AUTOMATISATION DE LA SEGMENTATION POUR LA LINGUISTIQUE DOCUMENTAIRE :
UNE NOUVELLE ÉVALUATION DES CAPACITÉS MULTILINGUES DES MODÈLES NEURONAUX PRÉ-ENTRAINÉS DE LA PAROLE

Clara Rosina Fernandez Séverine Guillaume Guillaume Wisniewski

INTRODUCTION

La diarisation de locuteurs·trices vise à segmenter des enregistrements audio pour **identifier «qui a parlé quand»**. Les modèles actuels, basés sur des réseaux neuronaux, démontrent une capacité de généralisation au-delà des langues utilisée pour l'apprentissage.

Nos contributions :

- **Évaluation** d'un modèle de diarisation sur des langues à faibles ressources.
- **Développement d'outils** pour faciliter la linguistique documentaire.

LES MODÈLES PRÉ-ENTRAÎNÉS DE LA PAROLE SONT-ILS CAPABLES DE GÉNÉRALISER SUR DES LANGUES RARES ?

DONNÉES

Pangloss

- Enregistrements de **terrain**
- **12 langues** différentes : mandarin de Beijing, boomu, tibétain Commun, français d'Abidjan, hayu, koyi, limbu, népali, newar, tibétain de l'Amdo, turc de Chypre et arabe yéménite
- Durée totale : **7h18min**

MODÈLE

pyannote

- pyannote/speaker-diarization-3.1
- Apprentissage **supervisé** basé sur PyTorch
- **État de l'art** (sep 2023)
- Multilingue

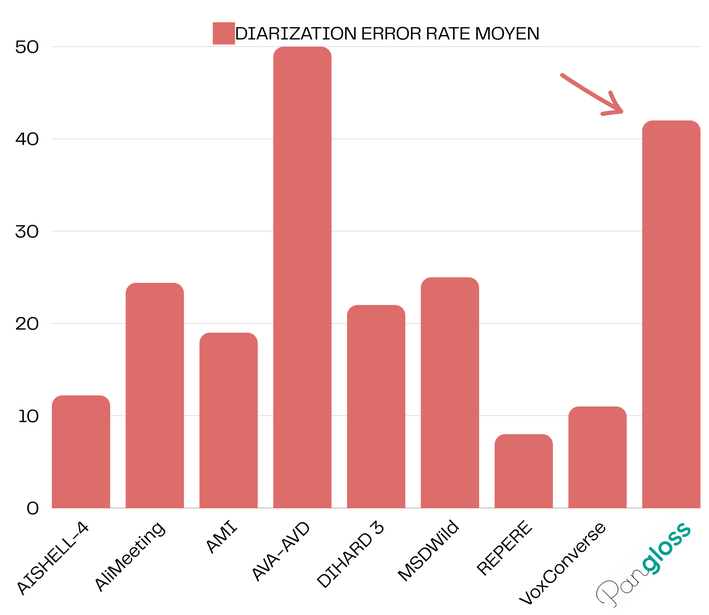
MÉTRIQUE

DIARIZATION ERROR RATE =

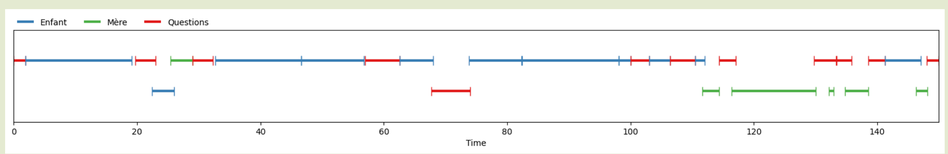
$$\frac{\text{FALSE ALARM} + \text{MISSED DETECTION} + \text{CONFUSION}}{\text{TOTAL}}$$

RÉSULTATS

Les DERs moyens obtenus par le modèle sur des langues «~usuelles~» varient entre 7,8% et 50,0% : des performances comparables à celles que nous observons. Sur **Pangloss**, les DERs variant entre 12,0% et 90% (moyenne : **42,12%**, médiane 36,98).



QUELLE PERFORMANCE ATTENDRE ?

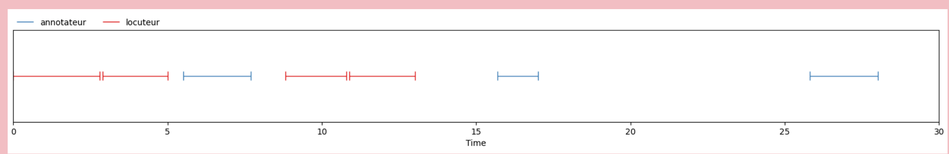


DER = 12%

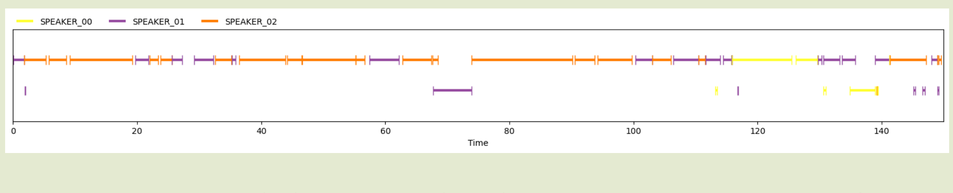
- Enregistrement de bonne qualité dans un environnement **calme**.
- Le **nombre de locuteurs** est correctement identifié.
- Les **tours de parole** principaux sont correctement identifiés.
- Les **chevauchements** sont souvent bien identifiés.

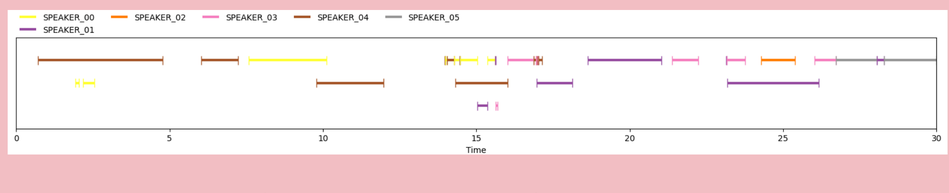
RÉFÉRENCE MANUELLE

PRÉDICTION AUTOMATIQUE



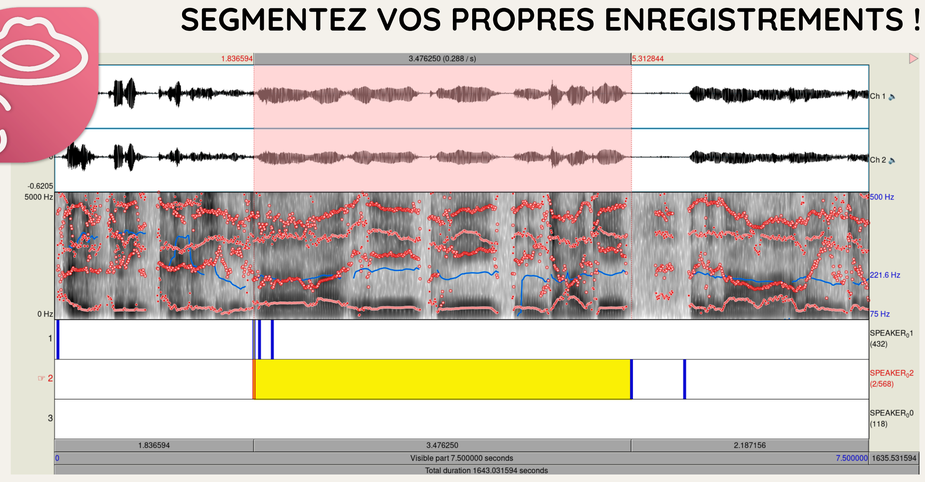
DER = 90%





- Enregistrement d'un environnement **bruyant.**
- Le **nombre de locuteurs** n'est pas correctement prédit.
- Les **tours de parole** principaux ne sont pas toujours bien délimités.
- Du **brouhaha** erronément perçu comme des chevauchements.

SEGMENTEZ VOS PROPRES ENREGISTREMENTS !



- Un fichier **.TextGrid segmenté**, prêt à être analysé, annoté, corrigé…

CONCLUSION

- Des **résultats comparables** à ceux obtenus sur les corpus d'évaluation en langues bien dotées et qui, de plus, sont présentes dans les données d'apprentissage.
- La capacité de généralisation peut donc être considérée comme suffisante pour l'**aide à la documentation** des langues rares.
- La **qualité et les conditions des enregistrements** ont un grand impact sur la performance du modèle.

Université Paris Cité
Laboratoire de linguistique formelle
cnrs
LACITO Langues et Civilisations à Tradition Orale