



Offre n°2024-07314

## Doctorant F/H Détection et clustering de la langue parlée

Type de contrat : CDD

Niveau de diplôme exigé : Bac + 5 ou équivalent

Fonction : Doctorant

### Contexte et atouts du poste

Inria Défense&Sécurité (Inria D&S) a été créé en 2020 pour fédérer les actions d'Inria répondant aux besoins numériques des forces armées et forces de l'intérieur. La thèse sera réalisée au sein de l'équipe de recherche en traitement de l'audio de Inria D&S, sous la direction de Jean-François Bonastre et co-encadrée par Raphaël Duroselle.

La thèse s'inscrit dans un projet visant au profilage vocal explicable et frugal. Le profilage vocal consiste à extraire des informations d'un enregistrement audio comme l'identité, la langue parlée, l'âge, l'origine géographique et ethnique, ou encore des marques socio/patho/physiologiques dans la voix. L'objectif de ce projet est d'apporter une explicabilité aux systèmes de profilage vocal sans perte de performance. L'explicabilité permet de conserver les opérateurs au centre du processus, en leur donnant les moyens d'une décision instruite.

### Mission confiée

#### Approche

L'approche envisagée pour la thèse repose sur la définition d'un jeu d'attributs vocaux génériques partagés par des groupes individus. Seule la présence ou l'absence d'un attribut dans un extrait vocal donné est utilisée pour prendre la décision, menant à une représentation binaire. Cette approche a été introduite pour la tâche de vérification du locuteur [1,2].

La thèse proposée vise à développer cette méthodologie en l'abordant selon l'objectif d'analyse de la langue parlée [3]. Le système vise à regrouper ensemble les segments relevant de la même langue et de détecter si celle-ci fait partie d'un panel de langues connues ou s'il s'agit d'une langue inconnue. Dans ce dernier cas, la proximité avec les langues connues devra être explicitement proposée, sur la base des attributs connus par le système.

Depuis l'apparition des modèles iVector [4] (initialement pour la reconnaissance du locuteur) dans la détection de la langue, le schéma général a peu évolué dans ce domaine. Il s'agit toujours de proposer un extracteur appris sur une grande masse de données et capable de représenter une séquence acoustique de durée quelconque par un vecteur de taille fixe, concentrant la variabilité utile à la tâche visée. Ensuite des classifieurs 1:1, comparant deux langues, ou 1:N, comparant N langues sont construits et un système de prise de décision, dit « back-end », se base sur ces classifieurs pour répondre aux diverses tâches visées. Les réseaux de neurones, comme les « bottleneck features » ont permis d'intégrer très bas (proche du niveau acoustique) des éléments de plus haut niveau, allant jusqu'aux modèles de langage, apportant un gain très significatif [5]. Puis les embeddings issus de modèles neuronaux, dit « xVector », ont remplacé les iVector et permis à la fois d'augmenter la taille des modèles (et la performance) et de simplifier l'apprentissage, avec un procédé unique réalisant la transformation d'une séquence acoustique de taille variable en un vecteur signifiant de taille contenue [6].

Plus récemment, l'usage des modèles pré-appris comme WavLM [7] ou MMS [8] a été étudié [9]. Par leur généralité, ces modèles permettent des gains intéressants, surtout quand peu de données sont disponibles dans la base d'entraînement pour certaines langues, au prix d'un accroissement important de la complexité en termes de nombres de paramètres.

Ces approches partagent des limitations communes : elles sont peu capables d'expliquer leur décision, les performances se dégradent très significativement quand le contexte d'utilisation s'éloigne du contexte d'apprentissage, les performances sont très variables suivant les couples de langues ou dialectes considérés, elles gèrent mal le déséquilibre entre les quantités de données d'apprentissage disponibles par langue et elles sont lourdes à adapter/réapprendre. Enfin, elles ne proposent rien ou peu dans le cas de langues inconnues.

Dans ce projet, nous proposons de partir de l'état de l'art puis d'adapter l'approche par attribut de voix au contexte de la détection de la langue parlée. Dans cette adaptation, une langue peut être représentée par un vecteur binaire correspondant à la présence/absence d'attributs dans cette langue ou par un

vecteur scalaire, indiquant la fréquence des attributs dans la langue. Les attributs eux-mêmes peuvent intégrer des informations de plus haut niveau, comme les niveaux phonotactiques et linguistiques). Cette architecture offre la possibilité de reconnaître une langue inconnue (au sens qu'aucune donnée correspondant à cette langue n'est présente dans la base d'apprentissage) et de situer sa proximité avec les langues connues en termes d'attributs explicites, permettant d'exploiter des connaissances en géolinguistique, par exemple. Un modèle de langue peut ainsi être construit dès le premier exemple de cette langue disponible, puis être adapté sans coût de calcul à chaque arrivée d'un exemple complémentaire. Si nécessaire, l'extracteur d'attributs peut être adapté en ajoutant un ou plusieurs attributs à partir des nouvelles données, sans nécessité de manière obligatoire de réapprendre la totalité du modèle. Les gains espérés sont donc importants, au niveau de l'explicitabilité, du traitement des langues inconnues et de l'adaptation au contexte.

## Objectifs

1. Appliquer l'approche par attribut décrite précédemment à la détection de la langue parlée ;
2. Développer la capacité à apprendre ou à étendre (nouvelle langue, nouveaux attributs) les modèles à partir de données peu ou pas annotées (par exemple, des données où seule la région d'enregistrement est connue) en optimisant le ratio « quantité de données/qualité des informations sur ces données » ;
3. Explorer la capacité de cette approche à renseigner sur des langues inconnues ;

Exploiter l'approche pour le regroupement en langues de documents audio, même quand tout ou partie des langues sont inconnues du système, incluant donc la découverte et la caractérisation de langues inconnues.

## Principales activités

- Etat de l'art, entraînement et évaluation de systèmes de reconnaissance de la langue parlée ;
- Deep learning, et notamment utilisation et adaptation de modèles pré-entraînés de traitement de l'audio, comme WavLM [7] ou MMS [8] ;
- Apprentissage semi-supervisé ;

Travail sur l'explicitabilité post-hoc de modèles de reconnaissance de la langue.

## Compétences

Compétences et connaissances souhaitées :

- Master 2 ou diplôme d'école d'ingénieur en informatique, mathématiques appliquées ou phonétique,
- Intérêt marqué pour la recherche appliquée,
- Maîtrise de l'anglais parlé et écrit,
- Connaissances en traitement du signal,
- Connaissances en apprentissage automatique de manière générale et dans les approches neuronales (deep learning) en particulier,
- Connaissance pratique d'outils comme Pytorch, Keras ou Scikit-learn,
- Expérience en traitement automatique de la parole, dont la connaissance de plateformes open-source comme Kaldi ou Speechb

## Références

- [1] Ben-Amor, I., & Bonastre, J. F. (2022, April). BA-LR: Binary-Attribute-based Likelihood Ratio estimation for forensic voice comparison. In 2022 International workshop on biometrics and forensics (IWBIF) (pp. 1-6). IEEE.
- [2] Ben-Amor, I., Bonastre, J. F., O'Brien, B., & Bousquet, P. M. (2023, August). Describing the phonetics in the underlying speech attributes for deep and interpretable speaker recognition. In Interspeech 2023.
- [3] Li, Haizhou, Bin Ma, et Kong Aik Lee. « Spoken Language Recognition: From Fundamentals to Practice ». *Proceedings of the IEEE* 101, n° 5 (mai 2013): 1136-1159. <https://doi.org/10.1109/JPROC.2012.2237151>.
- [4] Dehak, Najim, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, et Pierre Ouellet. « Front-End Factor Analysis for Speaker Verification ». *IEEE Transactions on Audio, Speech, and Language Processing* 19, n° 4 (mai 2011): 788-798. <https://doi.org/10.1109/TASL.2010.2064307>.
- [5] Fér, Radek, Pavel Matějka, František Grézl, Oldřich Plchot, Karel Veselý, et Jan Honza Černocký. « Multilingually trained bottleneck features in spoken language recognition ». *Computer Speech & Language* 46 (1 novembre 2017): 252-267. <https://doi.org/10.1016/j.csl.2017.06.008>.
- [6] Snyder, David, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Daniel Povey, et Sanjeev Khudanpur. « Spoken Language Recognition Using X-Vectors ». In *The Speaker and Language Recognition Workshop (Odyssey 2018)*, 105-111. ISCA, 2018. <https://doi.org/10.21437/Odyssey2018-15>.
- [7] Chen, Sanyuan, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, et al. « WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing ». *IEEE Journal of Selected Topics in Signal Processing* 16, no 6 (octobre 2022): 1505-1518. <https://doi.org/10.1109/JSTSP.2022.3188113>.
- [8] Pratap, Vineel, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, et al. 2023. « Scaling Speech Technology to 1,000+ Languages ». arXiv. <http://arxiv.org/abs/2305.13516>.
- [9] Alumäe, Tanel, et Kunnar Kukk. 2022. « Pretraining Approaches for Spoken Language Recognition: TalTech Submission to the OLR 2021 Challenge ». arXiv. <http://arxiv.org/abs/2205.07083>.

## Avantages

- Restauration subventionnée,
- Transports publics remboursés partiellement,
- Congés: 7 semaines de congés annuels + 10 jours de RTT (base temps plein) + possibilité d'autorisations d'absence exceptionnelle (ex : enfants malades, déménagement),
- Possibilité de télétravail (2 jours par semaine) et aménagement du temps de travail,
- Équipements professionnels à disposition (visioconférence, prêts de matériels informatiques, etc.),
- Prestations sociales, culturelles et sportives (Association de gestion des œuvres sociales d'Inria),
- Accès à la formation professionnelle,

## Rémunération

Année 1 & 2 = 2082 € bruts mensuels

Année 3 = 2190 € bruts mensuels

## Informations générales

- Ville : PARIS
- Centre Inria : [Siège](#)
- Date de prise de fonction souhaitée : 2024-05-01
- Durée de contrat : 3 ans
- Date limite pour postuler : 2024-09-01

## Contacts

- Équipe Inria : MIS-DEFENSE (DIRECTION)
- Directeur de thèse :  
Maillet Florence / [florence.maillet@inria.fr](mailto:florence.maillet@inria.fr)

## A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

**Attention:** Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

## Consignes pour postuler

Nous vous remercions d'adresser un CV accompagné d'une lettre de motivation.

### Sécurité défense :

Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

### Politique de recrutement :

Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.