# Déterminer la similarité entre deux langues à l'aide des modèles pré-entraînés de la parole. Une étude pilote

Séverine Guillaume, Guillaume Wisniewski

# Déterminer la similarité entre deux langues à l'aide des modèles pré-entraînés de la parole. Une étude pilote

Séverine Guillaume[1] Guillaume Wisniewski[2]
[1] Langues et Civilisations à Tradition Orale (LACITO), UMR 7107 CNRS - Sorbonne Nouvelle - INALCO
[2] Laboratoire de Linguistique Formelle (LLF), UMR 7110 CNRS - Université de Paris Cité
severine.guillaume@cnrs.fr guillaume.wisniewski@u-paris.fr

Résumé

Motivés par les résultats obtenus avec des modèles neuronaux comme `wav2vec` pour la transcription de langues « rares » et/ou « peu dotées », nous souhaitons pousser ces modèles encore plus loin en montrant qu'ils peuvent apporter un autre type d'aide aux linguistes dans leur travail de documentation et d'anlayse des langues en permettant d'extraire automatiquement des informations typologiques (inventaire de phonèmes, indices de complexité phonologique et morphosyntaxique, ...), d'enregistrements audio. Dans cette étude pilote, nous décrivons une première série d'expériences visant à déterminer dans quelle mesure les modèles de la parole multilingue peuvent être utilisés pour détecter des langues « similaires » au plan phonético-phonologique.

Abstract

**Probing `wav2vec` for Typological Signal. A Pilot Study**

Motivated by the results obtained with neural models such as `wav2vec` for the transcription of "rare", "under-resourced" languages, we wish to push these models even further by showing that they can bring another kind of help to linguists documenting and anlayzing languages by allowing the automatic extraction of typological information (phoneme inventory, phonological and morphosyntactic complexity phonological and morphosyntactic complexity, ...) from audio recordings. In this pilot study, we describe a first series of experiments to determine to what extent models of multilingual speech model can be used to detect languages that are phonetically/phonologically "similar".

Mots-clés : documentation computationnelle des langues, langues rares, apprentissage profond, typologie linguistique.

Keywords : computational language documentation, endangered languages, deep learning, linguistic typology.

## 1 Introduction

Les modèles de représentation de la parole multilingue appris de manière non supervisée par les réseaux de neurones comme `XLS-R` (Conneau *et al.*, 2021) et `HuBERT` (Hsu *et al.*, 2021) per-

mettent de développer des systèmes de reconnaissance de la parole de bonne qualité à partir de très peu de données annotées. Ces technologies ouvrent de nombreuses possibilités pour la linguistique documentaire computationnelle : de nombreux travaux (Foley *et al.*, 2018 ; Anastasopoulos *et al.*, 2020 ; Partanen *et al.*, 2020), dont les nôtres (Macaire *et al.*, 2021 ; Guillaume *et al.*, 2022b,a), ont montré qu'il est possible de développer, pour des langues rares ou peu documentées, des systèmes de traitement de la parole pour faciliter le travail de transcription et d'annotation des linguistes de terrain. Nous souhaitons désormais pousser ces modèles encore plus loin en montrant que les représentations au cœur des systèmes de l'état de l'art peuvent apporter un autre type d'aide aux linguistes. Il s'agit de déceler automatiquement des propriétés importantes relatives à la typologie d'une langue (telles que : inventaire de phonèmes, indices de complexité phonologique et morphosyntaxique...), et, au moyen de ces propriétés, d'estimer le degré de proximité d'une langue avec telle ou telle autre.

Le travail exploratoire présenté ici est une première étape vers l'objectif ambitieux exposé ci-dessus. Nous souhaitons, dans un premier temps, déterminer dans quelle mesure les modèles de la parole multilingue peuvent être utilisés pour détecter des langues « similaires » au plan phonético-phonologique. Plus précisément, nous cherchons à savoir si les représentations de `XLS-R` permettent de distinguer deux langues et, plus généralement, dans quelle mesure deux langues « proches » (par exemple deux dialectes d'une même langue) auront des représentations plus proches que deux langues « éloignées ».

## 2 Estimer la similarité phonético-phonologique entre langues

Les modèles multilingues de représentation de la parole comme `XLS-R` ou `HuBERT` permettent de représenter un segment audio par un vecteur de taille fixe. Nous souhaitons vérifier que ces représentations encodent une information qui permet de caractériser la langue du segment.

Pour cela, nous nous inspirons des tests ABX utilisés en psychologie [1] ou dans l'analyse des représentations neuronales (de Seyssel *et al.*, 2022) ou pour l'apprentissage de systèmes d'identification du locuteur (Bredin, 2017) pour déterminer de manière complètement non supervisée (c'est-à-dire sans aucune annotation) si un modèle est capable de distinguer deux langues ou non. Ce test consiste à considérer les représentations construites par un modèle donné de deux segments audio de taille fixe $A$ et $X$ d'une langue donnée $L_1$ et un segment $B$ d'une deuxième langue $L_2$, et à vérifier si la distance $d(A, X)$, mesurée par exemple par la distance cosinus, entre les deux vecteurs représentant $A$ et $X$ est plus petite que la distance $d(A, B)$ entre les représentations de deux segments dans des langues différentes.

Ce processus est répété pour un grand nombre de triplets $(A, B, X)$ et nous calculons le score ABX correspondant à la proportion de triplets pour lesquels la condition $d(A, X) < d(A, B)$ est bien vérifiée. Plus ce score est grand, plus les représentations construites par le modèle pour les segments de la langue $L_1$ sont différents des segments de la langue $L_2$. Intuitivement, ce score ABX peut également être utilisé pour mesurer une distance entre langues : deux

1. Notons toutefois que, contrairement à l'approche adoptée ici, les tests ABX en psychologie sont souvent réalisés sur des « paires minimales », c'est-à-dire des exemples identiques à une caractéristique prêt (p. ex. un phonème).

langues proches seront plus confondues par le modèle et auront un score ABX plus petit que deux langues très différentes.

# 3 Expérience

Nous mettons en œuvre ce principe sur six langues de la collection `Pangloss` : quatre langues sino-tibétaines (trois langues de Chine : na de Yongning, na de Shekua, japhug ; et le thulung, parlé au Népal) ; le cèmuhî (langue austronésienne, Nouvelle Calédonie) ; et le nashta (langue indo-européenne, Grèce). Les deux premières langues sont des langues pouvant être qualifiées de « proches » (elles sont identifiées comme dialecte dans Pangloss).

**Validation du protocole expérimental** Dans une première expérience visant à valider le protocole expérimental décrit dans la section précédente, nous ne considérons que la paire de langue na du Yongning/japhug, les deux langues utilisées dans nos précédents travaux (Macaire *et al.*, 2021). Nous reportons, à la Table 1, l'évolution du score ABX en fonction de la longueur des segments considérés pour une paire de langue donnée. Notons que tous ces scores sont calculés de manière symétrique : il y a autant de triplets contenant deux enregistrements de la 1$^{e}$ langue que de triplets contenant deux enregistrements de la 2$^{e}$ langue. Intuitivement, plus les segments sont longs plus le modèle disposera d'information pour distinguer les langues et on peut donc s'attendre à ce que les scores ABX augmentent avec la durée des segments. Considérer des segments de longueurs différentes permet également de prendre en compte différents types d'information acoustiques : le modèle ne peut capturer que des informations « bas niveau » (p.ex. phonémiques) dans les segments les plus courts et des informations de plus « haut niveau » (p.ex. prosodiques) ne sont accessibles que dans des segments plus longs.

Les résultats de la Table 1 montrent clairement que les représentations construites automatiquement par `wav2vec` encodent des informations sur la langue : les scores ABX sont systématiquement au dessus de 50% et deviennent même très bon dès que les segments considérés sont suffisamment longs (plus de 5 s). Ce résultat est d'autant plus intéressant que les deux langues considérées ne sont pas présentent dans le corpus d'apprentissage : les représentations multilingues « découvertes » par `wav2vec` sont donc *génériques* et capables d'extraire des informations pertinentes y compris sur de nouvelles langues.

**Mesure de la similarité entre langues** Dans une seconde expérience, nous reportons à la Table 2 les scores ABX obtenus lorsque l'on cherche à distinguer des segments de 20 secondes pour différentes paires de langues qui peuvent être formées à partir des 6 langues considérées.

Ces résultats (même si encore préliminaires !) confirment que les représentations multilingues apprises automatiquement par `wav2vec` permettent de distinguer des langues, même si celles-ci ne sont pas présentes dans le corpus d'apprentissage. De manière plus intéressante, les performances semblent effectivement liées à une similarité entre langues puisque les scores ABX entre langues sino-tibétaines sont plus faibles qu'entre une langue sino-tibétaine et une autre langue (cèmuhî, nashta) : la moyenne de tous les scores ABX entre les langues

| durée audio | score ABX |
|---|---|
| 50 ms | 59,2% |
| 1 s | 61,0% |
| 5 s | 82,2% |
| 10 s | 92,2% |
| 20 s | 94,3% |
| 40 s | 97,6% |
| 50 s | 97,0% |

TABLE 1 – Évolution du score ABX pour la paire japhug, na de Yongning en fonction de la durée de l'échantillon.

| | cèmuhî | nashta | na de Shekua | thulung | Japhug | na de Yongning |
|---|---|---|---|---|---|---|
| cèmuhî | — | 74,6% | 88,3% | 62,7% | 82,4% | 80,6% |
| nashta | — | — | 88,2% | 62,3% | 81,4% | 88,4% |
| na de Shekua | — | — | — | 79,8% | 87,5% | 87,2% |
| thulung | — | — | — | — | 79,8% | 76,2% |
| japhug | — | — | — | — | — | 94,3% |
| na de Yongning | — | — | — | — | — | — |

TABLE 2 – Score ABX pour différentes paires de langues. Les langues sino-tibétaine sont indiquées en orange, les autres langues en améthyste.

sino-tibétaine et les langues des autres familles est de $79,3\%$, alors qu'entre la moyenne de ces scores entre toutes les paires de langues sino-tibétaines est de $84,1\%$. Il est donc, de manière surprenante, plus difficile de distinguer les langues de deux familles différentes que les langues d'une même famille.

Pour corroborer les résultats de cette première expérience, nous nous sommes intéressés aux langues qianguiques, un sous-groupe de langues de la famille tibéto-birmane parlées en Chine, principalement sur le plateau de Qinghai. Les linguistes distinguent habituellement deux groupes de langues (Sims, 2016) : le qiang du Nord et le qiang du Sud. La collection Pangloss contient des enregistrements de 12 langues qiang, 5 identifiées comme des dialectes du qiang du Nord (luoduo, shuangliusuo, weicheng, waboliangzi et qugu) et 7 comme des dialectes du qiang du Sud (longxi, luobozhai, baishui, goukou, sanlong, heihu et baixi).

Nous avons considéré, pour chacune de ces 12 langues, 100 segments de 10 secondes, construit les représentations de ces segments à l'aide de `XSL-R` et mesuré la distance euclidienne entre toutes les paires de représentation. Nous avons alors calculé la moyenne des distances entre deux dialectes pour construire une matrice de distances entre dialectes et avons ensuite effectué un regroupement hiérarchique *(clustering)* de cette matrice à l'aide de l'algorithme du point le plus proche (Duda *et al.*, 2001) afin de construire un dendrogramme entre ces langues.

La hiérarchie et la matrice de similarité sont représentés à la Figure 1. Il apparaît clairement que les représentations construites par `XSL-R` permettent de retrouver la distinction entre les

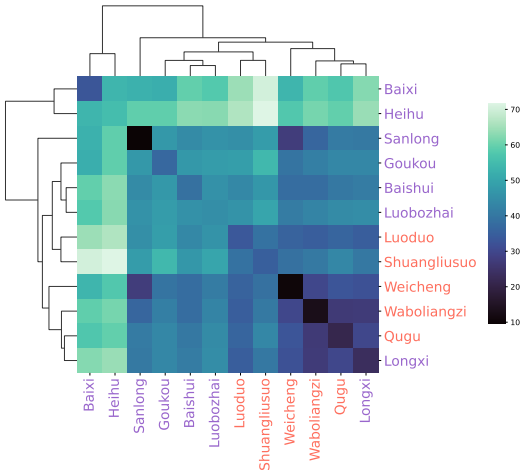


FIGURE 1 – Matrice de distances entre les langues qiang de la collection Pangloss et dendrogramme correspondant à la classification hiérarchique ascendante de celles-ci. Les dialectes du qiang du Nord sont représentés en orange et ceux du qiang du Sud en améthyste. Les dialectes ont été ordonnés automatiquement par similarité.

langues du qiang du Nord et celles appartenant au groupe du qiang du Sud (à une exception près). Ces représentations capturent donc bien des informations liées à la typologie des langues.

# 4 Conclusion

Les premiers résultats que nous présentons dans ce travail sont encourageants : s'ils soulèvent de nombreuses questions, ils ouvrent également la porte à de nombreuses perspectives particulièrement intéressantes. Ces résultats préliminaires devront toutefois être confirmés notamment pour garantir que la capacité du modèle à distinguer les langues repose bien sur des caractéristiques « linguistiques » et non uniquement sur des facteurs confondants (différences dans la manière dont les enregistrements ont été réalisés, locuteurs de genre différents, style de parole...).

# Remerciements

Un grand merci à Alexis Michaud pour avoir initié ce travail et pour son soutien constant. Ses nombreux commentaires tant sur le fond que sur la forme ont grandement amélioré ce travail.

Nous remercions le CNRS/TGIR HUMA-NUM et le Centre de Calcul IN2P3 (Lyon - France) pour la fourniture des ressources informatiques et de traitement des données nécessaires à ce travail.

Ce travail a bénéficié du soutien financier de l'Agence Nationale de la Recherche (projet « La documentation computationnelle des langues à l'horizon 2025 » [ANR-19-CE38-0015-04] et Labex « Fondements empiriques de la linguistique » [ANR-10-LABX-0083]) ainsi que de l'Institut des Langues Rares (ILARA-EPHE).

Une partie importante des ressources linguistiques utilisées dans le présent travail a été collectée dans le cadre du projet « Corpus parallèles en langues himalayennes » [ANR-12-CORP-0006].

## Références

Anastasopoulos A., Cox C., Neubig G. & Cruz H. (2020). Endangered languages meet modern NLP. In *Proceedings of the 28th International Conference on Computational Linguistics : Tutorial Abstracts*, p. 39–45, Barcelona, Spain (Online) : International Committee for Computational Linguistics. doi : 10.18653/v1/2020.coling-tutorials.7.

Bredin H. (2017). TristouNet : Triplet loss for speaker turn embedding. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, New Orleans, United States.

Conneau A., Baevski A., Collobert R., Mohamed A. & Auli M. (2021). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, p. 2426–2430. doi : 10.21437/Interspeech.2021-329.

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 - 23rd INTERSPEECH Conference*, Incheon, South Korea. hal : hal-03830470.

Duda R. O., Hart P. E. & Stork D. G. (2001). *Pattern Classification*. New York : Wiley, 2 édition.

Foley B., Arnold J., Coto-Solano R., Durantin G., Ellison T. M., van Esch D., Heath S., Kratochvíl F., Maxwell-Smith Z., Nash D., Olsson O., Richards M., San N., Stoakes H., Thieberger N. & Wiles J. (2018). Building speech recognition systems for language documentation : The CoEDL endangered language pipeline and inference system. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2018)*.

Guillaume S., Wisniewski G., Galliot B., Nguyễn M.-C., Fily M., Jacques G. & Michaud A. (2022a). Plugging a neural phoneme recognizer into a simple language model : a workflow for low-resource settings. In *Interspeech 2022 - 23rd Annual Conference of the International Speech Communication Association*, Proceedings of Interspeech 2022, Incheon, South Korea. doi : 10.5281/zenodo.5521111, hal : halshs-03625581.

Guillaume S., Wisniewski G., Macaire C., Jacques G., Michaud A., Galliot B., Coavoux M., Rossato S., Nguyễn M.-C. & Fily M. (2022b). Les modèles préentraînés à l'épreuve des langues rares : expériences de reconnaissance de mots sur la langue japhug (sino-tibétain). In *JEP 2022 - 34e Journées d'Études sur la Parole*, Actes des 34e Journées d'Études sur la Parole (JEP2022), Noirmoutier, France. hal : halshs-03625580.

Hsu W.-N., Bolte B., Tsai Y.-H. H., Lakhotia K., Salakhutdinov R. & Mohamed A. (2021). Hubert : Self-supervised speech representation learning by masked prediction

of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, **29**, 3451–3460. DOI : 10.1109/TASLP.2021.3122291.

Macaire C., Wisniewski G., Guillaume S., Galliot B., Jacques G., Michaud A., Rossato S., Nguyễn 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.

Partanen N., Hämäläinen M. & Klooster T. (2020). Speech recognition for endangered and extinct samoyedic languages. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, p. 523–533, Hanoi, Vietnam : Association for Computational Linguistics.

Sims N. (2016). Towards a more comprehensive understanding of Qiang dialectology. *Language and Linguistics*, **17**(3), 351–381.