2018-01201 - PhD Position F/M Privacy preserving and personalized transformations for speech recognition

Contract type : Public service fixed-term contract
Level of qualifications required : Graduate degree or equivalent
Fonction : PhD Position

Context

This PhD thesis fits within the scope of a collaborative project (project DEEP-PRIVACY funded by the French National Research Agency) involving the MULTISPEECH team of Inria Nancy - Grand-Est, the MAGNET team of Inria Lille - Nord Europe, the LIUM (Laboratoire d'Informatique de l'Université du Mans) and the LIA (Laboratoire Informatique d'Avignon).

This PhD position is in collaboration with the Le Mans University, and will be co-supervised by Denis Jouvet (https://members.loria.fr/DJouvet/) and Anthony Larcher (https://lium.univ-lemans.fr/team/anthony-larcher/). The selected candidate is expected to spend time in both teams over the course of the PhD.

Additional information:

- Ecole Doctorale IAEM Lorraine (http://iaem.univ-lorraine.fr/)
- Duration: 3 years
- Starting date: spring 2019

Assignment

Scientific Context

Over the last decade, great progress has been made in automatic speech recognition [Saon et al., 2017; Xiong et al., 2017]. This is due to the maturity of machine learning techniques (e.g., advanced forms of deep learning), to the availability of very large datasets, and to the increase in computational power. Consequently, the use of speech recognition is now spreading in many applications, such as virtual assistants (as for instance Apple's Siri, Google Now, Microsoft's Cortana, or Amazon's Alexa) which collect, process and store personal speech data in centralized servers, raising serious concerns regarding the privacy of the data of their users. Embedded speech recognition frameworks have recently been introduced to address privacy issues during the recognition phase: in this case, a (pre-trained) speech recognition model is shipped to the user's device so that the processing can be done locally without the user sharing its data. However, speech recognition technology still has limited performance in adverse conditions (e.g., noisy environments, reverberated speech, strong accents, etc.) and thus, there is a need for performance improvement. This can only be achieved by using large speech corpora that are representative of the actual users and of the various usage conditions. There is therefore a strong need to share speech data for improved training that is beneficial to all users, while preserving the privacy of the users, which means at least keeping the speaker identity and voice characteristics private[1].

[Saon et al., 2017] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall. English conversational telephone speech recognition by humans and machines. *Technical report*, arXiv:1703.02136, 2017.

[Xiong et al., 2017] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. Achieving human parity in conversational speech recognition. *Technical report*, arXiv:1610.05256, 2017.

[1] Note that when sharing data, users may want not to share data conveying private information at the linguistic level (e.g., phone number, person name, ...). Such privacy aspects also need to be taken into account, but they are out-of-the scope of this thesis.

Main activities

Missions:

Within this context, the objective of the proposed thesis is twofold. First, it aims at finding a privacy preserving transform of the speech data, and, second, it will investigate the use of additional personalized transforms, that can be applied on the user's terminal, to increase speech recognition performance.

In the proposed approach, the device of each user will not share its raw speech data, but a privacy preserving transformation of the user speech data. In such approach, some private computations will be handled locally, while some cross-user computations may be carried out on a server using the transformed speech data, which protect the speaker identity and some of his/her features (gender, sentiment, emotions...). More specifically, this will rely on a representation learning to separate the features of the user data that can expose private information from generic ones useful for the task of interest, i.e., here, the recognition of the linguistic content. We will build upon ideas of Generative Adversarial Networks (GANs) for proposing such a privacy preserving transform. Since a few years, GANs are getting more and more used in deep learning. They typically rely on both a generative network and a discriminative network, where the generator aims to output samples that the discriminator cannot distinguish from the true samples [Goodfellow et al., 2014; Creswell et al., 2018]. They have also been used as autoencoders [Makhzani et al., 2015] which are made of three mains blocks: encoder, generator and discriminator. In our case, the discriminators shall focus on discriminating between speakers and/or between voice-related classes (defined according to gender, emotions, etc). The training objective will be to maximize the speech recognition performance (using the privacy preserving transformed signal) while minimizing the available speaker or voice-related information measured by the discriminator.

As devices are getting more and more personal, it creates opportunities to make speech recognition

General Information

- **Town/city :** Villers-lès-Nancy
- **Inria Center :** CRI Nancy - Grand Est
- **Starting date :** 2019-07-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2019-06-30

Contacts

- **Inria Team :** MULTISPEECH
- **PhD Supervisor :**
  Jouvet Denis / denis.jouvet@inria.fr

About Inria

Inria, the French national research institute for the digital sciences, promotes scientific excellence and technology transfer to maximise its impact. It employs 2,400 people. Its 200 agile project teams, generally with academic partners, involve more than 3,000 scientists in meeting the challenges of computer science and mathematics, often at the interface of other disciplines. Inria works with many companies and has assisted in the creation of over 160 startups. It strives to meet the challenges of the digital transformation of science, society and the economy.

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.

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.

more personalized. This includes two aspects: adapting the model parameters to the speaker (and to the device) and introducing personalized transforms to help hiding the speaker voice identity. Both aspects will be investigated. Voice conversion approaches provide example of transforms aiming at modifying the voice of a speaker so that it sounds like the voice of another target speaker [e.g., Chen et al., 2014; Mohammadi & Kain, 2014]. Similar approaches can thus be applied to map speaker specific features to those of a standard (or average) speaker, which thus would help concealing its identity. To take benefit of the increased personal usage of terminals, speaker and environment specific adaptation will be investigated to improve speech recognition performance. Collaborative learning mixing speech and speaker recognition has been shown to benefit both tasks [Liu et al. 2018; Garimella et al. 2015] and provide a way to combine both information in a single framework. This approach will be compared to adaptation of deep neural networks-based models [e.g., Abdel-Hamid & Jiang, 2013] to handle best different amounts of adaptation data.

Bibliography:

[Abdel-Hamid & Jiang, 2013] Abdel-Hamid, O., & Jiang, H. Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code. In *ICASSP-2013, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7942-7946, 2013.

[Chen et al., 2014] Chen, L. H., Ling, Z. H., Liu, L. J., & Dai, L. R. Voice conversion using deep neural networks with layer-wise generative training. TASLP-2014, *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 22(12), pp. 1859-1872, 2014.

[Creswell et al., 2018] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine* 35, 1, 53-65, 2018.

[Garimella et al. 2015] Garimella, S., Mandal, A., Strom, N., Hoffmeister, B., Matsoukas, S., & Parthasarathi, S. H. K., Robust i-vector based adaptation of DNN acoustic model for speech recognition. In *INTERSPEECH*, 2015.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672-2680, 2014.

[Liu et al. 2018] Y. Liu, L. He, J. Liu, and M. Johnson, \Speaker Embedding Extraction with Phonetic Information," in *INTERSPEECH* , pp. 2247-2251, 2018

[Makhzani, 2015] Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. Adversarial autoencoders. *arXiv preprint* arXiv:1511.05644, 2015.

[Mohammadi & Kain, 2014] Mohammadi, S. H., & Kain, A. Voice conversion using deep neural networks with speaker-independent pre-training. In *SLT-2014, Spoken Language Technology Workshop* , pp. 19-23, 2014.

Skills

Technical skills and level required :

- Master in machine learning or in computer science
- Background in statistics, and in deep learning
- Experience with deep learning tools is a plus
- Good computer skills (preferably in Python)
- Experience in speech and/or speaker recognition is a plus

Benefits package

- Subsidised catering service
- Partially-reimbursed public transport
- Social security
- Paid leave
- Flexible working hours
- Sports facilities

Remuneration

Salary: 1982€ gross/month for 1st and 2nd year. 2085€ gross/month for 3rd year.

Monthly salary after taxes : around 1596,05€ for 1st and 2nd year. 1678,99€ for 3rd year. (medical insurance included).