Automatisation de la segmentation pour la linguistique documentaire : une nouvelle évaluation des capacités multilingues des modèles neuronaux pré-entrainés de la parole

Clara Rosina Fernández[2] Séverine Guillaume[1] Guillaume Wisniewski[2]
(1) LACITO, CNRS, Université Sorbonne Nouvelle, F-94800, Villejuif, France
(2) LLF, CNRS, Université Paris-Cité, F-75013, Paris, France
c.rosinafernandez@gmail.com,
severine.guillaume@cnrs.fr,
guillaume.wisniewski@u-paris.fr

1 Introduction

La diarisation de locuteurs·trices est la tâche qui consiste à étiqueter « qui a parlé quand » à partir d'un enregistrement audio. Plus précisément, un modèle de diarisation est capable de prédire une segmentation comme celle de la figure 1 qui distingue les silences des périodes contenant de la parole et identifie, pour chaque segment de parole, les locuteurs·trices de celui-ci [1]. Les modèles de diarisation de l'état de l'art reposent sur des architectures neuronales capables d'apprendre des représentations abstraites du signal, suffisamment générales pour obtenir de bonnes performances dans de nombreuses langues. Cela ouvre la possibilité d'utiliser ces modèles pour segmenter automatiquement des enregistrements de langues rares et en cours de documentation.

Dans cet article, nous proposons de tester un modèle de diarisation état de l'art, sur des enregistrements de terrain issus de la collection Pangloss (Michailovsky *et al.*, 2014). Les corpus oraux recueillis lors de travaux de terrain présentent des caractéristiques linguistiques, souvent sous-représentées, voire absents, dans les jeux de données utilisés dans les travaux de TAL (Bender, 2009), notamment l'alternance codique fréquente et les dialectes non standard. Un enjeu scientifique majeur est de déterminer si des modèles neuronaux, entraînés et testés uniquement sur un petit nombre de langues pour lesquelles il existe de nombreuses ressources, sont suffisamment robustes pour pouvoir être appliqués, sans ajustement ou affinage supplémentaire, à de nouvelles langues.

En effet, si les modèles de diarisation ont été testés sur de nombreux jeux de données, incluant des données naturelles, des configurations d'enregistrement audio variées, des environnements acoustiques divers et des styles de discours différents, ceux-ci se concentrent

1. L'objectif précis de la diarisation est de déterminer si deux segments ont été prononcés par la même personne sans chercher à identifier cette personne — ce qui relève de la tâche d'*identification du locuteur*.

principalement sur des langues disposant de nombreuses ressources, telles que l'anglais, le mandarin et le français. Il est important de savoir si les architectures neuronales utilisées sont capables de capturer la diversité phonétique présente dans l'ensemble des langues, bien au-delà des quelques dizaines de langues considérés lors de leur apprentissage. Nos travaux permettent de mieux comprendre les capacités de généralisation des modèles multilingues, en variant leur cadre d'utilisation et d'évaluation habituels.

Ce travail apporte une seconde contribution en introduisant un nouvel outil [2] pour la linguistique documentaire. Cet outils permet de produire automatiquement, à partir d'un enregistrement audio, un fichier `TextGrid` Praat compatible avec ELAN contenant la segmentation des enregistrements, qui peut facilement être utilisé pour analyser ou transcrire ceux-ci. En détectant automatiquement les silences et les tours de parole, cet outil permet aux linguistes de se concentrer sur les parties « intéressantes » de leurs enregistrements. Ainsi, en annotant seulement celles-ci, le flux de travail est considérablement amélioré.

2 Expériences

Données Nous utilisons dans nos expériences des corpus issus de la collection Pangloss (Michailovsky *et al.*, 2014). Celle-ci contient aujourd'hui plus de 1 180 heures d'enregistrements dans 252 parlers différents, dont près de la moité sont annotés (notamment par des transcriptions ou des traductions). L'objectif de notre expérience est d'évaluer la diarisation automatique sur toutes les données de Pangloss dont les tours de parole sont annotés. Plus précisément, nous avons besoin d'enregistrements contenant plusieur·e·s locuteurs·trices, qui ont été segmentés en tour de parole et dont les locuteurs·trices sont identifié·e·s. En filtrant la collection Pangloss suivant ces trois critères nous avons pu constituer un corpus de 12 langues totalisant 7h 18min d'enregistrements : mandarin de Beijing, boomu, tibétain Commun, français d'Abidjan, hayu, koyi, limbu, nepali, newar, tibétain de l'Amdo, turc de Chypre et arabe yéménite. Les enregistrements comprenaient entre 2 et 5 locuteurs·trices. Les données, issues du travail de linguistique de terrain, incluent des enregistrements en extérieur, dans des environnements calmes ainsi que bruyants, des entretiens et des dialogues. La figure 1 donne un exemple d'un enregistrement de 150 s (extrait d'un entretien de 27 minutes en tibétain commun impliquant trois locuteurs·trices : deux femmes, désignées respectivement comme « Question » (la linguiste) et « Mère », et un enfant. Ces annotations ont été extraites automatiquement des transcriptions disponibles dans Pangloss.

Modèle Dans toutes nos expériences, nous avons utilisé, `pyannote.audio` [3] (Plaquet & Bredin, 2023; Bredin, 2023), un modèle de diarisation « sur étagère » qui peut être utilisé

2. Le code développé est accessible ici : https://github.com/rfclara/diarization.
3. Plus précisément, l'identifiant du modèle utilisé est `pyannote/speaker-diarization-3.1`

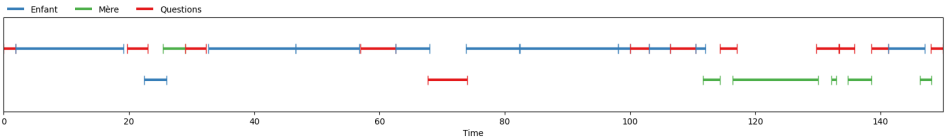

FIGURE 1 – Exemple d'annotation en tour de parole utilisé pour tester les systèmes de diarisation.

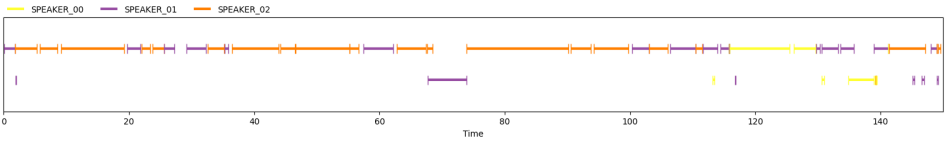

FIGURE 2 – Segmentation automatique du même enregistrement prédite par `pyannote.audio`

sans nécessiter de réglages supplémentaires [4]. Cet outils repose sur un réseau de neurones qui construite automatiquement des représentations *(embeddings)* les locteurs·trices adaptées à la tâche. Il est appris de manière supervisée à partir de corpus annotés multilingue issus des principales campagnes d'évaluation de la diarisation (Bredin, 2023). Ce modèle a prouvé son efficacité dans diverses compétitions, obtenant notamment, en 2022, la première place lors des challenges Ego4D (Grauman *et al.*, 2022) et Albayzin (Ortega *et al.*, 2022). Dans ce travail, nous nous intéressons à la performance de la pipeline pré-entraînée sur de nouvelles langues sans avoir procédé à un affinage préalable avant l'évaluation.

Les systèmes de diarisation de locuteurs·trices sont généralement évalués par le DER *(Diarization Error Rate)*. Il mesure le pourcentage d'erreurs commises en tenant compte des segments mal attribués aux locuteurs·trices, des omissions (voix non détectées) et des insertions (voix détectées à tort). Un DER plus bas indique une meilleure précision du modèle dans la tâche consistant à répondre à « qui parle quand ».

3 Résultats

Sur l'ensemble des enregistrements considérés `pyannote.audio` obtient un DER moyen de 42,1 % (médiane : 37,0 %). À titre de comparaison, la figure 3 représente l'évalaution du modèle à travers différents jeux de données. Les DERs obtenus par le modèle sur des langues « usuelles » varient entre 7,8 % et 50,0 %.

La collection Pangloss comporte des données très variées quant aux environnements d'enre-

4. `pyannote.audio` propose également une « recette » que les utilisateurs peuvent suivre pour affiner et améliorer les performances du modèle sur leurs propres jeux de données annotés manuellement

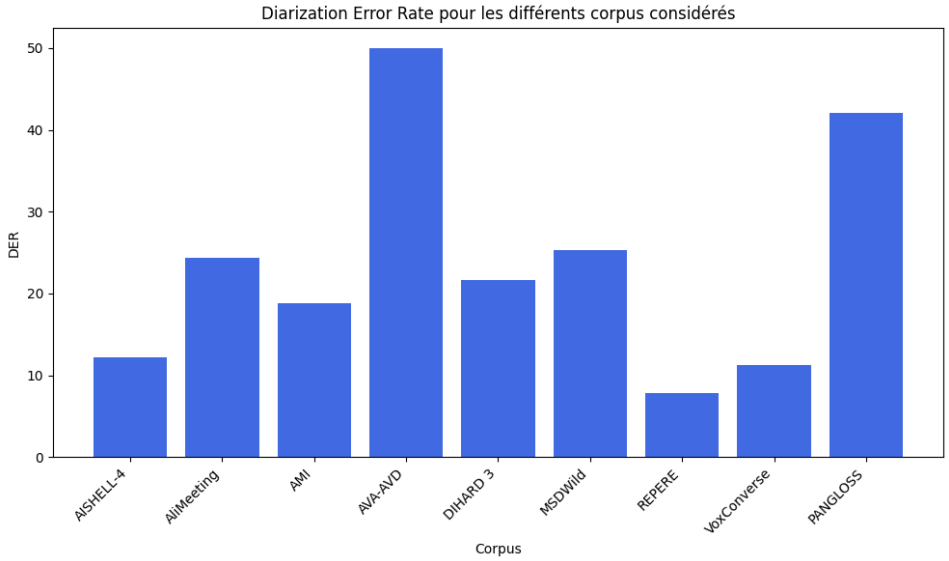

FIGURE 3 – Diarization Error Rate pour les différents corpus considérés.

gistrement et aux types de discours, ce qui peut expliquer des DERs variant entre 12,3 % et 90,7 % selon les enregistrement.

Afin de réaliser une étude plus qualitative, nous avons représenté, à la figure 2, la segmentation prédite par le modèle sur l'enregistrement utilisé comme exemple à la figure 1. Sur cet enregistrement, le modèle obtient un DER de 17,7 % et a correctement identifié le nombre de locuteurs·trices. La comparaison des deux segmentations montre que le modèle a été plus « sensible » aux silences que l'annotateur·trice, ce qui a entraîné des erreurs des faux-négatifs : là où l'annotateur·trice a segmenté l'extrait (150 s) en 29 segments, le modèle en a identifié 59. Plus généralement, sur la durée totale de l'enregistrement (27 mn), la référence comprend 392 segments et la prédiction 559. Si ce type d'erreur impacte le DER, il n'est pas forcément problématique pour l'utilisation des segmentations prédites.

La segmentation prédite présente également deux confusions de locuteurs·trices. Autour de la seconde 25, le changement de tour de parole n'a pas été détecté. Après correspondance manuelle, nous avons constaté que la confusion s'est produite entre les locutrices « Mère » et « Questions » (peut-être parce qu'elles sont de même genre). Une confusion similaire est également présente à la fin de l'extrait. Il est intéressant de noter que, dans cet exemple, il n'y a pas de faux positifs (cas où la prédiction détecte de la parole alors que la référence indique un silence). Si ce type d'erreur est absent dans cet enregistrement de bonne qualité, il est plus fréquent dans des enregistrements en extérieur ou avec plus de nuisances sonores.

Une analyse des enregistrements pour lesquels le DER est élevé, comme ceux en français d'Abidjan et arabe yéménite (dialecte de Sanaa) dont les DERs atteignent jusqu'à 90,7%,

montrent que ceux-ci comportent des chevauchements fréquents (locuteurs·trices parlant en même temps) et une mauvaise qualité audio. Ces difficultés sont connues pour rendre la diarisation particulièrement difficile : les DERs obetnus par `pyannote.audio` dans la campagne *AVA-AVD Audio-Visual Speaker Diarization in the Wild* (Xu *et al.*, 2022) visant à tester les systèmes de diarisation dans des conditions réelles dépassent également les 50,0 %.

4 Conclusion

Dans ce travail, nous avons utilisé un modèle de diarisation à l'état de l'art pour segmenter automatiquement des enregistrements de langues rares issues de la collection Pangloss. Les performances que nous avons obtenues varient davantage entre les différents enregistrements qu'entre les langues, indiquant que le DER est fortement impacté par la qualité audio, mais aussi par la « rigueur » des annotateurs·trices, notamment en ce qui concerne la segmentation des silences.

De manière générale, nos résultats sur les langues de la collection Pangloss, avec un DER moyen de 42,1 %, sont comparables aux résultats obtenus sur les corpus d'évaluation de diarisation habituels, qui ne contiennent que des langues à fortes ressources également présentes dans les données d'apprentissage. La capacité de généralisation peut donc être considérée comme suffisante pour l'aide à la documentation des langues en danger. Ainsi, l'outil que nous avons développé permet d'intégrer la diarisation automatique dans le flux de travail actuel des linguistes.

De manière plus générale, nous envisageons d'utiliser cet outil pour aider à enrichir des ressources existantes qui n'ont pas été segmentées manuellement, comme c'est le cas de plus de 200 corpus de la collection Pangloss.

5 Remerciements

Ces travaux ont été partiellement financés par le projet DIAGNOSTIC soutenu par l'*Agence d'Innovation de Défense* (contrat n° 2022 65 007) et le projet DEEPTYPO soutenu par l'*Agence Nationale de la Recherche* (ANR-23-CE38-0003-01).

Références

BENDER E. M. (2009). Linguistically naïve != language independent : Why NLP needs linguistic typology. In T. BALDWIN & V. KORDONI, Éds., *Proceedings of the EACL 2009*

Workshop on the Interaction between Linguistics and Computational Linguistics : Virtuous, Vicious or Vacuous ?, p. 26–32, Athens, Greece : Association for Computational Linguistics.

BREDIN H. (2023). pyannote.audio 2.1 speaker diarization pipeline : principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

GRAUMAN K., WESTBURY A., BYRNE E., CHAVIS Z., FURNARI A., GIRDHAR R., HAMBURGER J., JIANG H., LIU M., LIU X., MARTIN M., NAGARAJAN T., RADOSAVOVIC I., RAMAKRISHNAN S. K., RYAN F., SHARMA J., WRAY M., XU M., XU E. Z., ZHAO C., BANSAL S., BATRA D., CARTILLIER V., CRANE S., DO T., DOULATY M., ERAPALLI A., FEICHTENHOFER C., FRAGOMENI A., FU Q., GEBRESELASIE A., GONZÁLEZ C., HILLIS J., HUANG X., HUANG Y., JIA W., KHOO W., KOLÁŘ J., KOTTUR S., KUMAR A., LANDINI F., LI C., LI Y., LI Z., MANGALAM K., MODHUGU R., MUNRO J., MURRELL T., NISHIYASU T., PRICE W., RUIZ P., RAMAZANOVA M., SARI L., SOMASUNDARAM K., SOUTHERLAND A., SUGANO Y., TAO R., VO M., WANG Y., WU X., YAGI T., ZHAO Z., ZHU Y., ARBELÁEZ P., CRANDALL D., DAMEN D., FARINELLA G. M., FUEGEN C., GHANEM B., ITHAPU V. K., JAWAHAR C. V., JOO H., KITANI K., LI H., NEWCOMBE R., OLIVA A., PARK H. S., REHG J. M., SATO Y., SHI J., SHOU M. Z., TORRALBA A., TORRESANI L., YAN M. & MALIK J. (2022). Ego4d : Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 18995–19012.

MICHAILOVSKY B., MAZAUDON M., MICHAUD A., GUILLAUME S., FRANÇOIS A. & ADAMOU E. (2014). Documenting and researching endangered languages : the Pangloss Collection. *Language Documentation & Conservation*, **8**, 119–135. HAL : halshs-01003734.

ORTEGA A., MIGUEL A., LLEIDA E., BAZÀN V., PÉREZ C. & DE PRADA A. (2022). Albayzin evaluation iberspeech-rtve 2022 speaker diarization and identity assignment.

PLAQUET A. & BREDIN H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

XU E. Z., SONG Z., TSUTSUI S., FENG C., YE M. & SHOU M. Z. (2022). Ava-avd : Audio-visual speaker diarization in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, p. 3838–3847, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3503161.3548027.