Institut de Recherche
en Informatique de Toulouse

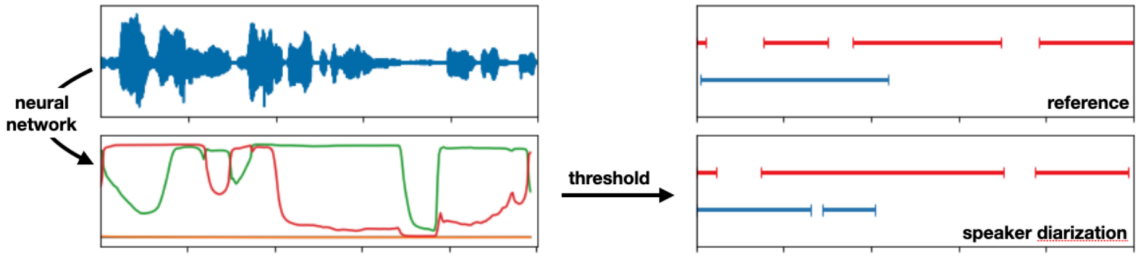
Stage Master 2 / Ingénieur 3ème année

Représentations neuronales auto-supervisées de parole pour reconnaître “qui parle quand ?” (*speaker diarization)*

Mots clés : traitement automatique de la parole, speaker diarization

Dans le cadre du projet **SeRiouSLy:** Segmentation, Regroupement, et Séparation des Locuteurs, nous proposons un stage pour étudier les représentations du signal de parole, issues de réseaux de neurones entraînés à des tâches dites auto-supervisées, très utilisées depuis récemment en reconnaissance automatique de la parole.

Qu’est-ce que la *speaker diarization ?* La “diarisation” du locuteur consiste à partitionner un flux audio en segments temporels homogènes en fonction de l'identité du locuteur. Les méthodes actuelles sont fondées sur un réseau de neurones dit de bout-en-bout, qui ingère l'enregistrement audio et sort directement la sortie de la diarisation (qui tient compte de chevauchements éventuels), comme représenté sur la figure ci-dssous.



Depuis quelques années, de nouvelles méthodes d’apprentissage de représentations de données audio de parole ont émergé récemment. Elles reposent sur un apprentissage dit auto-supervisé (*self-supervised)*, où un modèle de plusieurs centaines de millions de paramètres, voire plus, est entraîné à prédire le contenu d’une trame acoustique masquée. L’un des premiers modèles a été Wav2Vec [Schneider, 2019], et dès lors, des améliorations (Wav2Vec2, Wav2Vec2 XLS) ainsi que de nombreux autres modèles plus ou moins similaires ont été proposés, par exemple HuBERT [Hsu, 2021], WavLM [Chen 2022]. Les représentations qu’ils génèrent sont données en

entrée à d'autres modèles qui vont être spécialisés dans une tâche précise (reconnaissance de la parole, détection d'émotions dans la voix, etc.).

L'objet de ce stage est d'étudier ces représentations pour la tâche de segmentation et regroupement en locuteur (SRL). Il existe très peu de travaux reposant sur des représentations auto-supervisées. Une exception récente est WavLM [Chen 2022] qui modélise explicitement les locuteurs dans sa fonction de coût et montre que ceci est bénéfique pour les tâches de vérification du locuteur, de SRL.

En pratique, les recherches menées dans ce stage contribuera à la toolbox open source pyannote.audio (basée sur PyTorch), développée par Hervé Bredin, qui co-supervisera le stage. Les expériences feront bon usage du superordinateur Jean Zay.

Veuillez envoyer votre candidature (CV, notes, éventuellement des recommandations ou contacts de professeurs pouvant vous recommander) :
- à thomas.pellegrini@irit.fr
- avec pour objet "Stage self-sup diarisation".

Lieu : France, Toulouse, IRIT, équipe SAMoVA.
Date et durée : 5 à 6 mois.
Indemnité : environ 600 euros mensuels

Bibliographie

[Park 2022] Tae Jin Park et al. A Review of Speaker Diarization: Recent Advances with Deep Learning. Computer Speech & Language, 2022.

[Chen 2022] Sanyuan Chen et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. Journal of Selected Topics in Signal Processing

Framework pyannote.audio
https://github.com/pyannote/pyannote-audio