover, the magnitude of this displacement strongly predicts degradation in recognition performance. These results suggest that accent is not fully abstracted away in learned speech representations, which remain sensitive to phonological variation across speakers.

MOTS-CLÉS : reconnaissance automatique de la parole, accents non natifs, géométrie des représentations.

KEYWORDS: speech recognition, linguistic invariance, layer-wise analysis, accented speech.

1 Introduction

Des systèmes de reconnaissance automatique de la parole (RAP) tels que `wav2vec2` (Baevski *et al.*, 2020) et `HuBERT` (Hsu *et al.*, 2021), entraînés par apprentissage auto-supervisé, ainsi que `Whisper` (Radford *et al.*, 2023), entraîné de manière supervisée de bout en bout, atteignent désormais des performances proches de celles de l'humain pour de l'anglais lu ou conversationnel dans des conditions favorables. Ces résultats soulèvent une question fondamentale : les performances des systèmes de RAP de l'état de l'art reflètent-elles l'apprentissage de représentations linguistiques

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 322–333.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

abstraites de la parole, ou résultent-elles principalement du fait que des corpus d'entraînement massifs couvrent implicitement la plupart des variations observées lorsque ces modèles sont utilisés en conditions réelles ?

Pour répondre à cette question, nous examinons un phénomène bien identifié de variabilité de la parole : la robustesse des systèmes RAP à l'accent, et plus précisément à la parole non native (L2). La parole non native entraîne souvent une dégradation importante des performances : le taux d'erreur sur les mots (WER) varie sensiblement selon la langue maternelle (L1) des locuteur·rice·s. Ces effets ont été documentés à plusieurs reprises, par exemple par (Graham *et al.*, 2024; Yong *et al.*, 2025), et demeurent marqués même lorsque les données d'entraînement comprennent une quantité importante de parole accentuée. Les accents L2 constituent ainsi une sonde naturelle pour évaluer la capacité des modèles RAP à généraliser au-delà des variations phonétiques de surface : si les représentations apprises capturent réellement le contenu linguistique de manière abstraite, les différences d'accent ne devraient pas entraîner de dégradation systématique des performances. Nous considérons toutefois l'accent comme un continuum plutôt que comme une opposition binaire entre parole native et non native : les données étudiées incluent plusieurs variétés natives de l'anglais ainsi qu'un ensemble diversifié de groupes L1 non natifs, permettant d'observer différents degrés de variation acoustico-phonologique.

Dans ce travail, nous commençons par montrer que cette dégradation est à la fois importante et systématique. En nous concentrant sur l'anglais, nous constatons que les performances RAP varient considérablement selon les groupes L1, et que cette variabilité est remarquablement stable d'une architecture à l'autre et d'un paradigme d'entraînement à l'autre : des modèles entraînés avec des objectifs radicalement différents (à savoir des encodeurs auto-supervisés et systèmes séquence-à-séquence appris de manière entièrement supervisée) présentent en effet des profils de dégradation très similaires selon les L1 des locuteur·rice·s. Ce résultat suggère une limitation qui ne semble pas propre à une architecture ou à une fonction de coût particulière.

Pour comprendre l'origine de ce comportement, nous analysons les représentations internes apprises par ces modèles. Nous conduisons une analyse géométrique couche par couche afin d'évaluer si les représentations éliminent progressivement la variabilité liée à l'accent tout en préservant le contenu linguistique. Si une telle invariance était atteinte, les représentations de productions natives et non natives d'une même unité linguistique devraient converger dans l'espace des représentations à mesure que la profondeur du réseau augmente.

Notre analyse montre que ce n'est pas le cas. La variation liée à l'accent reste fortement encodée tout au long du réseau, induisant des décalages géométriques systématiques qui suivent de près les taux d'erreur en reconnaissance. Ces effets apparaissent de manière constante dans les deux familles de modèles que nous considérons, révélant ainsi une propriété structurelle partagée par les encodeurs de parole modernes.

Pris ensemble, ces résultats suggèrent que les bonnes performances des systèmes de RAP actuels n'impliquent pas l'émergence de représentations invariantes à l'accent. Au contraire, l'accent forme une structure géométrique persistante dans l'espace des représentations, au sein de laquelle contenu linguistique et patrons phonologiques propres au locuteur restent étroitement entremêlés, ce qui limite la robustesse face à la parole non native.

2 Protocole expérimental

Données Dans toutes nos expériences, nous utilisons le corpus *Speech Accent Archive* (SAA) (Weinberger, 2015), qui contient des enregistrements de parole lue produits par des locuteur·rice·s natif·ve·s (anglais L1) et non natif·ve·s (L2), tou·te·s prononçant le même paragraphe d'élicitation. Nous ne retenons que les langues pour lesquelles au moins 20 locuteur·rice·s distinct·e·s sont disponibles, ce qui donne quatre variétés natives de l'anglais (579 locuteur·rice·s, $\sim$5 heures) et 21 langues non natives ($>$7 heures au total). [1] Le principal intérêt du corpus SAA est que le contenu linguistique est fixé : tous les locuteur·rice·s lisent le même paragraphe. Cela permet de contrôler la variabilité lexicale et d'attribuer les différences entre représentations principalement à l'accent.

Prétraitement et alignement Tous les enregistrements audio sont rééchantillonnés à 16 kHz et normalisés en amplitude crête afin d'assurer des conditions d'entrée homogènes entre locuteur·rice·s. La segmentation au niveau du mot est obtenue à l'aide de MFA (McAuliffe *et al.*, 2017) et d'un modèle acoustique de l'anglais, qui permet d'estimer les frontières de mots.

À partir de ces frontières, chaque occurrence de mot est représentée par un vecteur de dimension fixe. Ce vecteur est obtenu en extrayant les états cachés au niveau des trames depuis l'encodeur, puis en appliquant un moyennage temporel au sein d'une couche donnée. Cette procédure produit ainsi un vecteur par mot et par couche, qui sert de base à nos analyses couche par couche.

Le moyennage est effectué à l'échelle du mot, c'est-à-dire sur un nombre relativement restreint de trames. Cela limite le risque d'effacer l'information temporelle fine, risque qui serait nettement plus élevé si l'agrégation portait sur des segments plus longs, tels que des syntagmes ou des énoncés entiers. Le moyennage à l'échelle du mot constitue donc un compromis raisonnable entre précision temporelle et stabilité des représentations. [2]

Modèles RAP et méthode d'évaluation Pour évaluer la robustesse des systèmes RAP à la parole accentuée, nous testons plusieurs modèles largement utilisés sur le corpus SAA, couvrant différents paradigmes d'apprentissage. Nous comparons notamment des systèmes fondés sur des encodeurs de parole auto-supervisés, affinés pour la transcription, à un système RAP encodeur–décodeur entraîné de bout en bout sur des données étiquetées.

Nous considérons d'abord `wav2vec2` (Baevski *et al.*, 2020), une famille de modèles auto-supervisés apprenant des représentations de parole à partir d'une tâche contrastive sur des états latents masqués, et évaluons plusieurs variantes librement disponibles : deux modèles monolingues pour l'anglais, `base` et `large`, ainsi qu'un modèle multilingue `xlsr-53` (Conneau *et al.*, 2021). Nous incluons également `HuBERT` (Hsu *et al.*, 2021), un autre modèle auto-supervisé de représentation de la parole reposant sur un clustering itératif et une prédiction masquée, en utilisant la variante `large`. Dans nos expériences, les checkpoints `wav2vec2`, `xlsr-53` et `HuBERT` utilisés pour le calcul du WER sont des versions déjà affinées pour la transcription de l'anglais, munies d'une tête CTC (*Connectionist*

1. Le nombre de locuteur·rice·s par L2 varie de 20 à 162 (médiane 35), avec une durée totale par langue allant de 10 minutes à 1,3 heure.

2. En principe, des comparaisons par alignement telles que le DTW pourraient être appliquées directement aux représentations au niveau des trames, sans moyennage temporel. Ces approches sont toutefois nettement plus coûteuses en calcul. Nos analyses préliminaires indiquent que le moyennage à l'échelle du mot suffit à capturer les effets liés à l'accent étudiés ici, rendant inutile le recours à des méthodes d'alignement plus coûteuses.

Temporal Classification) (Graves *et al.*, 2006). Les représentations analysées sont extraites avant cette tête de classification, à partir des couches de l'encodeur.

À l'inverse, nous évaluons `Whisper` (Radford *et al.*, 2023), un système RAP encodeur–décodeur entraîné de manière supervisée sur des données multilingues étiquetées à grande échelle. Nous considérons les variantes `base` et `large`, avec leur configuration de décodage par défaut, représentant des modèles dont les représentations internes sont directement optimisées pour la précision de transcription. Pour Whisper, les transcriptions sont produites par le modèle complet, tandis que les analyses géométriques portent sur les représentations de l'encodeur acoustique.

Pour chaque enregistrement, nous générons des transcriptions automatiques avec la procédure d'inférence associée à chaque checkpoint et calculons le taux d'erreur sur les mots (*word error rate*, WER) par rapport à la transcription de référence. Ce taux d'erreur est analysé en fonction de la langue maternelle du locuteur·trice et rapporté à la fois par langue et agrégé sur les groupes de locuteur·rice·s natif·ve·s et non natif·ve·s. En parallèle, nous extrayons les représentations cachées au niveau des trames pour toutes les couches de l'encodeur. Les checkpoints exacts et les détails d'inférence sont rapportés en Annexe A.

Géométrie des représentations : distance au natif Notre objectif est de comparer quantitativement les représentations de la parole produites par des locuteur·rice·s natif·ve·s et non natif·ve·s afin de caractériser les décalages géométriques induits par l'accent dans l'espace des représentations. Une comparaison directe de toutes les réalisations natives et non natives serait coûteuse en calcul. Nous adoptons donc une stratégie fondée sur une représentation de référence, qui fournit un point de comparaison stable et non biaisé.

En pratique, pour chaque couche ℓ et chaque mot w, nous construisons une représentation de référence native définie comme le centroïde des représentations des locuteur·rice·s natif·ve·s : $\boldsymbol{\mu}_w^{(\ell)} = \frac{1}{|\mathcal{N}_w|} \sum_{s \in \mathcal{N}_w} \mathbf{h}_w^{(\ell)}(s)$, où $\mathcal{N}_w$ désigne l'ensemble des locuteur·rice·s natif·ve·s américain·e·s pour lesquels un alignement valide est disponible pour le mot w. Le centroïde obtenu fournit un point d'ancrage opérationnel pour comparer les réalisations d'un même mot à la couche ℓ. Nous utilisons les locuteur·rice·s américain·e·s comme référence car les systèmes RAP considérés obtiennent leurs meilleures performances sur cette variété. Ce choix reflète donc la distribution à laquelle les modèles semblent le mieux adaptés, vraisemblablement en lien avec la composition de leurs données d'entraînement. Les distances rapportées doivent être interprétées relativement à cette référence américaine spécifique, plutôt que comme une mesure absolue de proximité à une parole native idéalisée.

Pour chaque locuteur·rice s, natif·ve ou non natif·ve, nous calculons ensuite une mesure de distance au natif en moyennant les distances cosinus entre les représentations du locuteur au niveau du mot et les centroïdes natifs correspondants :

$$d^{(\ell)}(s) = \frac{1}{|\mathcal{W}_s|} \sum_{w \in \mathcal{W}_s} \left(1 - \cos\left(\mathbf{h}_w^{(\ell)}(s), \boldsymbol{\mu}_w^{(\ell)}\right)\right), \tag{1}$$

où $\mathcal{W}_s$ est l'ensemble des mots produits par le locuteur·rice s. Cette formulation réduit fortement le coût de calcul par rapport aux comparaisons par paires, tout en évitant les biais qui résulteraient du choix d'un·e locuteur·rice natif·ve particulier·ère comme référence. Pour les locuteur·rice·s natif·ve·s, les centroïdes sont calculés en leave-one-out afin d'éviter les effets triviaux d'auto-correspondance.

Enfin, toutes les représentations sont normalisées avant le calcul des distances afin d'atténuer les effets d'anisotropie, une propriété bien documentée des représentations profondes de la parole et

Modèle	US nat.	Autre nat.	Non nat.	deu	fra	spa	cmn	ara	vie
`wav2vec2`									
base	7,6	11,6	25,7	15,8	23,5	26,5	32,5	26,7	42,6
large	5,1	7,9	19,9	12,2	17,5	20,8	25,8	21,8	36,8
xlsr-53	4,8	5,3	17,6	10,9	16,3	17,2	26,1	20,4	32,8
`HuBERT`									
large	2,6	4,4	13,2	6,8	10,0	14,5	17,9	15,4	26,6
`Whisper`									
base	2,9	11,4	32,1	5,1	46,1	32,3	22,6	32,3	48,4
large	1,6	3,9	18,9	2,0	8,5	17,6	20,4	35,4	28,9

TABLE 1 – Taux d'erreur moyen (%) par modèle RAP pour les locuteur·rice·s natif·ve·s et non natif·ve·s de l'anglais, avec décomposition par langue maternelle (L1). Les L1 reportées ont été sélectionnées pour illustrer des niveaux représentatifs de dégradation induite par l'accent.

du langage. Dans ce phénomène, les plongements tendent à se concentrer dans un cône étroit de l'espace vectoriel plutôt que d'y être uniformément distribués. Cette concentration déforme fortement les mesures de similarité, telles que la distance cosinus, en les rendant moins discriminantes et moins interprétables (Ethayarajh, 2019).

Pour y remédier, nous appliquons une procédure de normalisation standard consistant en une suppression de la moyenne couche par couche suivie d'une normalisation ℓ_2. Ce post-traitement améliore substantiellement les propriétés géométriques des espaces de représentation et la fiabilité des analyses fondées sur la similarité. Il est couramment adopté dans les études portant sur les représentations contextualisées (Zaiem *et al.*, 2023; Mu & Viswanath, 2018; Gao *et al.*, 2019).

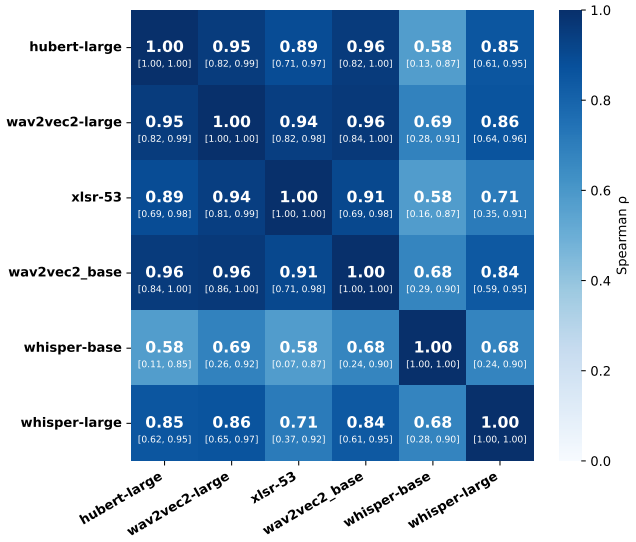

FIGURE 1 – Corrélation de Spearman inter-modèles (avec intervalles de confiance bootstrap à 95 %) des scores WER par locuteur·trice.

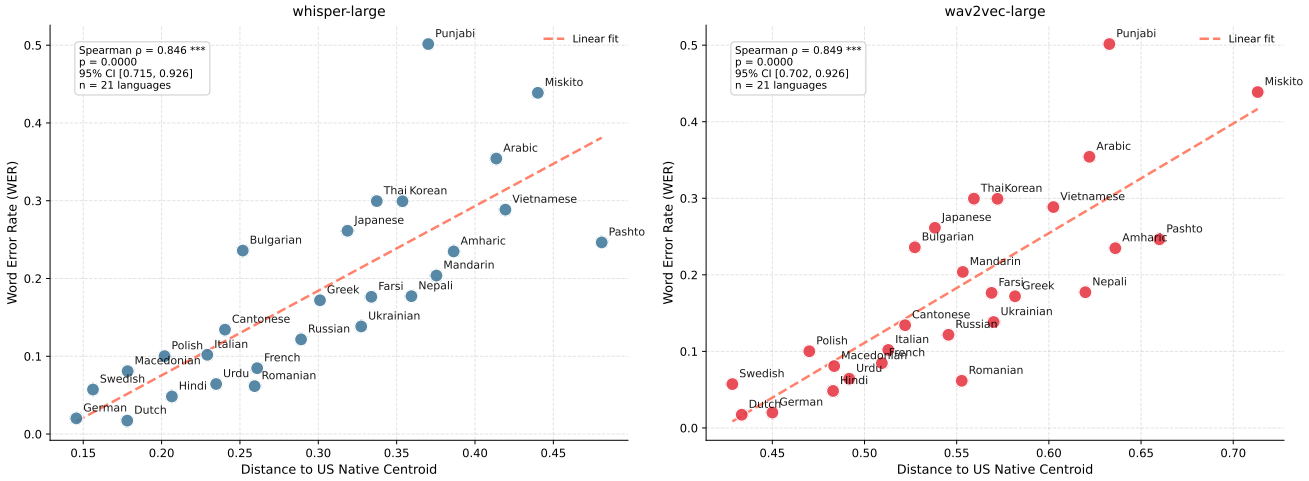

FIGURE 2 – Corrélation entre le déplacement représentationnel induit par l'accent et les taux d'erreur.

3 Résultats et discussion

Performances RAP sur la parole accentuée Le tableau 1 rapporte le taux d'erreur par mot moyen pour la parole native et L2, avec une décomposition selon six L1 choisies pour représenter des niveaux de dégradation faible, intermédiaire et sévère. Sur l'ensemble des modèles, le taux d'erreur augmente à mesure que la parole s'éloigne de la référence native américaine. Les locuteur·rice·s natif·ve·s d'autres variétés de l'anglais présentent déjà des taux d'erreur légèrement supérieurs à ceux des locuteur·rice·s américain·e·s, ce qui indique que la variabilité accentuelle au sein de l'anglais natif introduit déjà une difficulté mesurable pour les systèmes RAP. La dégradation devient toutefois nettement plus marquée pour la parole L2. Le rapport WER L1–L2 varie de $\times 1,8$ (`xlsr-53`) à $\times 5,4$ (`whisper-base`), la plupart des modèles affichant une augmentation d'environ un facteur trois. Bien que les modèles possédant plus de paramètres réduisent substantiellement les taux d'erreur absolus, l'écart de performances persiste.

La décomposition par L1 révèle en outre une hiérarchie cohérente entre les modèles. Les locuteur·rice·s germanophones correspondent au régime de dégradation le plus faible, tandis que les locuteur·rice·s arabophones, vietnamien·ne·s et mandarin produisent systématiquement les taux d'erreur les plus élevés. Le français et l'espagnol occupent un régime intermédiaire.

Pour déterminer si cette dégradation est propre à un modèle ou reflète une difficulté structurelle partagée, nous calculons des corrélations de Spearman par paires entre modèles à partir des scores WER par locuteur·rice, comme illustré à la Figure 1. Les corrélations sont uniformément élevées ($\rho > 0.80$ dans la plupart des cas, avec des intervalles de confiance bootstrap à 95 % étroits). Cette cohérence suggère que le profil de dégradation natif–non natif est largement stable d'un modèle à l'autre : les écarts de performances observés semblent refléter des propriétés linguistiques et acoustiques de la parole accentuée et du L1, plutôt que des effets idiosyncrasiques liés à une architecture, un objectif d'entraînement ou un corpus particulier. Cette observation nous motive l'étude des représentations internes des différents modèles : si la difficulté observée est bien structurelle, elle devrait également se manifester dans la façon dont la parole L2 est positionnée par rapport à la parole native dans les espaces de représentation des modèles.

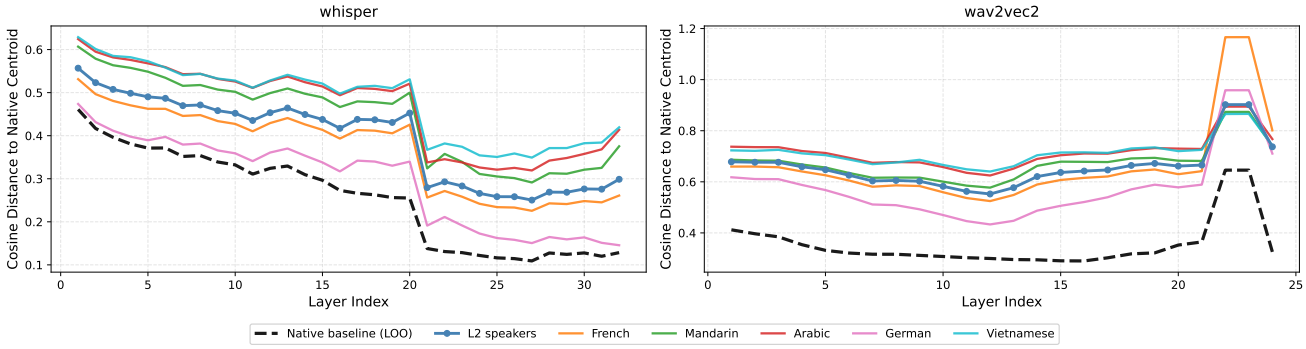

FIGURE 3 – Distance cosinus couche par couche au centroïde natif américain pour `wav2vec2` (droite) et `whisper` (gauche).

Lien entre représentations et dégradation du taux d'erreur Nous testons maintenant si le déplacement géométrique induit par l'accent dans l'espace des représentations prédit les performances de la RAP à travers les langues. En utilisant la méthode introduite à la Section 2, nous calculons la distance entre les représentations d'une langue donnée et celles des locuteur·rice·s natif·ve·s américain·e·s. Nous nous concentrons ici sur les représentations extraites de la dernière couche de l'encodeur, car elles constituent l'entrée directe de la tête CTC pour les modèles wav2vec2/HuBERT/XLSR-53 et du décodeur pour Whisper, et sont donc les plus pertinentes pour les performances en transcription.

Pour chaque groupe L1, nous mettons en relation sa distance moyenne au centroïde natif américain avec son taux d'erreur moyen. La Figure 2 présente les nuages de points correspondants pour `whisper` et `wav2vec2`. Dans les deux architectures, on observe une forte association monotone positive entre distance entre représentations et taux d'erreur : pour `whisper`, $\rho = 0.846$ ($p < 0.001$) ; pour `wav2vec2`, $\rho = 0.849$ ($p < 0.001$). Les langues géométriquement plus éloignées du centroïde natif dans l'espace de représentation de la dernière couche présentent systématiquement un taux d'erreur plus élevé. À titre de contrôle, nous avons répété l'analyse avec des descripteurs MFCC ; dans ce cas, la corrélation avec le taux d'erreur est nettement plus faible, ce qui indique que la relation observée émerge spécifiquement dans les représentations apprises.

Bien que l'existence d'une association positive entre déplacement représentationnel et taux d'erreur soit intuitivement attendue, la force et la cohérence de cet effet sont remarquables. Cela suggère que les décalages liés à l'accent dans l'espace de représentation de la dernière couche ne sont pas compensés par la tête CTC ou le décodeur : une fois que la parole L2 est géométriquement déplacée par rapport à la parole native, les couches de prédiction en aval semblent très sensibles à ce décalage et ne le corrigent pas efficacement. Ces résultats indiquent que le déplacement représentationnel induit par l'accent capture une structure qui affecte directement les performances en reconnaissance.

Distance des représentations à la parole native L'expérience précédente portait sur la dernière couche de l'encodeur et a montré que le déplacement géométrique induit par l'accent prédit fortement la dégradation des performances de RAP. Nous allons maintenant nous intéresser, au-delà de cette couche de sortie, à la manière dont cette structure émerge et évolue en profondeur. Une justification courante des architectures profondes est qu'elles construisent progressivement des représentations de plus en plus abstraites, en éliminant graduellement la variabilité de bas niveau tout en préservant l'information linguistiquement pertinente. Si tel était le cas, on pourrait s'attendre à ce que les différences liées à l'accent s'atténuent dans les couches profondes, à mesure que les représentations

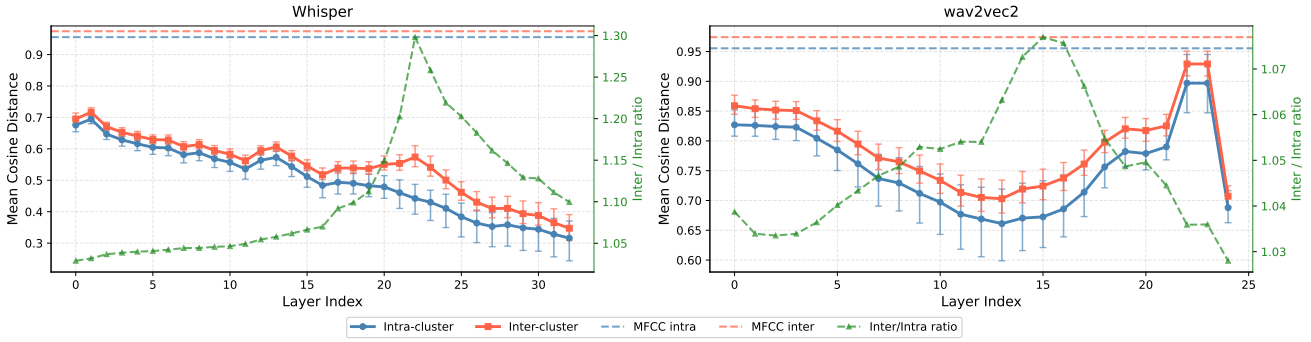

FIGURE 4 – Distances cosinus intra-cluster (cohésion) et inter-cluster (séparation) couche par couche pour les groupes L1 dans `whisper` et `wav2vec2`, comparées aux descripteurs MFCC.

deviennent plus invariantes.

Pour examiner cette évolution dans le réseau, nous répétons l'analyse précédente en considérant les représentations extraites de chaque couche de l'encodeur. Les résultats (Figure 3) montrent qu'une organisation géométrique cohérente apparaît dans les deux architectures. La structure observée dans la dernière couche (où la distance à la parole native reflète les performances RAP) est déjà visible dans les premières couches et reste largement inchangée tout au long du réseau. Autrement dit, l'organisation géométrique des langues selon leur distance à la parole native reflète la difficulté en reconnaissance et demeure remarquablement stable d'une couche à l'autre.

Pour vérifier que ce schéma n'est pas un artefact des quelques langues représentatives sélectionnées, nous menons une analyse plus systématique sur l'ensemble des groupes L1. Pour chaque couche, nous classons les langues selon leur proximité géométrique au centroïde natif américain et calculons la corrélation de rang entre ces classements couche par couche. Les corrélations sont uniformément élevées : la valeur minimale observée est $\rho = 0.991$, indiquant que l'ordre relatif de similarité entre langues est largement préservé en profondeur. La structure induite par l'accent dans l'espace des représentations émerge donc tôt et reste stable à mesure que les représentations sont progressivement affinées.

La dynamique des différentes couches diffère toutefois entre les modèles. Dans `whisper`, les distances diminuent dans les premières couches, évoluent dans les couches supérieures, puis se stabilisent près de la sortie. Dans `wav2vec2`, elles restent largement stables et n'augmentent nettement que dans les dernières couches. Cette différence reflète probablement leurs objectifs d'entraînement respectifs : l'objectif contrastif de `wav2vec2`, appliqué à des états latents masqués, peut encourager la préservation du détail acoustique dans les couches intermédiaires, tandis que l'objectif de transcription de bout en bout de `whisper` peut favoriser une compression plus précoce avant que les représentations ne soient réorganisées pour le décodage. Malgré ces différences, les deux modèles présentent une séparation native–non native persistante et un classement L1 stable, aligné avec les performances RAP.

Distances intra- et inter-clusters Si l'analyse précédente caractérisait les effets de l'accent par leur déplacement par rapport à la parole native, elle ne permettait pas de voir comment les accents non natifs s'organisent les uns par rapport aux autres dans l'espace des représentations. Nous proposons donc d'affiner notre analyse en examinant la structure interne de cet espace à l'aide des distances

intra-cluster (au sein d'une même L1) et inter-cluster (entre L1 différentes).

La Figure 4 rapporte ces quantités couche par couche pour `whisper` et `wav2vec2`. Dans les deux modèles, les locuteur·rice·s partageant la même L1 présentent systématiquement des distances intra-cluster plus faibles que les distances inter-cluster, ce qui indique que les représentations de locuteur·rice·s aux « schémas » accentuels similaires forment des clusters cohérents. Le rapport inter/intra reste nettement supérieur à 1 dans la plupart des couches, montrant que le regroupement par accent persiste tout au long de la succession de couches. Cela complète l'analyse précédente : si la Figure 3 montrait que les langues sont ordonnées selon leur distance à la parole native, la présente analyse révèle que les accents s'organisent également en clusters cohérents. Ces résultats indiquent que la structure accentuelle ne se réduit pas à une simple déviation scalaire par rapport à la parole native, mais façonne la géométrie globale de l'espace des représentations, les différents groupes de L1 occupant des régions distinctes.

Il est intéressant de noter que ce rapport tend à se rapprocher de 1 dans les dernières couches du réseau. Ce comportement ne doit toutefois pas être interprété comme une preuve d'invariance à l'accent. En effet, la distance à la parole native et le classement relatif des langues restent largement inchangés à travers les couches. Une explication plus plausible est que l'espace des représentations subit une compression géométrique dans les couches les plus proches de la tête de prédiction. Ces couches étant directement optimisées pour la fonction de coût considérée (perte CTC ou décodeur), l'espace d'enchâssement peut devenir plus homogène, réduisant le contraste entre distances intra- et inter-cluster sans pour autant éliminer la structure accentuelle sous-jacente.

Prises ensemble, les analyses des Figures 2, 3 et 4 brossent un tableau cohérent de la façon dont l'information accentuelle est organisée dans les encodeurs de parole modernes. Premièrement, les accents induisent un déplacement systématique par rapport à la parole native qui prédit fortement la dégradation du taux d'erreur. Deuxièmement, cette organisation géométrique apparaît dès les premières couches et reste stable tout au long du réseau. Troisièmement, les accents forment des clusters cohérents dans l'espace des représentations plutôt que de converger vers une représentation commune invariante à l'accent.

Ces observations suggèrent que les encodeurs de parole modernes n'éliminent pas progressivement la variation accentuelle à mesure que les représentations deviennent plus abstraites. Ils semblent au contraire construire des représentations dans lesquelles le contenu linguistique reste enchevêtré avec les patrons phonologiques propres à l'accent : les modèles apprennent des représentations de la forme $z = f(\text{contenu}, \text{accent})$ plutôt qu'une représentation purement invariante à l'accent $z = f(\text{contenu})$.

4 Conclusion

Nous avons étudié la façon dont les systèmes de RAP modernes représentent la parole accentuée. Toutes architectures confondues, la parole non native reste géométriquement déplacée par rapport à la parole native tout au long du réseau, et ce déplacement prédit fortement la dégradation du taux d'erreur à travers les langues. Plutôt que de converger vers des représentations invariantes à l'accent, les accents organisent l'espace des représentations en une structure géométrique stable qui émerge tôt et persiste d'une couche à l'autre.

Notre analyse se concentre sur la reconnaissance de l'anglais L2, ce qui permet de contrôler finement le contenu lexical grâce au Speech Accent Archive, mais limite la généralité typologique des conclusions.

Une extension naturelle consisterait à reproduire cette analyse pour d'autres langues cibles, à l'aide de corpus comparables et de systèmes RAP multilingues. Une autre perspective importante serait d'exploiter directement la structure géométrique mise en évidence dans cet article afin d'améliorer la robustesse à la parole accentuée. En particulier, un travail futur pourrait chercher à atténuer ou corriger le déplacement accentuel dans l'espace des représentations, tout en préservant le contenu linguistique nécessaire à la transcription.

Ces résultats suggèrent ainsi qu'améliorer la robustesse à la parole accentuée pourrait nécessiter de modéliser ou de dissocier explicitement la variation liée à l'accent lors de l'apprentissage des représentations.

5 Remerciements

Ce travail a été réalisé dans le cadre du projet DeepTypo financé par l'Agence Nationale de la Recherche (ANR-23-CE38-0003-01).

Références

BAEVSKI A., ZHOU H., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

ETHAYARAJH K. (2019). How contextual are contextualized word representations ? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, p. 55–65, Hong Kong, China.

GAO T., YAO X. & CHEN D. (2019). Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations (ICLR)*.

GRAHAM C. *et al.* (2024). Evaluating openai's whisper asr : Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, **4**(2), 025206.

GRAVES A., FERNÁNDEZ S., GOMEZ F. & SCHMIDHUBER J. (2006). Connectionist temporal classification : Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, p. 369–376.

HSU W.-N. *et al.* (2021). Hubert : Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 3451–3460.

MCAULIFFE M., SOCOLOF M., MIHUC S., WAGNER M. & SONDEREGGER M. (2017). Montreal forced aligner : Trainable text-speech alignment using kaldi. In *Proceedings of Interspeech*, p. 498–502.

MU J. & VISWANATH P. (2018). All-but-the-top : Simple and effective postprocessing for word representations. In *International Conference on Learning Representations (ICLR)*.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2023). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv :2212.04356*.

WEINBERGER S. (2015). Speech accent archive. https://accent.gmu.edu/.

YONG Z. *et al.* (2025). Effects of speaker count, duration, and accent diversity on zero-shot accent robustness in low-resource asr. In *Proc. INTERSPEECH 2025*.

ZAIEM S., KEMICHE Y., PARCOLLET T., ESSID S. & RAVANELLI M. (2023). Speech self-supervised representation benchmarking : Are we doing it right ? In *Proceedings of Interspeech 2023*, p. 2873–2877. DOI : 10.21437/Interspeech.2023-1087.

A Détails expérimentaux supplémentaires

Le Tableau 2 rapporte les checkpoints utilisés pour l'évaluation RAP. Tous les enregistrements sont convertis en mono, rééchantillonnés à 16 kHz et fournis aux modèles via le processeur associé à chaque checkpoint. Pour les modèles CTC, les logits sont décodés par maximisation locale, puis convertis en texte avec le processeur du modèle. Pour Whisper, les transcriptions sont produites avec la méthode de génération autoregressive par défaut.

Pour les modèles CTC, la tête de transcription correspond à la couche de projection linéaire fournie avec le checkpoint, appliquée aux états cachés de l'encodeur afin de produire une distribution sur le

vocabulaire du processeur associé. Nous n'entraînons pas de sonde supplémentaire et ne modifions pas cette tête de classification. Les représentations utilisées dans les analyses géométriques sont extraites avant cette tête, à partir des états cachés de l'encodeur.

Pour Whisper, les transcriptions sont produites avec le modèle encodeur–décodeur complet. Les analyses géométriques portent en revanche sur les états cachés de l'encodeur acoustique. Les états du décodeur ne sont pas inclus dans ces analyses, car ils sont conditionnés par les tokens précédemment générés et ne sont donc pas directement comparables aux représentations acoustiques des encodeurs CTC.

Modèle	Checkpoint	Inférence RAP
wav2vec2-base	`facebook/wav2vec2-base-960h`	CTC
wav2vec2-large	`facebook/wav2vec2-large-960h`	CTC
XLSR-53	`jonatasgrosman/wav2vec2-large-xlsr-53-english`	CTC
HuBERT-large	`facebook/hubert-large-ls960-ft`	CTC
Whisper-base	`openai/whisper-base`	génération autoregressive
Whisper-large	`openai/whisper-large-v3`	génération autoregressive

TABLE 2 – Checkpoints utilisés pour l'évaluation RAP. Les modèles wav2vec2, XLSR-53 et HuBERT correspondent à des checkpoints déjà affinés pour la transcription de l'anglais.