Offer #2024-07410

PhD Position F/M Spoken language detection and clustering

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

Context

Inria Défense&Sécurité (Inria D&S) was created in 2020 to federate Inria's actions for the benefit of military forces. The PhD will be carried out within the audio processing research team of Inria D&S, under the supervision of Jean-François Bonastre and co-supervised by Raphaël Duroselle.

The thesis is part of a project aimed at explainable and frugal voice profiling. Voice profiling consists in extracting information from an audio recording, such as identity, spoken language, age, geographical and ethnic origin, or socio/patho/physiological marks in the voice. The aim of this project is to make voice profiling systems explainable without degrading performance. Explainability maintains the central role of operators in the process, giving them the means to make informed decisions.

Assignment

Approach

The considered approach is based on the definition of a set of generic vocal attributes, which are shared by a group of individuals. Only the presence or absence of an attribute within a given speech utterance is used to make a decision. Thus the system uses binary representations. This approach was introduced for the speaker verification task [1,2].

The proposed thesis aims at developing this methodology working on the spoken language recognition task [3]. The system aims to group segments belonging to the same language and determine whether they belong to a recognized set of languages or if they are an unknown langage. In the latter case, the system should rely on the attributes it knows to analyse the proximity of the new unknown language to known languages.

Since the emergence of iVector models [4] (initially for speaker recognition) in spoken language recognition, the general scheme remains the same. The system relies on an embedding extractor, trained on a large dataset and responsible of representing an acoustic sequence of any duration by a fixed size vector. Then, 1:1 classifiers, comparing two languages, or 1:N classifiers, comparing N languages, are built, and a decision-making system relies on these classifiers to perform the various tasks. Neural networks, like bottleneck features have delivered very significant gains [5]. Subsequently, embeddings derived from neural models, known as "xVectors", replaced iVectors and made it possible both to increase model size (and performance) and to simplify the training recipe of the models [6]. More recently, pre-trained models such as WavLM [7] or MMS [8] have been used [9]. These generic models enable interesting gains, especially when the training dataset is small for some languages, but at the cost of a significant increase of the number of parameters.

These approaches have common limitations: they cannot explain their decision, they suffer a drop in performance over domain change, they have difficulty managing the imbalance between datasets of different languages and they are cumbersome to adapt or retrain. Finally, they offer little or nothing in the case of unknown languages.

In this project, we propose to start from a state-of-the-art model and to adapt the vocal attributes approach to the spoken language recognition task. A language could be represented by a binary vector corresponding to the presence/absence of attributes in that language, or by a scalar vector, indicating the frequency of attributes in the language. The attributes themselves can incorporate higher-level information, such as phonotactic and linguistic levels. With this architecture, an unknown language (in the sense that no data corresponding to this language is present in the training data) can be recognized and compared to known languages, for instance exploiting geo linguistic knowledge. A model of this new language can thus be built as soon as the first recording of that language is available, and then adapted without computational cost each time an additional recording is added. If necessary, the attribute extractor can be adapted by adding one or more attributes from the new data, without need to relearn the whole model. Therefore we expect significant gains, in terms of explainability, handling of unknown languages and context adaptation.

Goals

1. Apply the vocal attribute approach to spoken language recognition.

2. Study the capacity to learn or extend (new language or new attributes) a model from unlabelled or weakly labelled data (for instance only labelled with a region), optimizing the ration between quantity and quality of weak labels.
3. Explore this approach to analyse unknown languages.
4. Exploit this approach for the language clustering of audio documents, even when some languages are unknown.

Main activities

- Bibliography, development and evaluation of spoken language recognition systems.
- Deep learning, adaptation of self-supervised speech processing models such as WavLM [7] or MMS [8] ;
- Semi-supervised learning ;
- Explainability of deep learning models.

Skills

- Master level in computer science, mathematics or phonetics.
- Strong interest in applied research.
- Written and spoken English
- Signal processing
- Machine learning and deep learning
- Experience with deep learning toolkits such as pytorch or keras
- Speech processing experience, knowledge of open source toolkits such as kaldi or speechbrain

References

[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 (IWBF) (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▯59. 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▯98. 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▯67.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▯11. ISCA, 2018. https://doi.org/10.21437/Odyssey.2018-15.

[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▯18. 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.

Benefits package

- Subsidized meals,
- Partial reimbursement of public transport costs,
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.),
- Possibility of teleworking and flexible organization of working hours,
- Professional equipment available (videoconferencing, loan of computer equipment, etc.),
- Social, cultural and sports events and activities,

Remuneration

- 1st and 2nd year : 2082 € bruts - gross /month
- 3rd year : 2190 € bruts - gross /month

General Information

- **Town/city :** PARIS
- **Inria Center :** Siège
- **Starting date :** 2024-05-01
- **Duration of contract :** 3 years

Contacts

- **Inria Team :** MIS-DEFENSE (DIRECTION)
- **PhD Supervisor :**
  Maillet Florence / florence.maillet@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Defence Security :
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.