Détection de l'activité vocale dans des corpus audiovisuels à l'aide de représentations auto-supervisées

Stage de fin d'études d'Ingénieur ou de Master 2 – Année académique 2023-2024

Mots clés : deep learning, machine learning, self supervised models, voice activity detection, speech activity detection, wav2vec 2.0

Contexte

L'Institut National de l'Audiovisuel (INA) est un établissement public à caractère industriel et commercial (EPIC), dont la mission principale consiste à sauvegarder et promouvoir le patrimoine audiovisuel français à travers la vente d'archives et la gestion du dépôt légal. À ce titre, l'Institut capte en continu 180 chaînes de télévision et radio et stocke plus de 25 millions d'heures de contenu audiovisuel. L'INA assure également des missions de formation, de production et de recherche scientifique.

Le service de la recherche de l'INA mène depuis plus de 20 ans des travaux de recherche dans le domaine de l'indexation et de la description automatique de ces fonds selon l'ensemble des modalités : textes, sons et images. Le service participe à de nombreux projets collaboratifs de recherche que ce soit dans un cadre national et européen et accueille des stages de Master ainsi que des doctorants en co-encadrement avec des laboratoires nationaux d'excellence.

Ce stage est proposé au sein de l'équipe de recherche (https://recherche.ina.fr) et se place dans le cadre d'un projet collaboratif financé par l'ANR : Gender Equality Monitor (GEM).

D'autres sujets de stage sont également proposés dans l'équipe :
https://www.ina.fr/institut-national-audiovisuel/equipe-recherche/stages

Objectifs du stage

La détection d'activité vocale (Voice Activity Detection - VAD) est une tâche d'analyse audio qui vise à identifier les portions d'enregistrement contenant de la parole humaine, les distinguant des autres parties du signal contenant du silence, des bruits de fond ou de la musique. Souvent considérée comme un prétraitement, cette méthode utilisée en amont des tâches de reconnaissance automatique de la parole, des locuteurs ou des émotions.

Si les outils VAD existants permettent d'obtenir d'excellents résultats sur les programmes d'information ou les émissions de plateau [Dou18a, Bre23], les recherches récentes menées à l'INA ont révélé que les performances des systèmes état-de-l'art sont moindres pour un grand nombre de matériaux peu représentés dans les corpus de parole annotés. Ces contenus, qui ont fait l'objet d'une campagne d'annotation interne, incluent des émissions musicales, des dessins animés, du sport, des fictions, des jeux télévisés et des documentaires.

L'objectif du stage est de développer des modèles de détection d'activité vocale (VAD) en adoptant une approche fondée sur le paradigme d'apprentissage auto-supervisé et s'appuyant sur les architectures *transformers* telles que wav2vec 2.0 [Bae20]. Les modèles basés sur ces architectures permettent d'obtenir des résultats état de l'art sur de nombreuses tâches de traitement de la parole à l'aide de quantités d'exemples annotés limitées : transcription, compréhension, traduction, détection d'émotions, reconnaissance de locuteur, détection du langage, etc [Li22, Huh23, Par23].
Plusieurs études récentes ont démontré l'efficacité des approches auto-supervisées pour la VAD [Gim21, Kun23], mais ont à ce jour été entraînées et évaluées sur des données ne reflétant pas la diversité des contenus audiovisuels. Le stage proposé vise à exploiter les millions d'heures de contenu audiovisuel conservés à l'INA pour l'entraînement et l'amélioration des modèles.

Les modèles réalisés seront intégrés au logiciel open-source inaSpeechSegmenter, utilisé entre autres pour le décompte du temps de parole des femmes et des hommes dans les programmes à des fins de recherche ou de régulation du paysage audiovisuel [Dou18b, Arc23].

Valorisation du stage

Différentes stratégies de valorisation des travaux seront envisagées, en fonction de leur degré de maturité et des orientations envisagées pour la suite des travaux :

- Diffusion des modèles réalisés sous licence open-source sur HuggingFace et/ou le dépôt Github de l'INA : https://github.com/ina-foss
- Rédaction de publications scientifiques

Conditions du stage

Le stage se déroulera sur une période de 4 à 6 mois, au sein du service de la Recherche de l'Ina. Il aura lieu sur le site Bry 2, situé au 28 Avenue des frères Lumière, 94360 Bry-sur-Marne. La·le stagiaire sera encadré·e par Valentin Pelloin et David Doukhan. Un ordinateur équipé d'un GPU sera fourni ainsi qu'un accès au cluster de calcul de l'Institut.

Gratification : 760 € brut / mois + 50 % pass navigo
Télétravail : possible une journée par semaine

Contact

Pour soumettre votre candidature à ce stage, ou pour solliciter davantage d'informations, nous vous invitons à envoyer votre CV et votre lettre de motivation par e-mail aux adresses suivantes : vpelloin@ina.fr et ddoukhan@ina.fr.

Profil recherché

- Étudiant·e en dernière année d'un bac +5 dans le domaine de l'informatique et de l'IA
- Forte appétence pour la recherche académique
- Intérêt pour le traitement automatique de la parole
- Maîtrise de Python et expérience dans l'utilisation de bibliothèques de ML
- Capacité à effectuer des recherches bibliographiques
- Rigueur, Synthèse, Autonomie, Capacité à travailler en équipe

Bibliographie

[Arc23] ARCOM (2023). “La représentation des femmes à la télévision et à la radio - Rapport sur l'exercice 2022” [en ligne].

[Bae20] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” Neural Information Processing Systems, Jun. 2020.

[Bre23] Bredin, H. (2023). pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe, in INTERSPEECH 2023, ISCA, pp. 1983–1987.

[Dou18a] Doukhan, D., Carrive, J., Vallet, F., Larcher, A., & Meignier, S. (2018, April). An open-source speaker gender detection framework for monitoring gender equality. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5214-5218). IEEE.

[Dou18b] Doukhan, D., Poels, G., Rezgui, Z., & Carrive, J. (2018). Describing gender equality in french audiovisual streams with a deep learning approach. VIEW Journal of European Television History and Culture, 7(14), 103-122.

[Gim21] P. Gimeno, A. Ortega, A. Miguel, and E. Lleida, “Unsupervised Representation Learning for Speech Activity Detection in the Fearless Steps Challenge 2021,” in Interspeech 2021, ISCA, Aug. 2021, pp. 4359–4363.

[Huh23] Huh, J., Brown, A., Jung, J. W., Chung, J. S., Nagrani, A., Garcia-Romero, D., & Zisserman, A. (2023). Voxsrc 2022: The fourth voxceleb speaker recognition challenge. arXiv preprint arXiv:2302.10248.

[Kun23] M. Kunešová and Z. Zajíc, “Multitask Detection of Speaker Changes, Overlapping Speech and Voice Activity Using wav2vec 2.0,” in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Jun. 2023, pp. 1–5.

[Li22] Li, M., Xia, Y., & Lin, F. (2022, December). Incorporating VAD into ASR System by Multi-task Learning. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)* (pp. 160-164). IEEE.

[Par23] Parcollet, T., Nguyen, H., Evain, S., Boito, M. Z., Pupier, A., Mdhaffar, S., ... & Besacier, L. (2023). LeBenchmark 2.0: a Standardized, Replicable and Enhanced Framework for Self-supervised Representations of French Speech. *arXiv preprint arXiv:2309.05472*.