Annexes de l'article intitulé
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

Ce document présente les annexes de l'article intitulé *Mesure du niveau de proximité entre enregistrements audio et évaluation indirecte du niveau d'abstraction des représentations issues d'un grand modèle de langage*.

ABSTRACT

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

This document presents the appendices to the article untitled *Establishing degrees of closeness between audio recordings along different dimensions using large-scale cross-lingual models.*

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

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

A Métadonnées des expériences

La liste des métadonnées pour les expériences menées est fournie en Table 1 pour la série du *conte populaire*, Table 2 pour la série *répertoires de chansons* et Table 3 pour la série *phonétique*.

REC ID	Year	DUR (s)	MIC	ITV	Acoust.
V1	2006	518	Tab	out	ND
V2	2007	440	Tab	out	D
V3	2008	707	Tab	out	D
V4	2014	527	Hea	Na	D
V5	2014	423	Hea	Na	D
$V6_h$	2018	348	Hea	out	ND
$V6_t$	2018	348	Tab	out	ND
$V7_h$	2018	635	Hea	out	ND
$V7_t$	2018	635	Tab	out	ND

TABLE 1 – Métadonnées pour la série du *conte populaire*. MIC = microphone : casque (Hea) ou posé sur table (Tab) ; ITV = interviewer : *outsider* ou Na (local). Acoustique : *non-damped* (ND), ou *damped* (D).

REC ID	DUR (s)	% SONG
$S\text{-guqi}_1$	151	100
$S\text{-guqi}_2$	300	100
T-narrat	296	0
S-wmd	129	100
S+T-alili	194	49

TABLE 2 – Métadonnées pour la série *répertoires de chansons*, y compris la proportion de voix chantée dans l'enregistrement.

REC ID	DUR (s)	SPK	SESSION TYPE
AS_1	1567	AS (F)	Phonetic elicit.
AS_2	952	AS (F)	Phonetic elicit.
RS_1	681	RS (F)	Phonetic elicit.
RS_2	786	RS (F)	Phonetic elicit.
TLT	897	TLT (F)	Phonetic elicit.
AS_{Lex}	1216	AS (F)	Lexical elicit.

TABLE 3 – Métadonnées pour la série *phonétique*. SPK = locuteur ; (F) = *Female*. Toutes les données ont été collectées en 2019.

B Valeurs des moyennes et écarts-types obtenus sur deux types d'expériences avec deux durées d'extraits audio

La figure 1 montre les valeurs moyennes et l'écart type pour une comparaison entre les scores inter-enregistrements pour deux expériences (*série phonétique* et *conte populaire*) et les scores intra-enregistrement (*même enregistrement*), pour différentes quantité d'audio (durées des extraits) par vecteur. Pour toutes les durées d'extrait, le score ABX moyen inter-enregistrement est toujours significativement plus élevé que le score moyen intra-enregistrement, même pour la durée de 1 s. Cela montre que les tests ABX peuvent mesurer des différences dans ces expériences, que ce soit pour une durée de 1, 5, 10 et 20s. L'augmentation de la quantité d'audio par extrait a pour effet d'augmenter la variabilité des calculs de distances ABX ainsi que le score ABX en valeur absolue, pour du signal de parole, que les locuteurs diffèrent ou non.

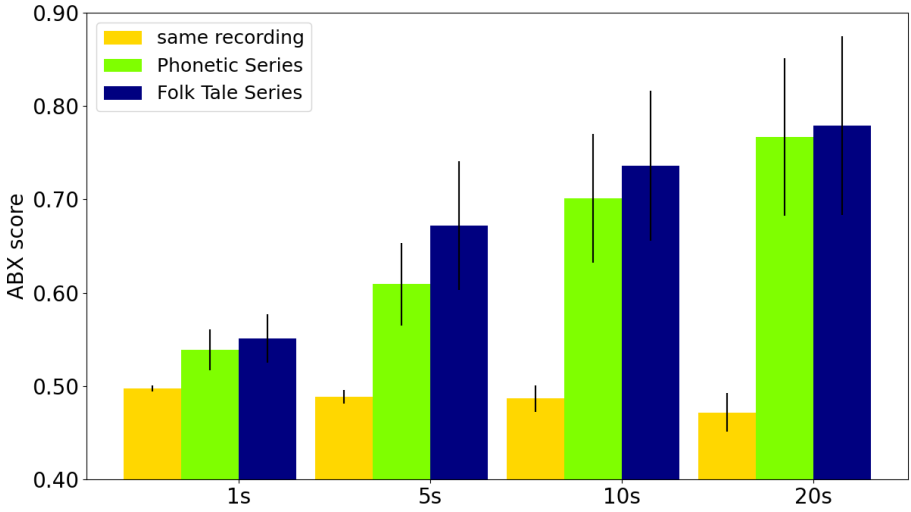

FIGURE 1 – Scores ABX moyens pour des extraits de 1, 5, 10, et 20 s.

C Effet de la couche sur les représentations vectorielles du modèle XLSR-53

Pour la *série du conte populaire*, nous avons effectué une analyse de l'effet du choix de la couche en sortie : les mêmes calculs que ceux décrits en section **??** ont été réalisés sur les 24 couches du modèles. Le résultat principal de cette étude de sensibilité est la très faible variabilité des scores ABX avec le numéro de couche (Figures 2 et 3). Ce résultat n'a été obtenu que pour des extraits de 10 s pour des raisons de temps calcul, mais il serait intéressant de vérifier si l'effet est nul pour des tailles d'audio de 1 s. Si cela était le cas, cela signifierait que ce qui est encapsulé sur 10 s d'audio, à savoir les variables extra-linguistiques, n'est pas sensible à la couche, tandis que les variables linguistiques (mieux repérées dans les extraits courts) le seraient.

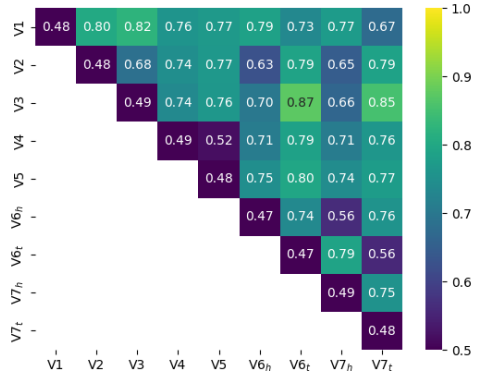

FIGURE 2 – Scores ABX moyens obtenus en faisant varier le numéro de couche (de 0 à 23), pour la série du conte populaire (taille des extraits : 10 s)

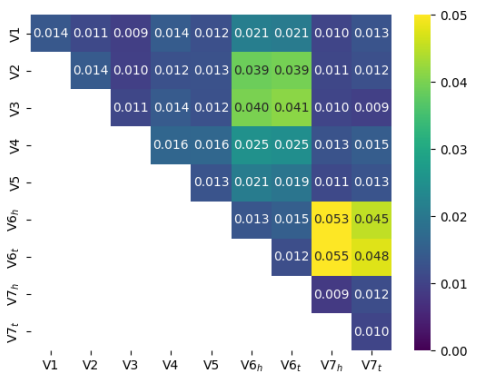

FIGURE 3 – écart type des scores ABX moyens obtenus en faisant varier le numéro de couche (de 0 à 23), pour la série du conte populaire (taille des extraits : 10 s)

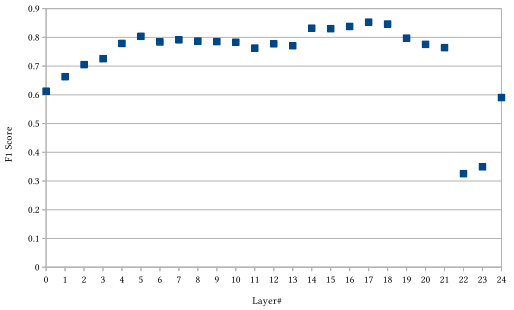

FIGURE 4 – Étude préliminaire de l'effet de la couche sur la performance en détection de langue par la méthode des sondes linguistiques (**?**).

D Scores ABX permettant de distinguer différentes versions de la série de *contes populaires*, par le même locuteur.

La valeur de 20 s pour la longueur d'extrait a été étudiée, et elle n'apporte pas beaucoup plus qu'une longueur de 10 s. En outre, une longueur de 20 s pour les extraits avec le max-pooling touche aux limites de la méthode du max-pooling. En effet, avec la méthode d'extraction max-pooling, chacun des 980 vecteurs avant pooling des 20 s d'audio n'occupera, en moyenne, que 1,04 cellule par vecteur final puisqu'il n'a que 1 024 composantes, ce qui ne donne qu'une composante en moyenne par vecteur final, ce qui est faible. Nous pensons qu'il y a une limite à la quantité d'audio que nous pouvons avoir dans un enregistrement, et ne voulons pas trop nous en approcher.

Les résultats sont visibles en figure 5 pour 20 s, figure 6 pour 10 s, figure 7 pour 5 s, et figure 8 pour les durées d'audio de 1 s par vecteur.

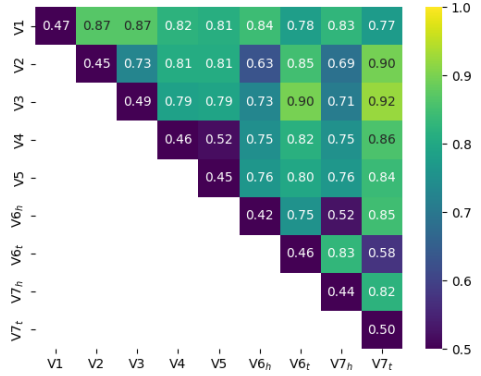

FIGURE 5 – Scores ABX pour la série du *conte populaire*. (Taille de l'extrait = 20 s).

FIGURE 6 – Scores ABX pour la série du *conte populaire*. (Taille de l'extrait = 10 s).

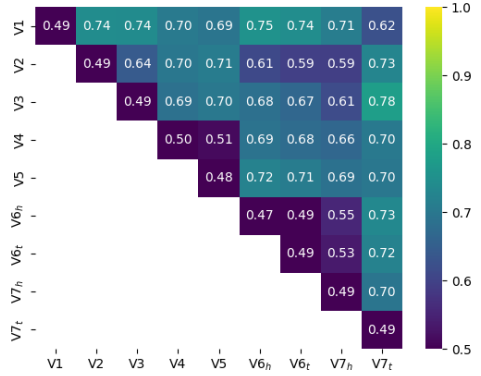

FIGURE 7 – Scores ABX pour la série du *conte populaire*. (Taille de l'extrait = 5 s).

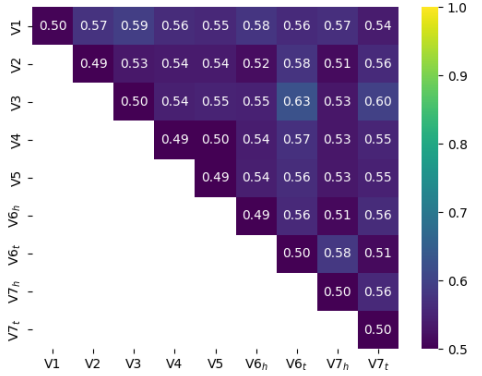

FIGURE 8 – Scores ABX pour la série du *conte populaire*. (Taille de l'extrait = 1 s).

E Scores ABX lors de l'expérience visant à distinguer les éléments de la série *phonétique*

Les résultats sont visibles en Figure 9 pour des extraits de 20 s, en Figure 10 pour des extraits de 10 s, Figure 11 pour des extraits de 5 s, et Figure 12 pour des extraits audio de 1 s par vecteur.

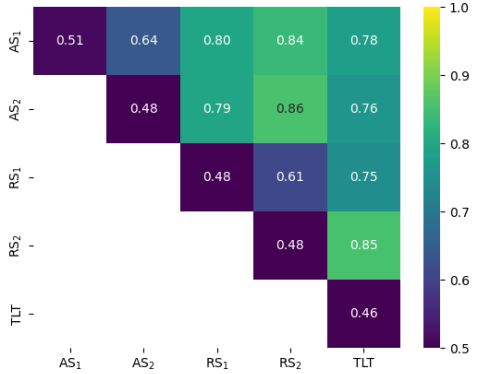

FIGURE 9 – Scores ABX pour la série *phonétique*. (Taille de l'extrait = 20 s).

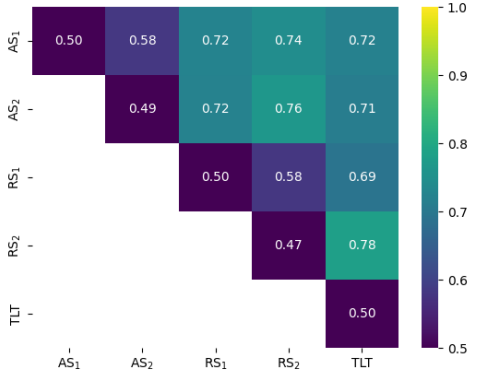

FIGURE 10 – Scores ABX pour la série *phonétique*. (Taille de l'extrait = 10 s).

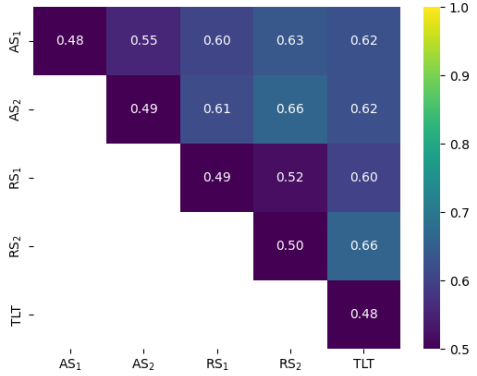

FIGURE 11 – Scores ABX pour la série *phonétique*. (Taille de l'extrait = 5 s).

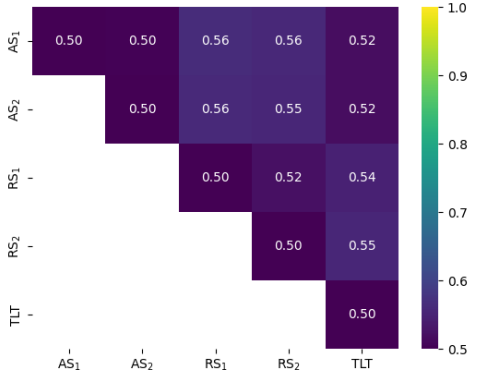

FIGURE 12 – Scores ABX pour la série *phonétique*. (Taille de l'extrait = 1 s).

F Ressources audio : liste des enregistrements utilisés pour l'étude, avec leurs DOI

Série du conte populaire :	
REC ID	DOI
V1	doi.org/10.24397/PANGLOSS-0004341
V2	doi.org/10.24397/PANGLOSS-0004343
V3	doi.org/10.24397/PANGLOSS-0004344
V4	doi.org/10.24397/pangloss-0004938
V5	doi.org/10.24397/pangloss-0004940
V6	doi.org/10.24397/pangloss-0007695
V7	doi.org/10.24397/pangloss-0007698

Série du répertoire de chansons :	
REC ID	DOI
S-guqi$_1$	doi.org/10.24397/pangloss-0004694
S-guqi$_2$	doi.org/10.24397/pangloss-0004697
T-narrat	doi.org/10.24397/pangloss-0004695
S-wmd	doi.org/10.24397/pangloss-0004698
S+T-alili	doi.org/10.24397/pangloss-0004699

Série phonétique	
REC ID	DOI
AS$_2$	doi.org/10.24397/pangloss-0008663
RS$_2$	doi.org/10.24397/pangloss-0008667
AS$_1$	doi.org/10.24397/pangloss-0008662
	doi.org/10.24397/pangloss-0008664
RS$_1$	doi.org/10.24397/pangloss-0008665
	doi.org/10.24397/pangloss-0008666
TLT	doi.org/10.24397/pangloss-0008668
	doi.org/10.24397/pangloss-0008669
AS$_{Lex}$	doi.org/10.24397/pangloss-0008670
	doi.org/10.24397/pangloss-0008671

TABLE 4 – Liste des DOIs pour les enregistrements de cette étude.

G Remerciements

Nous sommes reconnaissants aux communautés Na et Naxi et tenons à remercier tout particulièrement Mme Wang Sada et Mme Latami Dashilame pour le partage de leur expérience, leur générosité, leur confiance et leurs encouragements.

Cette recherche a été partiellement financée par le projet DIAGNOSTIC soutenu par l'Agence de l'Innovation de Défense (subvention n° 2022 65 007) et le projet DEEPTYPO soutenu par l'Agence Nationale de la Recherche (ANR-23-CE38-0003-01).

Nous remercions les sponsors des mission de terrain qui ont permis la collecte de données sur le naxi et le na (de 2002 à 2019). En particulier, nous souhaitons remercier le programme de mobilité internationale de l'IDEX de l'UGA qui a soutenu le travail de terrain sur na de Lataddi (Shekua) en 2019.