M2 internship offers
LORIA - MULTISPEECH

https ://team.inria.fr/multispeech/fr/

To apply, please send your CV and a short motivation letter directly to the supervisors of the corresponding offer.

Offer 1
Contrastive Learning for Hate Speech Detection

General information

Supervisors	Nicolas Zampieri, Irina Illina, Dominique Fohr
Address	LORIA, Campus Scientifique - BP 239, 54506 Vandœuvre-lès-Nancy
Phone	03 54 95 84 06
Email	`fistname.lastname@loria.fr`
Office	C 145

Motivation and context

The United Nations defines hate speech as "any type of communication through speech, writing or behavior, which denigrates a person or group based on who they are, i.e. their religion, ethnicity, nationality, or other identity factor.". We are interested in hate speech posted on social networks. With the expansion of social networks (Twitter, Facebook, etc.), the number of messages posted every day has increased dramatically. It is very difficult and expensive to process the millions of content posted every day in order to remove hateful content. Thus, automatic methods are required to moderate the influx. Automatic hate speech detection is a difficult task in the field of natural language processing (NLP) [6].

With the appearance of transformer-based language models like BERT [3], new state-of-the-art models have emerged for hate speech detection like HateBERT [1]. Current NLP models rely strongly on efficient learning algorithms. We are particularly interested in one of them : contrastive learning. Contrastive learning is employed to learn an embedding space such that pairs of similar sentences have close representations. [5] provide a summary of different models based on contrastive learning in language processing.

Goals and Objectives

The goal of this internship is to study contrastive learning in the context of hate speech detection. We believe that using this methodology will make the models more effective. Our model learns to estimate whether two sentences have the same sentiment or not. Based on the first model, the intern will explore other approaches of contrastive learning, such as SimCSE [4] or Dual Contrastive Learning [2] models. The studied methods will be validated on several datasets to assess the robustness of the approach. In our team, we have several labeled corpora from social networks.

The internship workplan is as follows : at the beginning the student will conduct a state-of-the-art study on recent developments in hate speech detection and contrastive learning in NLP. The student will implement the selected methods. Finally, the performance of the different implemented methods will be evaluated on several hate speech corpora and compared to the state-of-the-art.

Required Skills

The candidate should have an experience with Deep Learning, including a good practice in Python and an understanding of deep learning libraries like Keras, Pytorch or Tensorflow.

Additional information

The student intern will join the MULTISPEECH team. The team provides access to the computational resources (GPU, CPU and datasets) in order to carry out the research.

References

[1] Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. HateBERT : Retraining BERT for Abusive Language Detection in English.. *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (pp. 17-25). ACL. doi :10.18653/v1/2021.woah-1.3. August 2021.

[2] Chen, Q., Zhang, R., Zheng, Y., & Mao, Y. Dual Contrastive Learning : Text Classification via Label-Aware Data Augmentation.. https ://arxiv.org/abs/2201.08702. 2022.

[3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding.. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, (pp. 4171-4186). 2022.

[4] Gao, T., Yao, X., & Chen, D. SimCSE : Simple Contrastive Learning of Sentence Embeddings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. ACL. doi :10.18653/v1/2021.emnlp-main.552. 2021.

[5] Rethmeier, N., & Augenstein, I. A Primer on Contrastive Pretraining in Language Processing : Methods, Lessons Learned and Perspectives.. https ://arxiv.org/abs/2102.12982. 2021.

[6] Zampieri, N., Ramishc, C., Illina, I., & Fohr D. Identification of Multiword Expressions in Tweets for Hate Speech Detection.. *In Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 202-210, 2022. European Language Resources Association.

Offer 2
Diffusion-based Deep Generative Models for Audio-visual Speech Modeling

General information

Supervisors	Mostafa SADEGHI, Romain SERIZEL
Address	LORIA, Campus Scientifique - BP 239, 54506 Vandœuvre-lès-Nancy
Email	mostafa.sadeghi@inria.fr, romain.serizel@loria.fr

Motivation

Recently, diffusion models have gained much attention due to their powerful generative modeling performance, in terms of both the diversity and quality of the generated samples [1]. It consists of two phases, where during the so-called forward diffusion process, input data are mapped into Gaussian noise by gradually perturbing the data. Then, during a reverse process, a denoising neural network is learned that removes the added noise at each step, starting from pure Gaussian noise, to eventually recover the original clean data. Diffusion models have found numerous successful applications, particularly in computer vision, e.g., text-conditioned image synthesis, outperforming previous generative models, including variational autoencoders (VAEs), generative adversarial networks (GANs), and normalizing flows (NFs). Diffusion models have also been successfully applied to audio and speech signals, e.g., for audio synthesis [2] and speech enhancement [3].

Goal and objectives

Despite their rapid progress and application extension, diffusion models have not yet been applied to audio-visual speech modeling. This task involves joint modeling of audio and visual modalities, where the latter concerns the lip movements of the speaker, as there is a correlation between what is being said and the lip movements. This joint modeling effectively incorporates the complementary information of visual modality for speech generation. Such a framework has already been established based on VAEs [4]. Given the great potential and advantages of diffusion models, in this project, we would like to develop a diffusion-based audio-visual generative modeling framework, where the generation of audio modality, i.e., speech, is conditioned on the visual modality, i.e., lip images, similarly to text-conditioned image synthesis. This might then serve as an efficient representation learning framework for downstream tasks, e.g., audio-visual speech enhancement (AVSE) [4].

Background in statistical signal processing, computer vision, machine learning, and deep learning frameworks (Python, PyTorch) are favored. Interested candidates should send an email to the supervisors with a detailed CV and transcripts.

Work environment

This master internship is part of the REAVISE project : "Robust and Efficient Deep Learning based Audio-visual Speech Enhancement" (2023-2026) funded by the French National Research Agency (ANR). The general objective of REAVISE is to develop a unified AVSE framework that leverages recent methodological breakthroughs in statistical signal processing, machine learning, and deep neural networks in order to design a robust and efficient AVSE framework.

The intern will be supervised by Mostafa Sadeghi (researcher, Inria) and Romain Serizel (associate professor, University of Lorraine), as members of the MULTISPEECH team, and will benefit from the research environment, expertise, and computational resources (GPU & CPU) of the team.

References

[1] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M. H. Yang, Diffusion models : A comprehensive survey of methods and applications *arXiv preprint arXiv :2209.00796*, 2022.

[2] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, Diffwave : A versatile diffusion model for audio synthesis *arXiv preprint arXiv :2009.09761*, 2020.

[3] Y. J. Lu, Z. Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, Conditional diffusion probabilistic model for speech enhancement *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[4] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, Audio-visual speech enhancement using conditional variational auto-encoders *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1788?1800, 2020.

Offer 3
Efficient Attention-based Audio-visual Fusion Mechanisms for Speech Enhancement

General information

Supervisors	Mostafa SADEGHI	Romain SERIZEL
Address	LORIA, Campus Scientifique - BP 239, 54506 Vandœuvre-lès-Nancy	
Email	mostafa.sadeghi@inria.fr	romain.serizel@loria.fr

Motivation

Audiovisual speech enhancement (AVSE) is defined as the task of improving the quality and intelligibility of a noisy speech signal by utilizing the complementary information provided by the visual modality, i.e., lip movements of the speaker [1]. Visual modality is especially important in high-noise situations, as it is less affected by acoustic noise. Because of that, AVSE could be exploited in several practical applications, including hearing assistive devices. Numerous works have already studied the integration of visual modality with audio modality to improve the performance of speech enhancement. While the majority of audiovisual speech enhancement algorithms rely on deep neural networks and supervised learning, they require very large audiovisual datasets with diverse noise instances to have good generalization performance.

A recently introduced AVSE approach is based on unsupervised learning [2,3], where during a training phase, the statistical distribution of clean speech is learned from a clean audiovisual dataset. This is done using a deep generative model, e.g. variational autoencoder (VAE) [4]. Then, at test (inference) time, the learned distribution is combined with a noise model to estimate the clean speech signal from the available noisy speech observations.

Goal and objectives

An important element of AVSE is audio-visual feature fusion, which should robustly and efficiently combine the two modalities. Current fusion mechanisms used for unsupervised AVSE are based on simple feature concatenation, which is not effective, as it treats the two feature streams on an equal basis. In fact, the audio modality usually contributes more than the visual modality, but in general, their contributions should be robustly balanced and weighted. In this project, we are going to develop efficient feature fusion modules based on attention models [5], which have proven very successful in different applications. The designed fusion module is supposed to robustly and efficiently incorporate the potentially different uncertainty (reliability) levels of the two modalities. We will then evaluate its effectiveness for AVSE.

Background in statistical signal processing, probabilistic machine learning, optimization, and programming languages & deep learning frameworks (Python, PyTorch) are favored. Interested candidates should send an email to the supervisors with a detailed CV and transcripts.

Work environment

This master internship is part of the REAVISE project : "Robust and Efficient Deep Learning based Audio-visual Speech Enhancement" (2023-2026) funded by the French National Research Agency (ANR). The general objective of REAVISE is to develop a unified AVSE framework that leverages recent methodological breakthroughs in statistical signal processing, machine learning, and deep neural networks in order to design a robust and efficient AVSE framework.

The intern will be supervised by Mostafa Sadeghi (researcher, Inria) and Romain Serizel (associate professor, University of Lorraine), as members of the MULTISPEECH team, and will benefit from the research environment, expertise, and computational resources (GPU & CPU) of the team.

References

[1] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, An overview of deep-learning-based audio-visual speech enhancement and separation *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368?1396, 2021.

[2] M. Sadeghi and X. Alameda-Pineda, Switching variational auto-encoders for noise-agnostic audio-visual speech enhancement *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[3] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, Audio-visual speech enhancement using conditional variational auto-encoders *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1788?1800, 2020.

[4] D. P. Kingma and M. Welling, An introduction to variational autoencoders *Foundations and Trends in Machine Learning*, 2019.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need *Advances in neural information processing systems*, 2017.

Offer 4
Multi-modal Stuttering Detection Using Self-supervised Learning

General information

Supervisors	Shakeel Ahmad Sheikh, Slim Ouni
Address	LORIA, Campus Scientifique - BP 239, 54506 Vandœuvre-lès-Nancy
Email	`fistname.lastname@loria.fr`
Office	C 137

Motivation

Stuttering is a neuro-developmental speech disorder that starts appearing when language, speech, and emotion supporting neural connections are changing quickly [2]. In standard stuttering therapy sessions, the speech pathologists or speech therapists either manually examine and analyze the person who stutter (PWS) speech or their recordings. In order to rectify the stuttering, the speech therapists carefully observe and monitor the patterns in speech utterances of PWS. However, this convention of stuttering detection is very time consuming and strenuous. It is also biased towards the subjective belief of speech language therapists. Thus, it is important to build stuttering detection interactive tools that provide impartial objective assessment, and can be utilized to tune and improve various ASR virtual assistants for stuttered speech.

Deep learning has been used tremendously in domains like speech recognition [5], emotion detection [1], however, in stuttering domain, its application is limited. The acoustic cues embedded in the speech of PWS can be exploited by various deep learning methods in the detection of stuttering. Most of the existing stuttering detection techniques utilize spectral features such as spectrograms and MFCCs as an input representation of the stuttered speech [12, 11, 3]. The most common problem in the stuttering domain is the dataset issue. There are few stuttering datasets like UCLASS, FluencyBank, and SEP28K [3], which are small containing only a few dozens of speakers. While deep learning methods have shown substantial gains in domains like ASR, speaker verification, emotion detection, etc, however, the improvement in stuttering detection is very limited, most likely due to the miniature size of datasets.

The common strategy in dealing with training on small datasets is to apply transfer learning, where the pre-trained model (trained first on some auxiliary task on a large dataset) is used to enhance the performance of the desired task, for which data is very scarce. The deep learning model trained on some auxiliary task can be fine-tuned by re-training, or replacing some of its last layers, or it can also be employed as a feature extractor for the desired task, that we are trying to address. Transfer learning methodology has been explored in various fields like ASR, emotion detection [8], etc. Recently, self-supervised learning has shown significant improvement in stuttering detection [11, 18, 17, 16].

Multimodal Stuttering Detection

Stuttering can be characterized as an audio-visual problem. Cues are present both in the visual (e.g., head nodding, lip tremors, quick eye blinks, and unusual lip shapes) as well as in the audio modality [4]. This multimodal learning paradigm could be helpful in learning robust stutter-specific hidden representations across the cross-modality platform, and could also help in building robust automatic stuttering detection systems. Self-supervised learning can also be exploited to capture acoustic stutter-specific representations based on guided video frames. As proposed by Shukla *et al.* [14], this framework could be helpful in learning stutter-specific features from audio signals guided by visual frames or vice versa. Altinkaya and Smeulders [15] recently presented the first audio-visual stuttered dataset which consists of 25 speakers (14 male, 11 female). They trained ResNet-based RNN (gated recurrent unit) on the audio-visual modality for the detection of block stuttering type. The main idea in this internship is to explore the impact of further self supervised learning in stuttering detection in combination with audio-visual setup. The goal of the proposed study is to develop and evaluate audio-visual based self supervised stuttering detection classifiers, that will be able to distinguish among several stutter classes.

1. *Objective 1 :* Lliterature survey by looking at the existing work in stuttering detection.
2. *Objective 2 :* Developing a pre-trained stuttering classifier based on self-supervised learning ; Some initial experiments would be carried out. We would explore the self supervised models such as wav2vec 2.0, a modified version of wav2vec [9], and their variants such as Unispeech, HuBERT, etc. We would use wav2vec 2.0 either as a feature extractor or just fine tune it by replacing the last few layers and adapt it for stuttering detection.

3. The experiments would be carried out on the newly developed French stuttering dataset.
4. *Objective 3 :* Carrying out the actual experiments and the impact of fine-tuning and pre-trained features would be analyzed on the raw stuttered embedded audio-visual stuttered samples.

References

[1] Mehmet Berkehan Ak Cay and Kaya Oguz L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M. H. Yang, Speech emotion recognition : Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers" *Speech Communication*, 116 (2020) pp.56-76.

[2] Smith, Anne and Weber, Christine How stuttering develops : The multifactorial dynamic pathways theory" *Journal of Speech, Language, and Hearing Research*, 60 (2017) pp.2483–2505.

[3] Shakeel A. Sheikh, Md Sahidullah, Fabrice Hirsch, Slim Ouni, Machine learning for stuttering identification : Review, challenges and future directions, *Neurocomputing*, 514 (2022), pp 385-402,

[4] Guitar, Barry. Stuttering : An integrated approach to its nature and treatment. *Lippincott Williams & Wilkins*, 2013.

[5] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, "Speech Recognition Using Deep neural networks : A systematic review," *IEEE Access*, vol. 7, pp. 19143-19165, 2019.

[6] Latif, Siddique, Rajib Rana, Sara Khalifa, Raja Jurdak, Junaid Qadir, and Björn W. Schuller. "Deep representation learning in speech processing : Challenges, recent advances, and future trends." *arXiv preprint arXiv :2001.00378* (2020).

[7] Ning, Y., He, S., Wu, Z., Xing, C. and Zhang, L.J., 2019. A review of deep learning based speech synthesis. *Applied Sciences*, 9(19), p.4050.

[8] Wang, Yingzhi, Abdelmoumene Boumadane, and Abdelwahab Heba. "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding." *arXiv preprint arXiv :2111.02735* (2021).

[9] Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. "wav2vec 2.0 : A framework for self-supervised learning of speech representations." *Advances in Neural Information Processing Systems*, 33 (2020) : 12449-12460.

[10] Lea, Colin, Vikramjit Mitra, Aparna Joshi, Sachin Kajarekar, and Jeffrey P. Bigham. "Sep-28k : A dataset for stuttering event detection from podcasts with people who stutter." *In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6798-6802. IEEE, 2021.

[11] Sheikh, Shakeel A., Md Sahidullah, Slim Ouni, and Fabrice Hirsch. "End-to-End and Self-supervised learning for ComParE 2022 stuttering sub-challenge." *In Proceedings of the 30th ACM International Conference on Multimedia*, pp. 7104-7108. 2022.

[12] Sheikh, Shakeel A., Md Sahidullah, Fabrice Hirsch, and Slim Ouni. "Robust stuttering detection via multi-task and adversarial learning." *In 2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 190-194. IEEE, 2022.

[13] Ngiam, Jiquan, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. "Multimodal deep learning." *In ICML*. 2011.

[14] Shukla, Abhinav, Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. "Visually guided self supervised learning of speech representations." *In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6299-6303. IEEE, 2020.

[15] Altinkaya, Mehmet, and Arnold WM Smeulders. "A dynamic, self supervised, large scale audiovisual dataset for stuttered speech." *In Proceedings of the 1st International Workshop on Multimodal Conversational AI*, pp. 9-13. 2020.

[16] Mohapatra, Payal, Akash Pandey, Bashima Islam, and Qi Zhu. "Speech disfluency detection with contextual representation and data distillation." *In Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*, pp. 19-24. 2022.

[17] Grósz, Tamás, Dejan Porjazovski, Yaroslav Getman, Sudarsana Kadiri, and Mikko Kurimo. "Wav2vec2-based Paralinguistic Systems to Recognise Vocalised Emotions and Stuttering." *In Proceedings of the 30th ACM International Conference on Multimedia*, pp. 7026-7029. 2022.

[18] Bayerl, Sebastian P., Dominik Wagner, Elmar Nöth, and Korbinian Riedhammer. "Detecting Dysfluencies in Stuttering Therapy Using wav2vec 2.0." *arXiv preprint arXiv :2204.03417* (2022).

Offer 5
Dictionary learning for deep unsupervised speech separation

General information

Supervisors	Paul Magron	Mostafa Sadeghi
Address	LORIA, Campus Scientifique - BP 239, 54506 Vandœuvre-lès-Nancy	
Email	paul.magron@inria.fr	mostafa.sadeghi@inria.fr
Office	C 141	C 136

Motivation and context

Speech separation consists in isolating the signals that correspond to each speaker from an acoustic mixture where several persons might be speaking. This task is an important preprocessing step in many applications such as hearing aids or vocal assistants based on automatic speech recognition.

State-of-the-art separation systems rely on supervised deep learning, where a network is trained to predict the isolated speakers' signals from their mixture [1, 2]. However, these approaches are costly in terms of training data and have a limited capacity to generalize to unseen speakers.

Goal and objectives

The goal of this internship is to design a fully unsupervised system for speech separation, which is more data-efficient than supervised approaches, and applicable to any mixture of speakers. To that end, we propose to combine variational autoencoders (VAEs) with dictionary models (DMs). DM consist in decomposing a given input matrix (usually : an audio spectrogram) as the product of two interpretable factors : a *dictionary* of spectra and a temporal activation matrix). This family of methods has been extensively researched before the era of deep learning [3], but it is limited since real-world audio spectrograms cannot be decomposed using such simple models.

Therefore, we propose to leverage VAEs as a tool to learn a latent representation of the data which is regularized using DMs. Such a system can be cast as an instance of *transform learning* [4] : the key idea is to apply a (learned) transform to the data so that it better complies with a desirable property - here, decomposition on a dictionary. A first attempt was recently proposed and has shown promising results in terms of speech modeling [5], although it was using a fixed dictionary. This internship aims at extending this work by considering a system where both the VAE and the dictionary are learned jointly, and applying it to the task of speech separation.

Once trained, the resulting system operates in three stages : (i) the (mixture) audio spectrogram is projected through the encoder into some latent space ; (ii) this latent representation is decomposed efficiently using a DM learning algorithm, which provides a latent feature for each speaker ; (iii) these latent features are passed through the decoder to retrieve a spectrogram for each speaker. Such a system is promising since it is fully unsupervised (it can be trained without knowledge of specific mixtures), it yields an interpretable decomposition of the latent representation, and it can serve as a basis for other applications (including speaker diarization, speech enhancement or voice conversion).

A good practice in Python and basic knowledge about deep learning, both theoretical and practical (e.g., using PyTorch) are required. Some notions of audio/speech signal processing and machine learning is a plus.

Work environment

The trainee will be supervised by Paul Magron (Chargé de Recherche Inria) and Mostafa Sadeghi (Researcher, Inria Starting Faculty Position), and will benefit from the research environment and the expertise in audio signal processing of the MULTISPEECH team. This team includes many PhD students, post-docs, trainees, and permanent staff working in this field, and offers all the necessary computational resources (GPU and CPU, speech datasets) to conduct the proposed research.

References

[1] D. Wang and J. Chen, Supervised Speech Separation Based on Deep Learning : An Overview *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702-1726, 2018.

[2] Y. Luo and N. Mesgarani, Conv-TasNet : Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256-1266, 2019.

[3] T. Virtanen, Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066-1074, 2007.

[4] D. Fagot, H. Wendt and C. Févotte, Nonnegative Matrix Factorization with Transform Learning *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[5] M. Sadeghi, and P. Magron, A Sparsity-promoting Dictionary Model for Variational Autoencoders *Interspeech*, 2022.

Offer 6
Semantic latent space for expressive text-to-speech

General information

Supervisors	Vincent Colotte, Slim Ouni
Address	LORIA, Campus Scientifique - BP 239, 54506 Vandœuvre-lès-Nancy
Phone	03 54 95 20 74
Email	vincent.colotte@loria.fr, slim.ouni@inria.fr
Office	C141

Motivation

Over the last decades, text-to-speech synthesis (TTS) has reached good quality and intelligibility, and is now commonly used in information delivery services, as for instance in call center automation, and in navigation systems. In the past, the main goal when developing TTS systems was to achieve high intelligibility. The speech style was then typically a "reading style," which resulted from the style of the speech data used to develop TTS systems (reading of a large set of sentences). Recent research on speech synthesis focuses now on expressive speech to obtain generated speech more expressive or spontaneous. Almost all systems are now based on neural network methods. Therefore, to tackle expressiveness integration in a network, as numerous recent works in neural networks, training and testing step pass through several steps with specific latent spaces to condition the network or to propose a latent representation to control the expressiveness. In stochastic processes, the explanation of such a numeric representation is still difficult to extract [1]. Moreover, the use of new representations as Word2Vec for textual material or Wav2Vec for audio signal shows that we can find a representation with implicit linguistic and semantic information. The need of explanation still remains. The internship will take place in this framework.

Objectives and expected outcomes

The goal of the proposed study is to investigate the information contained in a latent representation dedicated to expressive speech. Previous work dealt with Variational Autoencoder (VAE) approach to explore this dimension in the audiovisual domain [2] without emotion tag the latent representation retrieved the emotional information. In addition, [3] used several representations of acoustic expressiveness to condition a network to transfer an emotion from a speaker to another sentence of another speaker. Moreover, [5] had jointly used acoustic and textual expressiveness representation. The textual representation was based on SBERT approach.

The internship work will consist to analyze latent representations of a TTS system (for instance Glow approach for audio speech or VAE for audio-visual speech). The second step will introduce semantic information by textual latent representation from simple tag [4], a description or the text itself. The objective is to jointly learn representations and analyze them to extract understanding for controlling the system.

Additional information and requirements

The internship will be carried out within the framework of the European project Humane-AI-Net. A good knowledge of Python and basic knowledge of neural network learning is required.

References

[1] Tits, N., Wang, F., Haddad, K.E., Pagel, V., Dutoit, T. *Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis,* in Proc. Interspeech, 2019

[2] S. Dahmani, V. Colotte, V. Girard and S. Ouni *Learning emotions latent representation with CVAE for Text-Driven Expressive AudioVisual Speech Synthesis,* in Neural Networks, Elsevier, 2021

[3] A. Kulkarni, V. Colotte, D. Jouvet, *Analysis of expressivity transfer in non-autoregressive end-to-end multispeaker TTS systems,* in Proc. Interspeech, 2022

[4] Kim, M., Cheon, S.J., Choi, B.J., Kim, J.J., Kim, N.S. *Expressive Text-to-Speech Using Style Tag.* Interspeech 2021

[5] Shin, Y., Lee, Y., Jo, S., Hwang, Y., Kim, T., *Text-driven Emotional Style Control and Cross-speaker Style Transfer in Neural TTS.* Interspeech 2022.

Offer 7
Disentanglement in Speech Data for Privacy Needs

General information

Supervisors	Emmanuel Vincent, Marc Tommasi
Address	LORIA, Campus Scientifique - BP 239, 54506 Vandœuvre-lès-Nancy
Email	emmanuel.vincent@inria.fr, marc.tommasi@inria.fr

Motivation and context

Large-scale collection, storage, and processing of speech data poses severe privacy threats [1]. Indeed, speech encapsulates a wealth of personal data (e.g., age and gender, ethnic origin, personality traits, health and socio-economic status, etc.) which can be linked to the speaker's identity via metadata or via automatic speaker recognition. Speech data may also be used for voice spoofing using voice cloning software. With firm backing by privacy legislations such as the European general data protection regulation (GDPR), several initiatives are emerging to develop and evaluate privacy preservation solutions for speech technology. These include voice anonymization methods [2] which aim to conceal the speaker's voice identity without degrading the utility for downstream tasks, and speaker re-identification attacks [3] which aim to assess the resulting privacy guarantees, e.g., in the scope of the VoicePrivacy challenge series [4].

Goals and objectives

The internship will tackle the objective of speech anonymization. Previous works have shown that simple adversarial approaches that aim at removing speaker identity from speech signals do not provide sufficient privacy guaranties [5]. An interpretation of this failure can be that adversaries were not strong enough. Moreover, there is no clear evidence that a transformation that removes speaker identity is informative enough to allow the reconstruction of intelligible speech signals. These observations raise a classical trade-off between privacy and utility that is essential in many privacy preservation scenarios. Instead of trying to remove speaker information, another option is to replace it by another one. To do so, a sub-objective is to disentangle speech signals, that is to isolate speech features that contribute to the success of speaker identification. Disentanglement is understood in this project as the process of embedding voice data in a new representation where different types of information (speaker identity, linguistic content, or even traits like age, gender or ethnicity) are separated and associated with disjoint sets of features. Variational autoencoders are supposed to naturally support disentanglement [6]. Additionally, variational approaches can also be used to make attackers stronger by introducing more diversity. Those two ways of improving adversarial approaches for learning a private representation of speech will be investigated.

References

[1] A. Nautsch, A. Jimenez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, M. Gomez-Barrero, D. Petrovska, G. Chollet, N. Evans, T. Schneider, J.-F. Bonastre, B. Raj, I. Trancoso, and C. Busch. *Preserving privacy in speaker and speech characterisation,* in Computer Speech and Language, 2019

[2] B. M. L. Srivastava, M. Maouche, M. Sahidullah, E. Vincent, A. Bellet, M. Tommasi, N. Tomashenko, X. Wang, and J. Yamagishi. *Privacy and utility of x-vector based speaker anonymization,* in IEEE/ACM Transactions on Audio, Speech and Language Processing, 30 :2383–2395, 2022.

[3] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent. *Evaluating voice conversion-based privacy protection against informed attackers,* in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020.

[4] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien, A. Chanclu, J.-F. Bonastre, M. Todisco, and M. Maouche. *The VoicePrivacy 2020 Challenge : Results and findings,* Computer Speech and Language, 2022.

[5] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent. *Privacy-preserving adversarial representation learning in ASR : Reality or illusion ?,* in Proc. Interspeech, 2019

[6] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda. *Dynamical Variational Autoencoders : A Comprehensive Review,* in Foundations and Trends in Machine Learning, vol. 15, 2021