Mesure du niveau de proximité entre enregistrements audio et évaluation indirecte du niveau d'abstraction des représentations issues d'un grand modèle de langage

Maxime Fily, Guillaume Wisniewski, Séverine Guillaume, Gilles Adda, Alexis Michaud

Mesure du niveau de proximité entre enregistrements audio et évaluation indirecte du niveau d'abstraction des représentations issues d'un grand modèle de langage

Maxime Fily[1, 2] Guillaume Wisniewski[1] Séverine Guillaume[2] Gilles Adda[3]
Alexis Michaud[2]
(1) LLF, CNRS, Université Paris-Cité, F-75013, Paris, France
(2) LACITO, CNRS, Université Sorbonne Nouvelle, F-94800, Villejuif, France
(3) LISN, CNRS, Université Paris-Saclay, F-91405, Orsay, France
maxime.fily@gmail.com, guillaume.wisniewski@u-paris.fr,
{severine.guillaume, alexis.michaud}@cnrs.fr, gilles.adda@limsi.fr

RÉSUMÉ

Nous explorons les représentations vectorielles de la parole à partir d'un modèle pré-entraîné pour déterminer leur niveau d'abstraction par rapport au signal audio. Nous proposons une nouvelle méthode non-supervisée exploitant des données audio ayant des métadonnées soigneusement organisées pour apporter un éclairage sur les informations présentes dans les représentations. Des tests ABX déterminent si les représentations obtenues via un modèle de parole multilingue encodent une caractéristique donnée. Trois expériences sont présentées, portant sur la qualité acoustique de la pièce, le type de discours, ou le contenu phonétique. Les résultats confirment que les différences au niveau de caractéristiques linguistiques/extra-linguistiques d'enregistrements audio sont reflétées dans les représentations de ceux-ci. Plus la quantité d'audio par vecteur est importante, mieux elle permet de distinguer les caractéristiques extra-linguistiques. Plus elle est faible, et mieux nous pouvons distinguer les informations d'ordre phonétique/segmental. La méthode proposée ouvre de nouvelles pistes pour la recherche et les travaux comparatifs sur les langues peu dotées.

ABSTRACT

Establishing degrees of closeness between audio recordings along different dimensions using large-scale cross-lingual models

In the highly constrained context of low-resource language studies, we examine vector representations of speech from a pretrained model to determine their level of abstraction from the audio signal. We propose a new unsupervised method using ABX tests on audio recordings with carefully curated metadata to shed light on the type of information present in the representations. ABX tests determine whether the representations computed by a multilingual speech model encode a given characteristic. Three experiments are devised : one on room acoustics aspects, one on linguistic genre, and one on segmental phonetic aspects. The results confirm that the representations extracted from recordings with different linguistic/extra-linguistic characteristics differ along the same lines. Embedding more audio into a vector better discriminates extra-linguistic characteristics, whereas shorter snippets are better for distinguishing segmental information. The method is fully unsupervised, potentially opening up new avenues of research for comparative work on under-documented languages.

MOTS-CLÉS : TAL, langues peu dotées, méthodes non-supervisées.

KEYWORDS: NLP, under-documented languages, unsupervised methods.

1 Introduction

La parole, lorsqu'elle se présente sous la forme d'enregistrements audio, est multi-factorielle : une voix enregistrée transmet un message, une intention, une émotion. L'enregistrement contient également des informations au delà du signal de parole, sur l'environnement, l'identité du locuteur ou de la locutrice par exemple. Ce travail aborde la question de la nature de l'information encodée dans les représentations vectorielles produites par un réseau de neurones appris de manière auto-supervisée comme `wav2vec2` (Baevski *et al.*, 2020). Notre objectif est d'étudier le niveau d'abstraction présent dans les représentations construites par ces modèles et plus précisément de vérifier si celles-ci ne contiennent que des informations liées au contenu linguistique ou si elles contiennent également des informations *indexicales* (Foulkes, 2010, 7).

Notre dispositif expérimental s'appuie sur des jeux de données « sur mesure », sélectionnées à partir de corpus de linguistique documentaire afin d'évaluer comment une différence donnée dans le signal d'entrée se reflète dans les vecteurs construits par le réseau de neurones. Considérer des données issues de ce type de corpus nous permet d'accéder à des métadonnées riches sur les conditions d'enregistrement. Nous utilisons des tests `ABX` pour mesurer l'impact de certains facteurs (locuteur·rice, microphone, ...) sur les distances entre représentations à travers trois jeux de données décrits en section 2.

Les résultats fournissent un aperçu sur la nature des informations encodées dans les représentations d'un modèle tel que XLSR-53 (Baevski *et al.*, 2020; Babu *et al.*, 2021), et suggèrent que les tests `ABX` peuvent être exploités pour faire ressortir des différences dans la configuration acoustique (salle, microphone), certaines caractéristiques de la voix, ou dans le contenu linguistique. Une étude paramétrique montre que le traitement de l'audio par extraits de 10 s est suffisant pour faire ressortir les différences dans la configuration acoustique et dans les propriétés de la voix, tandis que des extraits de 1 s sont meilleurs pour explorer les caractéristiques segmentales.

Cette étude fournit une méthode innovante pour détecter des facteurs de confusion dans des corpus destinés à de l'apprentissage automatique, ainsi qu'un moyen pour accélérer la classification d'enregistrements (p.ex. par niveau de bruit, par type) selon des informations fines, qui ne sont pas toujours présentes dans les métadonnées.

2 Méthode expérimentale

Notre méthode repose sur : (i) des tests de similarité pour déterminer si une caractéristique d'un enregistrement audio est encodée dans la représentation vectorielle d'un enregistrement ou non, et (ii) des corpus audio avec des métadonnées riches, qui nous permettent de construire des ensembles de données partageant ou non **une caractéristique à la fois** : langue, locuteur·rice, acoustique de la pièce, type de microphone, caractéristiques de la voix, contenu segmental, etc.

Tests ABX Pour déterminer, de manière non supervisée, si un modèle de parole encode une caractéristique $\mathcal{C}$ du signal de parole, nous utilisons un test `ABX` (Carlin *et al.*, 2011; Schatz *et al.*, 2013). Ce test s'appuie sur les représentations vectorielles construites par un modèle pré-entraîné de

trois audios : deux extraits, notés A et X, partagent une caractéristique $\mathcal{C}$, et un, noté B, ne la possède pas. Le test ABX consiste simplement à vérifier si la distance [1] $d(A, X)$ est plus petite que $d(A, B)$.

Le score ABX correspond à la proportion de triplets pour lesquels la condition $d(A, X) < d(A, B)$ est vraie. Un score ABX proche de 50 % indique qu'en moyenne, la distance entre A et X est supérieure ou égale à la distance entre A et B, ce qui suggère que $\mathcal{C}$ n'est pas encodé dans la représentation vectorielle de l'enregistrement. Inversement, plus le score est proche de 100 %, plus la représentation capture la caractéristique $\mathcal{C}$.

Les tests ABX sont intéressants pour les scénarios à faibles ressources (comme le nôtre) car ils ne nécessitent pas de données pour entraîner un classifieur (contrairement aux sondes linguistiques (Belinkov & Glass, 2019) souvent utilisées pour analyser les représentations vectorielles) : toutes les données disponibles sont utilisées pour tester si la propriété est encodée dans la représentation ou non.

Corpus Notre étude s'appuie sur des enregistrements en na et en naxi, deux langues parlées dans le sud-ouest de la Chine. Le na est la langue maternelle d'environ 50 000 personnes. Le naxi est plus répandu, puisqu'il est la langue maternelle d'environ 200 000 personnes. Typologiquement similaires (langues SOV, de structure (C)(G)V+T [2], isolantes) le naxi et le na sont des langues apparentées, sans être pour autant mutuellement intelligibles. Les systèmes tonals diffèrent au niveau de la complexité des phénomènes morpho-tonologiques qui est très importante en na (Michaud, 2017, 425). Au plan phonologique, le na a une nasalité de type vocalique derrière les consonnes /h/ et plus marginalement derrière une attaque vide (Michaud *et al.*, 2012). Ces deux langues sont aujourd'hui en déclin, car progressivement remplacées par le mandarin, langue officielle utilisée dans les écoles, les administrations et les médias (Michaud & Latami, 2011; Zhao, 2022).

Tous les enregistrements proviennent de la collection Pangloss, une archive en libre accès de « langues peu documentées », et sont consultables via les DOI fournis en annexe (voir https://hal.science/hal-04583516). Trois séries d'enregistrements sélectionnées pour leurs caractéristiques sont examinées :

(i) La ***série du conte populaire*** consiste en sept sessions d'enregistrement d'un même conte en na, raconté par une même locutrice. La i^e session d'enregistrement sera désignée par V_i. Cette série se concentre sur l'effet des conditions d'enregistrement, qui sont légèrement différentes d'une version à l'autre, et que nous chercherons à distinguer à l'aide de tests ABX sont effectués. Plus précisément, nous nous concentrons sur trois lots :

- Le premier lot étudié comprend trois versions : V_1, V_2 et V_3. V_1 a été enregistré dans une salle avec une réverbération perceptible, tandis que V_2 et V_3 ont été enregistrés dans une salle mieux isolée phoniquement.
- Le deuxième lot est composé de V_6 et V_7. Ces deux versions ont été enregistrées dans les mêmes conditions acoustiques. L'audio a été capté simultanément par deux microphones : un micro-casque et un micro à main inséré dans un petit support.
- Le troisième lot compare V_4 et V_5 à tous les autres enregistrements de la *série du conte populaire*. Les enregistrements V_4 et V_5 ont pour récepteur [3] un auditeur natif alors que dans les autres enregistrements le récepteur était le linguiste collectant les données.

1. Nous avons utilisé une distance cosinus dans toutes nos expériences.
2. G = *Glide*, T = Ton.
3. Ici, récepteur est le destinataire d'un message, en cohérence avec la terminologie de Shannon (1948) qui propose l'idée très *traitement du signal* qu'il existe un canal de communication *(communication channel)* entre un émetteur et un récepteur. Ce canal est doublé d'un canal de rétroaction en sens inverse *(communication backchannel)*, qui influence la façon dont l'émetteur livre son message.

Ces enregistrements sont particulièrement intéressants car certaines variables (le sujet de conversation et le·la locuteur·rice) sont contrôlées, ce qui permet de se concentrer sur l'influence d'autres facteurs spécifiques (par exemple, l'acoustique de la pièce).

(ii) La ***série des répertoires de chant*** consiste en cinq enregistrements d'une même chanteuse professionnelle Naxi. Trois enregistrements sont de la chanson seule, un enregistrement est un récit, et un enregistrement comporte les deux genres (« Alili », qui est 50% texte et 50% chanson). Notre objectif est de comparer ces enregistrements. Un chanteur entraîné présente des propriétés vocales très différentes selon qu'il chante ou qu'il parle : en particulier, le timbre des voyelles et la tessiture sont affectées (Castellengo, 2016, 458). Ces différences sont perçues par les auditeurs (Castellengo, 2016, 187). Cette expérience vise à vérifier si ces différences sont reflétées dans les représentations.

(iii) La ***série phonétique*** est composée de cinq enregistrements d'élicitations phonétiques basés sur un même corpus et d'un enregistrement de mots dans une phrase porteuse basé sur un corpus différent, en langue na. Trois locutrices identifiées comme `AS`, `RS` et `TLT` sont prises en compte. Nous avons inclus deux sessions d'enregistrement, ce qui permet une comparaison inter et intra-locuteurs.

En utilisant ces enregistrements, nous adoptons une approche à l'opposé de ce qui se fait dans de nombreux travaux portant sur l'analyse des représentations neuronales : plutôt que de considérer de « grands » corpus avec de fortes variabilités (approche *big data*), nous privilégions un corpus de petite taille mais dans lequel de nombreux facteurs sont contrôlés (approche *beautiful data*).

Modèle Dans toutes nos expériences, nous utilisons le modèle `XLSR-53`[4], une architecture `wav2vec2` entraînée sur 56 kh de données audio (brutes) dans 53 langues (Conneau *et al.*, 2020), y compris des langues à tons (p.ex. mandarin, cantonais, lao, ...). Ni le na ni le naxi ne sont présents dans les données de pré-entraînement de ce modèle, mais Macaire *et al.* (2021) ont montré que ce modèle pouvait être affiné pour faire de la reconnaissance de la parole sur du na (après affinage, le CER en transcription est de 8 %), et qu'il était donc capable de gérer la diversité des réalisations de surface de cette langue, et notamment les tons.

Pour les comparaisons, nous considérons des extraits audio d'une durée de 1, 5, 10 et 20 secondes afin d'étudier l'effet de la longueur de l'extrait sur notre test `ABX`. Nous utilisons une stratégie de *max-pooling* pour construire un unique vecteur représentant l'extrait. Nous créons ensuite des cartes thermiques des scores `ABX`, pour comparer un à un les enregistrements.

Nous utilisons les représentations de la couche 21. Ce choix est basé sur les conclusions de Pasad *et al.* (2021, 2023) et Li *et al.* (2022, 2023) ainsi que sur notre pré-étude de sensibilité (méthode des sondes linguistiques), qui montrent toutes que la capacité de `wav2vec2` à capturer l'information linguistique diminue fortement dans les trois dernières couches (voir annexes : `https://hal.science/hal-04583516`).

3 Résultats

Avec les tests `ABX`, nous étudions si les représentations audio calculées par `wav2vec2` capturent ou non des informations spécifiques dans des signaux audio soigneusement sélectionnés. Notre approche est non-supervisée ce qui nous permet de l'appliquer à de petits jeux de données. Nos expériences

4. L'API HuggingFace a été utilisée (signature du modèle : `facebook/wav2vec2-large-xlsr-53`).

préliminaires (voir les annexes : https://hal.science/hal-04583516) montrent que si le choix de la couche (21 dans nos expérience) a un impact fort sur les résultat d'une sonde linguistique, ce n'est plus le cas avec notre approche.

3.1 Étude des différentes versions d'un même conte populaire

Notre première expérience a pour objectif de déterminer si des variables extra-linguistiques telles que l'acoustique de la pièce et le type de microphone sont présentes dans les représentations neuronales. Pour ce faire, nous utilisons un test ABX pour distinguer les différents enregistrements de la *série de contes populaires* : ces scores sont calculés à partir de triplets composés de deux extraits de 10 s d'un même enregistrement et d'un extrait de 10 s d'un enregistrement différent. [5]

La figure 1 montre que, dans la plupart des cas, avec une longueur d'extrait de 10 s, il est possible de distinguer les différents enregistrements, bien qu'il s'agisse toujours du même locuteur racontant la même histoire : hormis quelques rares exceptions, qui seront abordées plus loin, les scores rapportés sont largement supérieurs à 50 %. De plus, les scores sur la diagonale, correspondant à des tests où tous les extraits proviennent du même enregistrement, sont tous proches de 50 %. Cela indique clairement que les différences constatées dans les autres tests ABX ne sont pas dues au contenu linguistique (les mots prononcés), mais plutôt à la configuration acoustique et suggère que les représentations neuronales capturent bien plus que les informations linguistiques nécessaires à la compréhension de la parole : elles encodent également des informations liées aux conditions d'enregistrement.

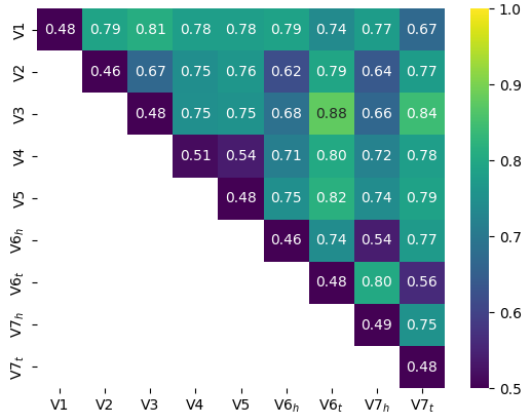

FIGURE 1 – Scores ABX pour l'étude de la *série du conte populaire*. Taille des extraits : 10 s.

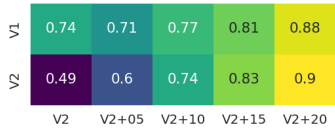

FIGURE 2 – Reproduction de la réverbération de la pièce V_1 avec une réverbération artificielle appliquée sur V_2. Taille des extraits : 5 s.

Une mise en perspective de ces observations avec les conditions d'enregistrement (rendue possible par une connaissance fine des enregistrements) permet de mieux comprendre les informations capturées (ou non) par les représentations. Ainsi, la comparaison de V_1, V_2 et V_3 montre que les représentations de V_2 et V_3 ne sont que peu discernables entre elles (score ABX de 0,67) mais peuvent être distinguées de celles de V_1 (scores de 0,79 et 0,81). La principale différence entre ces trois enregistrements est

5. Les résultats pour d'autres durées d'extraits sont rapportés en annexe (voir https://hal.science/hal-04583516).

liée au lieu d'enregistrement : V_2 et V_3 ont été enregistrés dans un lieu moins réverbérant que le lieu où V_1 a été enregistré. Pour confirmer l'influence de ce paramètre, nous avons réalisé une expérience de contrôle en ajoutant artificiellement de la réverbération [6] aux enregistrements V_2 et en mesurant le score ABX entre les enregistrements V_1 et V_2 modifiés. La figure 2 montre l'évolution du score ABX en fonction de la quantité de réverbération ajoutée. Lorsque la quantité de réverbération dans V_2 augmente, le score ABX diminue d'abord avant d'augmenter à nouveau. Cela signifie que V_1 est plus proche de V_2 avec 5 % de réverbération, ce qui suggère une relation de causalité entre la quantité de réverbération et le degré de proximité entre les enregistrements de ce lot.

Une seconde comparaison intéressante est celle entre les différentes versions des enregistrements V_6 et V_7. Ces enregistrements ont été réalisés avec un micro casque (enregistrements étiquetés h pour *headset*) et un micro à main inséré dans un support placé sur une table (enregistrements étiquetés t pour *table*). La figure 1 montre que les représentations peuvent être distinguées avec précision sur le type de microphone. Par exemple, les scores ABX entre $V_{6,h}$ et $V_{6,t}$ sont parmi les plus élevés de notre expérience, alors que pour deux enregistrements différents réalisés avec le même microphone (c'est-à-dire $V_{6,h}$-$V_{7,h}$ et $V_{6,t}$-$V_{7,t}$), les scores ABX ne sont que légèrement meilleurs que les scores pour le même enregistrement. Cela suggère que les représentations issues d'extraits de 10 s dépendent fortement du microphone utilisé : deux vecteurs représentant le même signal audio mais issus de microphones différents sont plus dissemblables que ceux représentant deux signaux audio différents enregistrés par le même microphone.

La figure 1 fait également ressortir une similitude inattendue entre les enregistrements V_4 et V_5. Le score ABX entre ces deux enregistrements n'est que de 54 %, alors qu'il n'est jamais inférieur à 71 % entre V_4, V_5 et toutes les autres paires. Or, V_4 et V_5 sont les seuls enregistrements dans lesquels le récepteur était un locuteur de la communauté linguistique. Cela ressemble à un cas d'adaptation linguistique (Piazza *et al.*, 2022) et suggère qu'il serait possible de générer automatiquement des hypothèses sur le contexte d'énonciation, par exemple, sur la base des méta-données.

Dans cette série d'expériences, nos observations sont plus visibles avec des extraits de 10 s. Cela semble être le réglage approprié pour mettre en évidence des différences au niveau de l'acoustique globale. Cette taille d'extrait semble aussi convenir pour révéler les différences au niveau du rythme de parole, de la prosodie. D'autres expériences sont néanmoins nécessaires pour confirmer cela.

3.2 Étude de différents répertoires de chansons

Le but de cette expérience est d'explorer si les paramètres d'extraction qui fonctionnent le mieux dans l'expérience précédente permettent ou non d'explorer les représentations en fonction des caractéristiques de voix du locuteur ou de la locutrice. Plusieurs enregistrements d'une chanteuse professionnelle Naxi sont comparés les uns aux autres : une chanson dans le style « Alili », deux dans le style « Guqi », une dans le style « Wo Menda », et un récit. Les chansons contenaient à l'origine une introduction non chantée qui a été supprimée pour les comparaisons, sauf pour la chanson de style « Alili », qui est pour moitié du texte et pour moitié une chanson.

La figure 3 montre que toutes les chansons se distinguent fortement du récit, à l'exception de l'enregistrement « Alili », qui est mi-texte mi-chanson. Il est intéressant de noter que celui-ci ne correspond ni aux chansons ni au récit : il se situe à mi-chemin entre les deux. Quant aux deux chansons dans le style « Guqi », elles présentent le score ABX le plus bas (0,57) même si leur contenu

6. Nous utilisons `Audacity` pour ajouter 5, 10, 15 ou 20 % de réverbération.

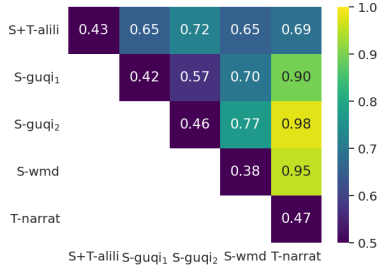

FIGURE 3 – Scores ABX pour les comparaisons entre les différents genres (T=texte, S=chanson). Des chansons dans trois styles différents sont interprétées par une chanteuse professionnelle Naxi. Taille des extraits : 10 s.

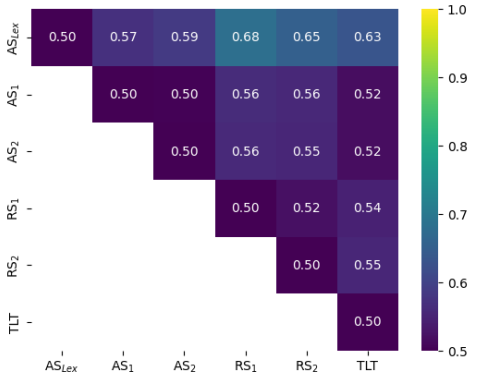

FIGURE 4 – Scores ABX pour les comparaisons entre les éléments de la *série phonétique*. La locutrice AS a trois enregistrements (AS_1, AS_2, AS_{Lex}), RS en a deux (RS_1, RS_2) et TLT en a un. Taille des extraits : 1 s.

linguistique est différent, ce qui suggère que le style de la chanson peut être détecté.

Ces résultats suggèrent que les propriétés de la voix sont présentes dans les représentations, puisque nous pouvons distinguer entre une narration et différents styles de chansons pour un même locuteur, et même regrouper par style de chanson. Ces résultats fournissent des pistes pour de futures études visant à utiliser des modèles neuronaux pour réaliser des études prosodiques.

3.3 Étude d'un corpus de phonétique

S'il est raisonnable de penser que deux phrases au contenu linguistique différent dans des conditions parfaitement contrôlées apparaîtront comme différentes lorsqu'elles seront soumises à un test ABX, la réponse n'est pas immédiate lorsqu'il s'agit d'un enregistrement entier. Il n'est pas non plus évident que deux phrases différentes prononcées par deux locuteurs·rices différent·e·s soient distinguées uniquement en raison d'une différence de contenu linguistique : l'identité du locuteur ou de la locutrice agit comme un facteur de confusion.

L'objectif de cette expérience est de comparer des données présentant des différences d'ordre « phonétique », mais où le paramètre de l'identité de la locutrice varie. Pour ce faire, nous nous appuyons sur un corpus phonétique enregistré de manière contrôlée, où chaque locutrice a reçu les mêmes instructions. Ces enregistrements ont le même contenu ($AS_{1,2}$, $RS_{1,2}$, TLT). Un enregistrement a un contenu différent (AS_{lex}). Les scores sont calculés à partir de triplets composés de deux extraits de 1 s issus du même enregistrement et d'un extrait issu d'un enregistrement différent. [7]

La figure 4 montre qu'avec une longueur de 1 s, il est difficile de distinguer les différents enregistrements des mêmes phrases (c.-à-d. issus du même corpus, mais prononcé par des locutrices différentes, ou simplement différentes répétitions d'un même enregistrement), et ce même lorsque les locutrices

7. Les résultats obtenus pour d'autres tailles d'extraits sont présentés en annexe (voir https://hal.science/hal-04583516).

diffèrent. l'écart à locutrice fixe varie de 0,0 à 0,02, tandis que, tandis que l'écart « cross-locutrices » varie de 0,02 à 0,06. Cela suggère que même avec des extraits d'1 s, l'identité de la locutrice est toujours reflétée, mais d'une manière difficilement repérable pour cette taille d'extrait. En revanche, l'information locuteur·rice ressort extrêmement clairement pour une taille d'extrait de 10 s (voir https://hal.science/hal-04583516), ce qui suggère que les représentations neuronales, pour de petits extraits, « centrifugent » mieux les informations extra-linguistiques. Cette observation n'est pas surprenante étant donné la façon dont les modèles sont pré-entraînés (Baevski *et al.*, 2020), et elle constitue une amorce intéressante pour la deuxième partie de l'analyse, qui consiste à comparer ces enregistrements de phrases identiques à un autre enregistrement avec des phrases différentes.

Les résultats de la première ligne de la Figure 4, par les différences observées relativement aux autres lignes de la figure, suggèrent que les tests ABX révèlent des différences lorsque le contenu linguistique diffère. Ces différences viennent vraisemblablement s'additionner à des différences d'identité de locuteur·rice, mais celles-ci sont suffisamment réduites pour laisser apparaître les différences au niveau du contenu.

Dans cette étude, les scores ABX sont calculés en moyenne sur l'ensemble d'un enregistrement. Pour les différences phonétiques, il serait intéressant de pouvoir effectuer des comparaisons par phrase, mais cela reviendrait à s'écarter d'une approche entièrement non supervisée.

4 Discussions et conclusions

Entreprendre de comparer des représentations vectorielles de parole, en mode non-supervisé, s'apparente à une gageure tant la parole est multi-factorielle et sujette à variation. Nous avons adopté une méthode expérimentale pour soumettre un modèle donné à différentes expériences avec des variables de test, en contrôlant par ailleurs un certain nombre de paramètres. Il faut tout d'abord mentionner que la connaissance des données étudiées et de leurs métadonnées a été déterminante car elle nous a permis d'obtenir des résultats dans un contexte où les corpus sont *de facto* de petite taille puisqu'ils concernent des langues peu documentées, et donc difficilement exploitables en TAL. Cette étude nous a permis dans un premier temps de mieux comprendre le degré d'abstraction des représentations de la parole issues de grands modèles de langage.

Ainsi, nous avons montré que le contenu des représentations n'était pas exclusivement centré sur le contenu linguistique dans la mesure où, en particulier lorsque les représentations encodent plusieurs secondes d'audio, il nous a été possible de corréler les résultats des tests ABX avec des informations relatives à l'acoustique de la pièce, le type de microphone, et le type de discours. Nous voyons donc que la méthode employée pourrait servir pour distinguer des enregistrements en fonction d'un grand nombre de critères extra-linguistiques.

Enfin, lors de l'étude de représentations d'extraits audio de petite taille, nous avons constaté que le poids des facteurs extra-linguistiques diminuait dans les tests ABX, et que des critères phonétiques/segmentaux prenaient au contraire en importance, ce qui est encourageant dans la perspective de comparaison typologiques de langues voisines entre elles, à l'aide d'approches non-supervisées.

L'utilisation de méthodes non-supervisées, notamment faisant intervenir des mesures telles que la distance cosinus, trouvent des applications pour l'amélioration des méthodes de reconnaissance automatique pour les langues rares (San *et al.*, 2024). Notre communication permet de mieux comprendre le lien entre l'audio et ses représentations vectorielles.

Références

BABU A., WANG C., TJANDRA A., LAKHOTIA K., XU Q., GOYAL N., SINGH K., VON PLATEN P., SARAF Y., PINO J. *et al.* (2021). XLS-R : Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv :2111.09296*.

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, **33**, 12449–12460.

BELINKOV Y. & GLASS J. (2019). Analysis methods in neural language processing : A survey. *Transactions of the Association for Computational Linguistics*, **7**, 49–72.

CARLIN M. A., THOMAS S., JANSEN A. & HERMANSKY H. (2011). Rapid evaluation of speech representations for spoken term discovery. In *Twelfth Annual Conference of the International Speech Communication Association*.

CASTELLENGO M. (2016). *Ecoute musicale et acoustique*. Paris : Eyrolles.

CONNEAU A., BAEVSKI A., COLLOBERT R., MOHAMED A. & AULI M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, **abs/2006.13979**.

FOULKES P. (2010). Exploring social-indexical knowledge : A long past but a short history. *Laboratory Phonology*, **1**(1), 5–39.

LI Y., BELL P. & LAI C. (2022). Fusing ASR outputs in joint training for speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 7362–7366 : IEEE.

LI Y., MOHAMIED Y., BELL P. & LAI C. (2023). Exploration of a self-supervised speech model : A study on emotional corpora. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, p. 868–875 : IEEE.

MACAIRE C., WISNIEWSKI G., GUILLAUME S., GALLIOT B., JACQUES G., MICHAUD A., ROSSATO S., NGUYÊN M.-C. & FILY M. (2021). Spécialisation de modèles neuronaux pour la transcription phonémique : premiers pas vers la reconnaissance de mots pour les langues rares. In *Journées scientifiques du Groupement de recherche "Linguistique informatique, formelle et de terrain" (GDR LIFT)*, Journées GDR LIFT 2021, Grenoble, France. HAL : halshs-03475443.

MICHAUD A. (2017). *Tone in Yongning Na : lexical tones and morphotonology*. Volume 13 de Studies in Diversity Linguistics. Berlin : Language Science Press. 10.5281/zenodo.439004.

MICHAUD A., JACQUES G. & RANKIN R. L. (2012). Historical transfer of nasality between consonantal onset and vowel : from c to v or from v to c ? *Diachronica*, **29**(2), 201–230.

MICHAUD A. & LATAMI D. (2011). A description of endangered phonemic oppositions in Mosuo (Yongning Na). In T. DE GRAAF, X. SHIXUAN & C. BRASSETT, Éds., *Issues of language endangerment*, p. 55–71. Beijing : Intellectual Property Publishing House.

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 : IEEE.

PASAD A., SHI B. & LIVESCU K. (2023). Comparative layer-wise analysis of self-supervised speech models.

PIAZZA G., MARTIN C. D. & KALASHNIKOVA M. (2022). The acoustic features and didactic function of foreigner-directed speech : A scoping review. *Journal of Speech, Language, and Hearing Research*, **65**(8), 2896–2918.

SAN N., PARASKEVOPOULOS G., ARORA A., HE X., KAUR P., ADAMS O. & JURAFSKY D. (2024). Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens.

SCHATZ T., PEDDINTI V., BACH F., JANSEN A., HERMANSKY H. & DUPOUX E. (2013). Evaluating speech features with the minimal-pair ABX task : Analysis of the classical MFC/PLP pipeline. In *INTERSPEECH 2013 : 14th Annual Conference of the International Speech Communication Association*, p. 1–5.

SHANNON C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**(3), 379–423. DOI : 10.1002/j.1538-7305.1948.tb01338.x.

ZHAO S. (2022). *Looking for a disappearing voice : place making, place-belongingness, and Naxi language vitality in Lijiang Ancient Town*. Thèse de doctorat, Massey University, Wellington, New Zealand.