RESEARCH CENTRE

Inria Nancy - Grand Est Center

IN PARTNERSHIP WITH:

CNRS, Université de Lorraine

2022

ACTIVITY REPORT

Project-Team

MULTISPEECH

Speech Modeling for Facilitating Oral-Based Communication

IN COLLABORATION WITH: Laboratoire lorrain de recherche en informatique et ses applications (LORIA)

DOMAIN

Perception, Cognition and Interaction

THEME

Language, Speech and Audio

Inria

Contents

Project-Team MULTISPEECH

Creation of the Project-Team: 2015 July 01

Keywords

Computer sciences and digital sciences

A3.4. – Machine learning and statistics

A3.4.6. – Neural networks

A3.4.8. – Deep learning

A3.5. – Social networks

A4.8. – Privacy-enhancing technologies

A5.1.5. – Body-based interfaces

A5.1.7. – Multimodal interfaces

A5.6.2. – Augmented reality

A5.7. – Audio modeling and processing

A5.7.1. – Sound

A5.7.3. – Speech

A5.7.4. – Analysis

A5.7.5. – Synthesis

A5.8. – Natural language processing

A5.9. – Signal processing

A5.9.1. – Sampling, acquisition

A5.9.2. – Estimation, modeling

A5.9.3. – Reconstruction, enhancement

A5.10.2. – Perception

A5.11.2. – Home/building control and interaction

A6.2.4. – Statistical methods

A6.3.1. – Inverse problems

A6.3.5. – Uncertainty Quantification

A9.2. – Machine learning

A9.3. – Signal analysis

A9.4. – Natural language processing

A9.5. – Robotics

Other research topics and application domains

B8.1.2. – Sensor networks for smart buildings

B8.4. – Security and personal assistance

B9.1.1. – E-learning, MOOC

B9.5.1. – Computer science

B9.5.2. – Mathematics

B9.5.6. – Data science

B9.6.8. – Linguistics

B9.6.10. – Digital humanities

B9.10. – Privacy

1 Team members, visitors, external collaborators

Research Scientists

- Anne Bonneau [CNRS, Researcher]
- Antoine Deleforge [INRIA, Researcher]
- Dominique Fohr [CNRS, Researcher]
- Denis Jouvet [INRIA, Senior Researcher, until Aug 2022, HDR]
- Yves Laprie [CNRS, Senior Researcher, HDR]
- Paul Magron [INRIA, Researcher]
- Mostafa Sadeghi [INRIA, Researcher]
- Emmanuel Vincent [INRIA, Senior Researcher, HDR]

Faculty Members

- Slim Ouni [Team leader, UL, Associate Professor, HDR]
- Vincent Colotte [UL, Associate Professor]
- Irina Illina [UL, Associate Professor, HDR]
- Romain Serizel [UL, Associate Professor, HDR]

Post-Doctoral Fellows

- Felix Gontier [INRIA]
- Ama Marina Kreme [INRIA, from Mar 2022]
- Marie-Anne Lacroix [UL, until Aug 2022 and INRIA, from Sep 2022]

PhD Students

- Louis Abel [UL]
- Tulika Bose [UL]
- Pierre Champion [INRIA]
- Can Cui [INRIA]
- Stephane Dilungana [INRIA]
- Sandipana Dowerah [INRIA]
- Adrien Dufraux [FACEBOOK, until Apr 2022]
- François Effa [Univ de Lyon]
- Ashwin Geet D'sa [UL, until May 2022]
- Mickaella Grondin-Verdon [CNRS]
- Seyed Hosseini [UL]
- Nasser-Eddine Monir [UL, from Dec 2022]

- Sewade Ogun [INRIA]
- Michel Olvera [INRIA]
- Robin San Roman [Meta AI, CIFRE, from Oct 2022]
- Shakeel Sheikh [UL]
- Vinicius Souza Ribeiro [UL, until Aug 2022]
- Tom Sprunck [INRIA]
- Prerak Srivastava [INRIA]
- Nicolas Zampieri [INRIA, ATER]
- Georgios Zervakis [INRIA]

Technical Staff

- Sofiane Azzouz [CNRS, Engineer]
- Hugo Bergerat [UL, Technician, from Jul 2022 until Aug 2022]
- Théo Biasutto-Lervat [INRIA, Engineer]
- Joris Cosentino [Inria, until Oct 2022]
- Louis Delebecque [UL, until Mar 2022 and CNRS, Engineer, from Mar 2022]
- Hubert Nourtel [INRIA, Engineer]
- Francesca Ronchini [Inria, Engineer, until May 2022]

Interns and Apprentices

- Louis Bahrman [INRIA, Intern, from Mar 2022 until Sep 2022]
- Colleen Beaumard [INRIA, Intern, from Mar 2022 until Jul 2022]
- Hugo Bergerat [UL, Intern, from Apr 2022 until Jun 2022]
- Hugo Breniaux [UL, Intern, until Apr 2022]
- Emre Canbazer [UL, Intern, from Feb 2022 until Jul 2022]
- Romain Chevret [UL, Intern, from May 2022 until Jun 2022]
- Valentin Gerard [UL, Intern, from Oct 2022]
- Ali Golmakani [INRIA, Intern, from Apr 2022 until Aug 2022]
- Soklay Heng [UL, Intern, from Mar 2022 until Aug 2022]
- Rasul Jasir Dent [INRIA, Intern, from Feb 2022 until Aug 2022]
- Thibault Odor [UL, Intern, from May 2022 until Jul 2022]
- Maeva Touchet [UL, Intern, from Mar 2022 until Aug 2022]

Administrative Assistants

- Helene Cavallini [INRIA]
- Delphine Hubert [UL]
- Anne-Marie Messaoudi [CNRS]

2 Overall objectives

Since the beginning of the year 2022, we have started to implement the new team project. Although we have not changed the name of the team, the objectives and the organization of its axes are substantially modified.

In Multispeech, we consider speech as a multimodal signal with different facets: acoustic, facial, articulatory, gestural, etc. Historically, speech was mainly considered under its acoustic facet, which is still the most important one. However, the acoustic signal is a consequence of the temporal evolution of the shape of the vocal tract (pharynx, tongue, jaws, lips, etc.), that is the articulatory facet of speech. The shape of the vocal tract is partly visible on the face, that is the main visual facet of speech. The face can provide additional information on the speaker's state through facial expressions. Speech can be accompanied by gestures (head nodding, arm and hand movements, etc.), that help to clarify the linguistic message. In some cases, such as in sign language, these gestures can bear the main linguistic content and be the only means of communication.

The general objective of Multispeech is to study the analysis and synthesis of the different facets of this multimodal signal and their multimodal coordination in the context of human-human or human-computer interaction. While this multimodal signal carries all of the information used in spoken communication, the collection, processing and extraction of meaningful information by a machine system remains a challenge. In particular, to operate in real-world conditions, such a system must be robust to noisy or missing facets. We are especially interested in designing models and learning techniques that rely on limited amounts of labeled data and that preserve privacy.

Therefore, Multispeech addresses data-efficient, privacy-preserving learning methods, and the robust extraction of various streams of information from speech signals. These two axes will allow us to address multimodality, i.e., the analysis and the generation of multimodal speech and its consideration in an interactional context.

The outcomes will crystallize into a unified software platform for the development of embodied voice assistants. Our main objective is that the results of our research feed this platform, and that the platform itself facilitates our research and that of other researchers in the general domain of human-computer interaction, as well as the development of concrete applications that help humans to interact with one another or with machines. We will focus on two main application areas: language learning and health assistance.

3 Research program

3.1 Axis 1 — Data-efficient and privacy-preserving learning

A central aspect of our research is to design machine learning models and methods for multimodal speech data, whether acoustic, visual or gestural. By contrast with big tech companies, we focus on scenarios where the amount of speech data is limited and/or access to the raw data is infeasible due to privacy requirements, and little or no human labels are available.

3.1.1 Axis 1.1 — Integrating domain knowledge

State-of-the art methods for speech and audio processing are based on discriminative neural networks trained for the targeted task. This paradigm faces major limitations: lack of interpretability, large data requirements, inability to generalize to unseen classes or tasks. Our approach is to combine the representation power of deep learning with our acoustic expertise to obtain smaller generative models

describing the probability distribution of speech and audio signals. Particular attention will be paid to designing physically-motivated input layers, output layers, and unsupervised representations that capture complex-valued, multi-scale spectro-temporal dependencies. Given these models, we derive computationally efficient inference algorithms that address the above limitations. We also explore the integration of deep learning with symbolic reasoning and common-sense knowledge to increase the generalization ability of deep models.

3.1.2 Axis 1.2 — Learning from little/no labeled data

While supervised learning from fully labeled data is economically costly, unlabeled data are inexpensive but provide intrinsically less information. Our goal is to learn representations that disentangle the attributes of speech by equipping the unsupervised representation learning methods above with supervised branches exploiting the available labels and supervisory signals, and with multiple adversarial branches overcoming the usual limitations of adversarial.

3.1.3 Axis 1.3 — Preserving privacy

To preserve privacy, speech must be transformed to hide the users' identity and other privacy-sensitive attributes (e.g., accent, health status) while leaving intact those attributes which are required for the task (e.g., phonetic content for automatic speech recognition) and preserving the data variability for training purposes. We develop strong attacks to evaluate the privacy. We also seek to hide personal identifiers and privacy-sensitive attributes in the linguistic content, focusing on their robust extraction and replacement from speech signals.

3.2 Axis 2 — Extracting information from speech signals

In this axis, we focus on extracting meaningful information from speech signals in real conditions. This information can be related (1) to the linguistic content, (2) to the speaker, and (3) to the speech environment.

3.2.1 Axis 2.1 — Linguistic speech content

Speech recognition is the main means to extract linguistic information from speech. Although it is a mature research area, performance drops in real-world environments pursue the development of speech enhancement and source separation methods to effectively improve robustness in such real-world scenarios. Semantic content analysis is required to interpret the spoken message. The challenges include learning from little real data, quickly adapting to new topics, and robustness to speech recognition errors. The detection and classification of hate speech in social media videos will also be considered as a benchmark, thereby extending the work on text-only detection. Finally, we also consider extracting phonetic and prosodic information to study the categorization of speech sounds and certain aspects of prosody by learners of a foreign language.

3.2.2 Axis 2.2 — Speaker identity and states

Speaker identity is required for personalization of human-computer interaction. Speaker recognition and diarization are still challenging in real-world conditions. The speaker states that we aim to recognize include emotion and stress, which can be used to adapt the interaction in real time.

3.2.3 Axis 2.3 — Speech environment information

We develop audio event detection methods that exploit both strongly/weakly labeled and unlabeled data, operate in real-world conditions, can discover new events, and provide a semantic interpretation. Modeling of the temporal, spatial and logical structure of ambient sound scenes over a long duration is also considered. We are also interested in the lesser studied problem of inferring the acoustic properties of the environment from impulse response measurements or from multichannel recordings of unknown sound sources.

3.3 Axis 3 — Multimodal Speech: generation and interaction

In our project, we consider speech as a multimodal object, where we study (1) multimodality modeling and analysis, focusing on multimodal fusion and coordination, (2) the generation of multimodal speech by taking into account its different facets (acoustic, articulatory, visual, gestural), separately or combined, and (3) interaction, in the context of human-human or human-computer interaction.

3.3.1 Axis 3.1 - Multimodality modeling and analysis

The study of multimodality concerns the interaction between modalities, their fusion, coordination and synchronization for a single speaker, as well as their synchronization across the speakers in a conversation. We focus on audiovisual speech enhancement to improve the intelligibility and quality of noisy speech by considering the speaker's lip movements. We also consider the semi/weakly/self-supervised learning methods for multimodal data so as to obtain interpretable representations that disentangle in each modality the attributes related to linguistic and semantic content, emotion, reaction, etc. We also study the contribution of each modality to the intelligibility of spoken communication.

3.3.2 Axis 3.2 - Multimodal speech generation

Multimodal speech generation refers to articulatory, acoustic, and audiovisual speech synthesis techniques which output one or more facets. Articulatory speech synthesis relies on 2D and 3D modeling of the dynamics of the vocal tract from real-time MRI (rtMRI) data. We consider the generation of the full vocal tract, from the vocal folds to the lips, first in 2D then in 3D. This comprises the generation of the face and the prediction of the glottis opening. We also consider audiovisual speech synthesis. Both the animation of the lower part of the face related to speech and of the upper part related to the facial expression are considered, and development continues towards a multilingual talking head. We investigate further the modeling of expressivity for both audio-only and audiovisual speech synthesis, for a better control of expressivity, where we consider several disentangled attributes at the same time.

3.3.3 Axis 3.3 — Interaction

Interaction is a new field of research for our project-team that we will approach gradually. We start by studying the multimodal components (prosody, facial expressions, gestures) used during interaction, both by the speaker and by the listener, where the goal is to simultaneously generate speech and gestures by the speaker, and generating regulatory gestures for the listener. We will introduce different dialog bricks progressively: Spoken language understanding, Dialog management, and Natural language generation. Dialog will be considered in a multimodal context (gestures, emotional states of the interlocutor, etc.) and we will break the classical dialog management scheme to dynamically account for the interlocutor's evolution during the speaker's response.

3.4 Software platform: Multimodal Voice assistant

The outcomes of the approaches and models in this research program will crystallize into a unified software platform for the development of embodied voice assistants. Our main objective is that the results of our research feed this platform, and that the platform itself facilitates our research and that of other researchers in the general domain of human-computer interaction, as well as the development of concrete applications that help humans to interact with one another or with machines. We will focus on two main application areas: language learning and health assistance.

4 Application domains

The approaches and models developed in Multispeech will have several applications to help humans interact with one another or with machines. Each application will typically rely on an embodied voice assistant developed via our generic software platform or on individual components, as presented above. We will put special effort on two application domains: language learning and health assistance. We chose

these domains mainly because of their economic and social impact. Moreover, many outcomes of our research will be naturally applicable in these two domains, which will help us showcase their relevance.

4.1 Language Learning

Learning a second language, or acquiring the native language for people suffering from language disorders, is a challenge for the learner and represents a significant cognitive load. Many scientific activities have therefore been devoted to these issues, both from the point of view of production and perception. We aim to show the learner (native or second language) how to articulate the sounds of the target language by illustrating articulation with a talking head augmented by the vocal tract which allows animating the articulators of speech. Moreover, based on the analysis of the learner's production, an automatic diagnosis can be envisaged. However, reliable diagnosis remains a challenge, which depends on the accuracy of speech recognition and prosodic analysis techniques. This is still an open question.

4.2 Health Assistance

Speech technology can facilitate healthcare access to all patients and it provides an unprecedented opportunity to transform the healthcare industry. This includes speech disorders and hearing impairments. For instance, it is possible to use automatic techniques to diagnose disfluencies from an acoustic or an audiovisual signal, as in the case of stuttering. Speech enhancement and separation can enhance speech intelligibility for hearing aid wearers in complex acoustic environments, while articulatory feedback tools can be beneficial for articulatory rehabilitation of cochlear implant wearers. More generally, voice assistants are a valuable tool for senior or disabled people, especially for those who are unable to use other interfaces due to lack of hand dexterity, mobility, and/or good vision. Speech technologies can also facilitate communication between hospital staff and patients, and help emergency call operators triage the callers by quantifying their stress level and getting the maximum amount of information automatically thanks to a robust speech recognition system adapted to these extreme conditions.

5 Social and environmental responsibility

A. Deleforge chairs the *Commission pour l'Action et la Responsabilité Ecologique* (CARE), formerly called the *Commission Locale de Développement Durable*, a joint entity between Loria and Inria Nancy. Its goals are to raise awareness, guide policies and take action at the lab level and to coordinate with other national and local initiatives and entities on the subject of the environmental impact of science, particularly in information technologies.

M.-A. Lacroix is working on the compression of large Wav2vec 2.0 audio models for embedded devices. Her work is applied to bird monitoring.

T. Biasutto-Lervat also paid special attention to the memory and computational footprint of speech recognition and synthesis models in the context of the development of the team's software platform for embodied voice assistants.

R. Serizel et al. [75] performed an extensive study about energy consumption used to train a sound event detection model for different GPU types and batch sizes. The goal was to identify which aspects can have an impact on the estimation of the energy consumption and should be normalized for a fair comparison across systems. Additionally, they proposed an analysis of the relationship between the energy consumption and the sound event detection performance that calls into question the current way to evaluate systems.

6 Highlights of the year

6.1 Awards

E. Vincent received the ISCA Award for the best paper published in Computer Speech and Language (2017–2021) [79].

B. M. L. Srivastava's startup project "Nijta" was awarded one of the 10 Grand Prizes of the i-PhD Innovation Challenge organized by the French Ministry of Higher Education, Research and Innovation in partnership with Bpifrance.

7 New software and platforms

7.1 New software

7.1.1 ASTALI

Name: Automatic Speech-Text Alignment Software

Keyword: Speech-text alignment

Functional Description: ASTALI is a software for aligning a speech signal with its corresponding orthographic transcription (given in simple text file for short audio signals or in .trs files as generated by transcriber for longer speech signals). Using a phonetic lexicon and automatic grapheme-to-phoneme converters, all the potential sequences of phones corresponding to the text are generated. Then, using acoustic models, the tool finds the best phone sequence and provides the boundaries at the phone and at the word levels. The web application makes the service easy to use, without requiring any software downloading.

News of the Year: The application has been migrated to a new, more robust web server. The application is still operational and actively used.

URL: http://astali.loria.fr/

Contact: Theo Biasutto–lervat

7.1.2 Asteroid

Name: Asteroid: The PyTorch-based audio source separation toolkit for researchers.

Keywords: Source Separation, Deep learning

Functional Description: Asteroid is an open-source toolkit made to design, train, evaluate, use and share neural network based audio source separation and speech enhancement models. Inspired by the most successful neural source separation systems, Asteroid provides all neural building blocks required to build such a system. To improve reproducibility, Kaldi-style recipes on common audio source separation datasets are also provided. Experimental results obtained with Asteroid's recipes show that our implementations are at least on par with most results reported in reference papers.

Release Contributions: [0.6.0] - 2022-06-28 Upgraded source code and recipes to PyTorch-Lighting 1.5+.

[0.5.3] - 2022-06-26 Added: [egs&tests] MixIT loss function (#595)

Changed: [docs] Update RTD version to 0.5.1

Fixed: [src] Fix FasNetTAC loading (#586) [src] Fix device in padding for sudormrf. Fix #598 (#603) [docs] Fix deep_clustering_loss docs example (#607) [src] Remove torch.complex32 usage for torch 1.11.0 (#609) [egs] Fix package install script for WHAMR! (#613) [src] Fix shape mismatching in SuDORMRF's masknn (#618) [CI] Fix CI Restrict torchmetrics version to under 0.8.0 (#619) [docs] Fix docs Restrict jinja2>=3.0.0,<3.1.0 (#620)

News of the Year: - Upgraded source code and recipes to PyTorch-Lighting 1.5+. - Added the MixIT loss function

URL: https://github.com/asteroid-team/asteroid

Contact: Antoine Deleforge

Participants: Manuel Pariente, Mathieu Hu, Joris Cosentino, Sunit Sivasankaran, Mauricio Michel Olvera Zambrano, Fabian Robert Stoter

7.1.3 VocabLearn

Name: Web-based Pronunciation Learning Application

Keywords: Pronunciation training, Talking head, Second language learning

Scientific Description: This platform highlights our work on realistic animation of a talking head from speech (also called lipsync). Our lipsync system is operational for German. The evaluation of pronunciation is based on our work on speech recognition. The work on evaluation is not fully completed.

Functional Description: This web-based application is dedicated to foreign language pronunciation learning (current version was developed for the German language). It is intended for high school and middle school students. There are two types of exercises that are integrated in this application. (1) Flashcards: Cards are presented, then a virtual teacher (a 3D talking head) pronounces the words and sentences corresponding to these cards. Students can practice and make an evaluation of their word comprehension. (2) Speech recognition. The application displays a list of words/phrases that the student pronounces and the system gives feedback on the quality of the pronunciation. This application is composed of two modules: one for students (described above) and one for teachers, allowing them to create lessons, and to follow the results and progress of student evaluations.

News of the Year: A first online version of the flash cards application is deployed. Both versions: Student and teacher are fully running. New features have been added to the administration interface to add cards and to generate new videos of the virtual teacher.

Contact: Slim Ouni

Participants: Theo Biasutto–lervat, Slim Ouni, Denis Jouvet, Louis Abel

7.1.4 Voice Transformer 2

Keywords: Speech, Privacy

Scientific Description: The implemented method is inspired from the speaker anonymisation method proposed in [Fan+19], which performs voice conversion based on x-vectors [Sny+18], a fixed-length representation of speech signals that form the basis of state-of-the-art speaker verification systems. We have brought several improvements to this method such as pitch transformation, and new design choices for x-vectors selection

[Fan+19] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.F. Bonastre. “Speaker Anonymization Using x-vector and Neural Waveform Models”. In: Proceedings of the 10th ISCA Speech Synthesis Workshop. 2019, pp. 155–160. [Sny+18] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. “X-vectors: Robust DNN embeddings for speaker recognition”. In: Proceedings of ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018, pp. 5329–5333.

Functional Description: Voice Transformer increases the privacy of users of voice interfaces by converting their voice into another person’s voice without modifying the spoken message. It ensures that any information extracted from the transformed voice can hardly be traced back to the original speaker, as validated through state-of-the-art biometric protocols, and it preserves the phonetic information required for human labelling and training of speech-to-text models.

Release Contributions: This version only retains the anonymization method using x-vectors and neural models from the previous version. The previous architecture inherited from kaldi based on bash scripts has been completely revised for a modular architecture, based on REST micro-services, implemented only in python and C++.

Contact: Nathalie Vauquier

Participants: Brij Mohan Lal Srivastava, Nathalie Vauquier, Emmanuel Vincent, Marc Tommasi

7.1.5 Web Multimodal Annotation

Name: Web-based Multimodal Data Annotation

Keywords: Annotation tool, Audio segmentation, Audiovisual, Web Application, Speech

Scientific Description: Web Multimodal Annotation is a web application that was designed to manually segment and label audio or video data and to visualize the audio data. In addition, this web application also allows for multi-level labeling and collaboration between multiple annotators. This web application can be useful to annotators, researchers in the field of multimodal speech or gestures.

Functional Description: Web Multimodal Annotation is a web application that was designed to manually segment and label audio or video data and to visualize the audio data. In addition, this web application also allows for multi-level labeling and collaboration between multiple annotators. This web application can be useful to annotators, researchers in the field of multimodal speech or gestures.

Release Contributions: The software is deployed on a web server. This version is quite mature, but requires evaluation by users.

Contact: Slim Ouni

Participants: Slim Ouni, Sofiane Azzouz

Partners: CNRS, Université de Lorraine

7.2 New platforms

7.2.1 Multimodal Voice assistant

Voice assistants and voice interfaces have become a key technology, simplifying the user experience and increasing the accessibility of many applications, and their use will intensify in the coming years. However, this technology poses two major problems today: on the one hand, the quasi-hegemony of large technology companies (mainly American) raises questions about European digital sovereignty, and on the other hand, the commonly used client-server architecture raises privacy risks. To simultaneously address these two problems, we are currently developing an open-source platform for the creation of embedded virtual assistants.

This platform will provide the main speech processing and natural language processing bricks that are necessary to build a voice interface, such as denoising, recognition or speech synthesis. The generated assistant will be fully embedded in the users' terminal. The data being processed locally, we ensure the protection of their private lives. We envisage a multiplatform solution on PC (Windows, Linux, MacOS) as well as on mobile (Android, iOS).

During this first year of development, many points have been addressed and evaluated, in particular:

- benchmarking and choice of software tools (wasm, ONNX...)
- benchmarking (performance and computing power) of speech recognition and synthesis models (from NVIDIA's NeMo repository)
- prototyping of a cross-platform PC solution (Python)
- prototyping of an Android solution (Kotlin)
- preliminary design of the API

8 New results

8.1 Axis 1 — Data-efficient and privacy-preserving learning

Participants: Antoine Deleforge, Denis Jouvet, Emmanuel Vincent, Vincent Colotte, Irina Illina, Romain Serizel, Marie-Anne Lacroix, Pierre Champion, Adrien Dufraux, Ajinkya Kulkarni, Sewade Olaolu Ogun, Robin San Roman, Georgios Zervakis, Hubert Nourtel, Mostafa Sadeghi, Paul Magron, Marina Krémé.

8.1.1 Axis 1.1 — Integrating domain knowledge

Integration of signal processing knowledge. We developed a probabilistic framework to incorporate prior knowledge of latent variables in a VAE-based speech generative model, using a sparsity promoting dictionary model [52]. This is different from the standard assumption of normal prior and it allows us to put some structure on the latent space, which could result in more efficient and interpretable models. Experiments on generative speech modeling have shown that the proposed approach outperforms previous techniques, since it increases sparsity without compromising the quality of the generated speech.

Integration of symbolic knowledge. Transformer based language models embed a vast amount of linguistic and commonsense knowledge, but their ability to reason over this knowledge is limited. We proposed a new neural network architecture for the task of target sense verification, which consists of deciding whether the sense of a word in a given context matches a candidate definition and corresponding hypernyms. The proposed architecture implements a form of analogical reasoning via a two-layer convolutional neural network which outperforms existing approaches and increases interpretability [60]. In parallel, we tackled the issue of linguistic ambiguities arising from changes in entities in videos. Focusing on instructional cooking videos as a challenging use case, we proposed new guidelines to annotate recipes for the anaphora resolution task which reflect such changes, and we designed and evaluated an end-to-end multimodal anaphora resolution system against this new annotation scheme [47].

Interfacing optimization and deep learning. Optimization-based approaches for speech processing are interesting since they require little or no training data, and generally exhibit more robustness to acoustic conditions and other variability factors than deep learning systems. We derived optimization-based algorithms for data-efficient speech signal restoration [68]. In such a framework, light neural networks can be used as a proxy to obtain appropriate prior information and/or initialization. We also interfaced optimization- and learning-based algorithms in a deep unfolding paradigm in order to design a very efficient system for audio signal recovery [21].

8.1.2 Axis 1.2 - Learning from little/no labeled data

Learning from noisy data. Training of multi-speaker text-to-speech (TTS) systems relies on high-quality curated datasets, which lack speaker diversity and are expensive to collect. As an alternative, we proposed to automatically select high-quality training samples from large, readily available crowdsourced automatic speech recognition (ASR) datasets using a non-intrusive perceptual mean opinion score estimator. Our approach results in improved quality with respect to training on a curated dataset, and it opens the door to automatic TTS dataset curation for a wider range of languages [46].

Learning from noisy labels. ASR systems are typically trained in a supervised fashion on manually labeled data. Semi-supervised learning and transfer learning approaches reduce the labeling cost but achieve limited performance. In his PhD thesis [64], Adrien Dufraux explored the middle ground where the training data are neither accurately labeled nor unlabeled but a not-so-expensive "noisy" transcription is available instead. In particular, he expressed the Lead2Gold loss he had previously proposed in the more flexible weighted finite state transducer formalism, and he collected a real crowdsourced labeled

dataset. Along a similar line, we proposed a sampling method to train and adapt Transformer based language models on uncertain ASR hypotheses [53].

Self-supervised learning. We started working on reducing the footprint of Wav2vec 2.0 in the line of [41]. We also started working on learning compressed representations for audio signals and efficient reconstruction from these compressed representations with diffusion approaches.

Transfer learning applied to speech synthesis. We worked on the disentanglement of speaker, emotion and content for transferring expressivity information from one speaker to another one, particularly when only neutral speech data is available for the latter. A deep metric learning framework based on multiclass n-pair loss has been used for improving the latent representation of expressivity in a multispeaker text-to-speech system setting, which results in improved expressivity transfer. Multi-stage attention was introduced for fine-grained expressivity transfer [40]. Also, expressivity transfer has been analyzed in detail in the framework of non-autoregressive end-to-end multispeaker TTS systems that rely on deep generative Flow models and diffusion probabilistic models [39].

8.1.3 Axis 1.3 - Preserving privacy

Speech signals convey a lot of private information. To protect speakers, we pursued our investigation of x-vector based voice anonymization, which relies on splitting the speech signal into speaker (x-vector), phonetic and pitch features and resynthesizing the signal with a different target x-vector. We conducted an extensive study of the impact of the target x-vector selection process (speaker distance metric, target region of x-vector space, target gender, speaker- or utterance-level target selection) on privacy and utility [19]. To reduce the amount of residual speaker information in the phonetic and pitch features, we explored the use of vector quantization [28, 27] and Laplacian noise [16] inspired from differential privacy. We also evaluated the impact of voice anonymization on emotional speech data [45]. Finally, we showed that slicing utterances into shorter segments further improves privacy at no cost in utility [42].

In a complementary line of work, we studied the adaptation of ASR language models trained on anonymized text data to the statistics of the original text data [57].

We analyzed the results of the 1st Voice Privacy Challenge which we had organized in 2020 in an article [20] and a detailed technical report [78]. We co-organized the 2nd Voice Privacy Challenge [77], which improves the evaluation of voice anonymization systems in several ways, namely ranking based on multiple privacy-utility tradeoffs, semi-informed attack models, and an intonation preservation constraint.

8.2 Axis 2 — Extracting information from speech signals

Participants: Antoine Deleforge, Dominique Fohr, Denis Jouvet, Emmanuel Vincent, Irina Illina, Romain Serizel, Anne Bonneau, Félix Gontier, Ama Marina Krémé, Tulika Bose, Can Cui, Stephane Dilungana, Sandipana Dowerah, François Effa, Nasser-Eddine Monir, Mauricio Michel Olvera Zambrano, Tom Sprunck, Prerak Srivastava, Nicolas Zampieri, Joris Cosentino, Louis Delebecque.

8.2.1 Axis 2.1 — Linguistic speech content

Dyslexia and non-native speech It is generally admitted that dyslexic people have a lower performance than non-dyslexic individuals with respect to the categorization and discrimination of phonemes in their L1. Yet, some experiments tend to show that dyslexic people exhibit a better allophonic (infra-phonemic) discrimination, a phenomenon that has been explained by the hypothesis that people with dyslexia would not loose, during speech perception development, the infants' ability to detect universal features. This raises the question of how dyslexic people discriminate sounds of a foreign language. Do they perform better, as the hypothesis of a "better sensitivity to universal features" suggests, or do they keep a lower

level of discrimination/categorisation, as in their L1? We investigated the perception of German fricatives by French dyslexic subjects. The targeted sounds were /s/ and /sh/, present in the French and German systems, and the voiceless palatal /ç/ (the final sound in “ich”), absent in French. The sounds were presented in an AX discrimination task and a categorization task to 20 non-dyslexic and 20 dyslexic young adults with no knowledge of German. Results confirm the existence of a (slight) deficit in categorization for dyslexic people, and show that the new sound (/ç/) was more difficult to discriminate for them, which invalidates, for the sounds investigated here, the hypothesis of a better sensitivity to new contrasts. Large variations among dyslexic individuals were observed [62].

Detection of hate speech in social media. The wide usage of social media has given rise to the problem of online hate speech. Deep neural network-based classifiers have become the state-of-the-art for automatic hate speech classification. The performance of these classifiers depends on the amount of available labelled training data. However, most hate speech corpora have a small number of hate speech samples. We aimed to jointly use multiple hate speech corpora to improve hate speech classification performance in low-resource scenarios. We harness different hate speech corpora in a multi-task learning setup by associating one task to one corpus. This multi-corpus learning scheme is expected to improve the generalization, the latent representations, and the domain adaptation of the model. Our work evaluated multi-corpus learning for hate speech classification and domain adaptation. We showed significant improvements in classification and domain adaptation in low-resource scenarios [33] [65].

In [51], we presented the M-Phasis corpus, a corpus of 9k German and French user comments collected from migration-related news articles. It goes beyond the "hate"-"neutral" dichotomy and is instead annotated with 23 features, which in combination become descriptors of various types of speech, ranging from critical comments to implicit and explicit expressions of hate. The annotations are performed by 4 native speakers per language and achieve high ($0.77 \leq k \leq 1$) inter-annotator agreements.

Multiword expression (MWE) identification in tweets is a complex task due to the complex linguistic nature of MWEs combined with the non-standard language use in social networks. MWE features were shown to be helpful for hate speech detection (HSD). In [59] [58], we presented joint experiments on these two related tasks on English Twitter data: first we focus on the MWE identification task, and then we observed the influence of MWE-based features on the HSD task. For MWE identification, we compared the performance of two systems: lexicon-based and deep neural network-based (DNN). The DNN-based system outperformed the lexicon-based one thanks to its superior generalisation power, yielding much better recall. For the HSD task, we proposed a new DNN architecture for incorporating MWE features. We confirmed that MWE features are helpful for the HSD task.

Hate speech classifiers exhibit substantial performance degradation when evaluated on datasets different from the source. This is due to learning spurious correlations between words that are not necessarily relevant to hateful language, and hate speech labels from the training corpus. We proposed to automatically identify and reduce spurious correlations using attribution methods with dynamic refinement of the list of terms that need to be regularized during training. Our approach is flexible and improves the cross-corpora performance over previous work independently and in combination with pre-defined dictionaries [25].

In the context of out-of-domain settings, we proposed a domain adaptation approach that automatically extracts and penalizes source-specific terms using a domain classifier, which learns to differentiate between domains, and feature-attribution scores for hate-speech classes, yielding consistent improvements in cross-domain evaluation [24].

In [26], we proposed a novel training strategy that allows flexible modeling of the relative proximity of neighbors retrieved from a resource-rich corpus to learn the amount of transfer. In particular, we incorporated neighborhood information with Optimal Transport, which allows us to exploit the geometry of the data embedding space.

Introduction of semantic information in an ASR system We aim to improve ASR performance by modeling long-term semantic relations. We proposed to perform this through DNN-based rescoring of the ASR n-best hypotheses, that combine semantic, acoustic, and linguistic information. Our DNN rescoring models are aimed at selecting hypotheses that have better semantic consistency and therefore lower word error rate. We investigated a powerful representation as part of input features to our DNN

model: dynamic contextual embeddings from BERT [34].

8.2.2 Axis 2.2 — Speaker identity and states

Speaker recognition. We investigated the robustness of speaker recognition systems with respect to environmental noises and reverberation. One approach was based on the investigation of speaker embeddings (x-vectors) that are robust to noise [44]. A comprehensive exploration of noise robustness and noise compensation in ResNet and TDNN-based speaker recognition systems has been conducted [43]. The other approach relies on enhancing the multichannel speech signal before feeding the enhanced signal to the speaker verification system. Deep neural networks are used in the speech enhancement process. Experiments have shown that speech enhancement preprocessing improves speaker verification performance provided that trial and enrollment utterances exhibit similar SNR levels, and are processed in the same way [31]. Diffusion probabilistic models have also been investigated for multichannel speech enhancement as a front-end for a state-of-the-art ECAPA-TDNN speaker verification system. Results show that a joint training of the two modules leads to better performance than separate training of the enhancement and of the speaker verification models [30].

Identifying disfluency in stuttered speech. Stuttering is a speech disorder during which the flow of speech is interrupted by involuntary pauses and repetition of sounds. Stuttering identification is an interesting interdisciplinary domain research problem which involves pathology, psychology, acoustics, and signal processing that makes it hard and complicated to detect [17]. Within the ANR project BENEPHIDIRE, the goal is to automatically identify typical kinds of stuttering disfluency using acoustic and visual cues for their automatic detection. This year, we proposed end-to-end and speech embedding based systems trained in a self-supervised fashion. In particular, we exploited the embeddings from the pre-trained Wav2Vec2.0 model for stuttering detection (SD) on the KSoF dataset. After embedding extraction, we benchmarked with several methods for SD. Our proposed self-supervised based SD system achieved good performance during the ACM Multimedia 2022 ComParE Challenge, specifically the stuttering sub-challenge [55].

We have also investigated the impact of multi-task (MTL) and adversarial learning (ADV) to learn robust stutter features. This is the first-ever preliminary study where MTL and ADV have been employed in stuttering identification (SI). We have evaluated our system on the SEP-28k stuttering dataset consisting of 20 hours of data from 385 podcasts. Our methods showed promising results and outperform the baseline in various disfluency classes. [54]

8.2.3 Axis 2.3 — Speech in its environment

Ambient sound recognition. Audio scene analysis systems face performance degradation when trained and tested on data recorded by different devices. To address this issue, we thoroughly analyzed the impact of normalization and moment matching strategies to compensate for the linear distortion introduced by the recording device and integrated them with adversarial domain adaptation to handle the remaining non-linear distortion [48]. Michel Olvera successfully defended his PhD on the broader topic of robust audio scene analysis [48].

Pursuing our involvement in the community on ambient sound recognition, we co-organized a task on sound event detection and separation as part of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2022 Challenge and published a detailed analysis of the submissions to the previous iteration of this task in 2021 [81]. In 2022, we introduced an energy consumption metric in order to raise awareness about the footprint of algorithms. In relation with this aspect, we measured the energy consumption of the baseline on several devices and for different hyperparameter values in order to define good practices to compare energy consumption of challenge submissions [75].

We also continued working on the automatic audio captioning. We participated in the organization of the audio-captioning task within the DCASE challenge. We also worked on proposing new metrics to evaluate captioning systems.

Speech enhancement. Following the work done by Nicolas Furnon during his PhD , we investigated to which extent using signals obtained in simulated acoustic environments is relevant to evaluate speech

enhancement approaches compared to using real recorded signals. This study focused in particular on distributed algorithms. It was shown that simulated acoustic environments that do not take the head and torso of the person wearing hearing devices into account can provide unreliable performance estimation. A parallel corpus with simulated signals and recorded signals under similar acoustic conditions was designed for these experiments and will be released.

Acoustic Parameter Estimation. To estimate acoustic quantities of interest from speech signals such as the speech source locations or acoustic properties of the room (surface, volume, reverberation time), state-of-the-art methods rely on supervised deep learning. Due to the lack of sufficiently diverse and large annotated real datasets to train the models on, most existing approaches rely partially or exclusively on simulated data. However, very few studies carefully examine the impact of acoustic simulation realism on the generalization of trained models to real data. We contributed two such studies, one on blind room acoustic parameter estimation [56] and one on speech direction of arrival estimation [76], revealing that improving the realism of source, microphone and wall responses at train time consistently and significantly improved generalization to real data for all considered tasks.

In another line of work, we developed an optimization-based method to estimate the acoustic responses of the walls, floor and ceiling of a room from measured room impulse responses given the approximate room and setup geometry [29], and an optimization-based method to estimate all the acoustic reflection paths up to order 8 from a single multichannel room impulse response [18]. Both methods have been extensively tested on simulated data under idealized source, microphone and wall responses, and were shown to be robust to noise under random geometries.

8.3 Axis 3 — Multimodal Speech: generation and interaction

Participants: Théo Biasutto-Lervat, Vincent Colotte, Yves Laprie, Slim Ouni, Mostafa Sadeghi, Emmanuel Vincent, Louis Abel, Mickaella Grondin-Verdon, Seyed Ahmad Hosseini, Vinicius Souza Ribeiro, Sofiane Azzouz.

8.3.1 Axis 3.1 — Multimodality modeling and analysis

Face frontalization for visually assisted speech processing. Speech processing tasks that utilize visual information from a speaker's lips, such as enhancement and separation, typically require a front-facing view of the speaker to extract as much useful information as possible from the speaker's lip movements. Previous methods have not taken this into account and instead rely on data augmentation to improve robustness to different face poses, which can lead to increased complexity in the models. Recently, we developed a robust statistical frontalization technique [36] that alternates between estimating a rigid transformation (scale, rotation, and translation) and a non-rigid deformation between an arbitrarily viewed face and a face model. This technique has been tested and found to be effective for audio-visual speech enhancement, and has been further extended in [11] to include a dynamic face deformation model. The method has been extensively evaluated and compared with other state-of-the-art frontalization techniques, including those that use modern deep learning architectures, for lip-reading and speech enhancement tasks. The results confirmed the benefits of the proposed framework over the previous works.

8.3.2 Axis 3.2 — Multimodal speech generation

Construction of a rt-MRI (real-time Magnetic Resonance Imaging) dataset for French. This year, in collaboration with the IADI laboratory (P.-A. Vuissoz and K. Isaieva) , we recorded a very large dataset of 2,100 sentences in rt-MRI for a female and a male speaker. The image of the mid-sagittal slice of the vocal tract was acquired at 50 Hz when the subject was reading the sentences aloud. This database required the development of a blocking foam adapted to the speaker's head and the MRI antenna to ensure that the speaker had the same posture for each of the six acquisitions that were required. These data were

phonetically segmented and manually corrected to obtain a highly accurate segmentation over time and in accordance with the phonemes articulated.

Prediction of the vocal tract shape from a sequence of phonemes to be articulated. Last year we developed an approach to predict the shape of the vocal tract from the sequence of phonemes to be articulated [15]. The training is performed on an rt-MRI articulatory dataset (the images of the mid-sagittal slice acquired at 50 Hz together with the denoised speech signal) in which the articulators have been automatically tracked. Each articulator, for example the tongue, is represented by a vector of points and it is therefore difficult to impose a phonetic constraint because these points are independent. We have therefore chosen to approximate the shape of the tongue using an autoencoder whose latent space size has been chosen according to previous work on articulatory models. Thanks to this new model [49] it is possible to easily impose constraints on the critical articulators.

Accelerating the centerline processing of vocal tract shapes for articulatory synthesis Acoustic simulations used in the articulatory synthesis of speech take a series of vocal tract shapes as an input. Acoustic simulations assume a plane wave propagation, simplifying and limiting the calculation time. It is, therefore, necessary to split 2D vocal tract shapes into small tubes perpendicular to the centerline simulating the plane wave propagation. The algorithm developed previously used a time-consuming regularization step whose computation time was close to that of acoustic simulations. Therefore, we explored the possibility of using deep learning to perform this step and accelerate the whole synthesis process. We used a dataset with a large number of rt-MRI images and our regularizing algorithm for training. A deep learning regression strategy was tested [37] and turned out to predict the centerline with a very good accuracy.

Sign language Sign languages are rich visual languages with complex grammatical structures. As with any other natural language, they have their own unique linguistic and grammatical structures, which often do not have a one-to-one mapping to their spoken language counterparts. Computational sign language research lacks the large-scale datasets that enables immediate applicability. To date, most datasets have been suffering from small domains of discourse, e.g., weather forecasts, lack of the necessary inter- and intra-signer variance on shared content, limited vocabulary and phrase variance, and poor visual quality due to a low resolution, a motion blur and interlacing artifacts. We collected a large dataset that includes over 300 hours of signing News video footage of a German broadcaster. We processed the video to extract spatial human skeletal features for the face, hands and body, and textual transcription of the signing content. We have analyzed the data (signer-based sample labeling, statistical outlier distribution, measurement of undersigning quality, and calculation of landmark error rate). We proposed a multimodal Transformer-based cross-attention framework to annotate our corpus with the existing glossary annotations extracted from the DGS (mDGS) dataset.

Expressive speech synthesis This year we continued to work on expressive speech synthesis. The goal was to generate expressive speech by transfer learning. Quality evaluation of generated expressiveness has been conducted (in addition to quality evaluation of the transfer described in 8.1.2). All details are available in the thesis PhD [66].

8.3.3 Axis 3.3 — Interaction

Speaker gesture generation. Our goal is to study the multimodal components (prosody, facial expressions, gestures) used during interaction. We consider the simultaneous generation of speech and gestures by the speaker, where we consider non-verbal and verbal gestures.

In Louis Abel's PhD, we focus on non-verbal gesture generation (upper body and arms) from the acoustic signal. The motion representation is firstly autoencoded from positions to obtain a latent representation. The model is based on Graphical Neural Networks to leverage the constraints of the skeleton. To perform learning in the context of interaction, we aim to integrate dyadic gesture information for conditioning the networks. It is challenging to find this information in a corpus. In a second step, the motion itself is predicted with a flow-based architecture from the motion (latent) representation and

audio speech feature context. In Mickaëlla Grondin-Verdon's PhD, we focus more on dyadic gestures by analyzing specific parts of the gestures, the strokes (duration, intensity, alignment...). The goal is to feed models with this information to predict namely and generate gestures in a dialog context.

9 Bilateral contracts and grants with industry

9.1 Bilateral grants with industry

9.1.1 Meta AI

- Company: Meta AI (France)
- Duration: Nov 2018 – Feb 2022
- Participants: Adrien Dufraux, Emmanuel Vincent
- Abstract: This CIFRE grant funded the PhD of Adrien Dufraux on cost-effective weakly supervised learning for automatic speech recognition.

9.1.2 Vivoka

- Company: Vivoka (France)
- Duration: Oct 2021 – Oct 2024
- Participants: Can Cui, Mostafa Sadeghi, Emmanuel Vincent
- Abstract: This contract funds the PhD of Can Cui on joint and embedded automatic speech separation, diarization and recognition for the generation of meeting minutes.

9.1.3 Meta AI

- Company: Meta AI (France)
- Duration: May 2022 – Apr 2025
- Participants: Robin San Roman, Antoine Deleforge, Romain Serizel
- Abstract: This CIFRE grant funds the PhD of Robin San Roman on self-supervised disentangled representation learning of audio data for compression and generation.

10 Partnerships and cooperations

10.1 International research visitors

10.1.1 Visits of international scientists

- Archontis Politis, assistant professor at Tampere University (Finland), visited the team from Sep. 12 to Nov. 4, 2022.

10.2 European initiatives

10.2.1 H2020 & Horizon Europe

ADRA-E

Title: AI, Data and Robotics Ecosystem

Duration: Jul 2022 – Jun 2025

Partners:

- Universiteit van Amsterdam (Netherlands)
- Universiteit Twente (Netherlands)
- ATOS Spain SA (Spain)
- ATOS IT (Spain)
- Commissariat à l'énergie atomique et aux énergies alternatives (France)
- Trust-IT SRL (Italie)
- Commpla (Italie)
- Linkopings Universitet (Sweden)
- Siemens Aktiengesellschaft (Germany)
- Dublin City University (Ireland)
- Deutsche Forschungszentrum für Künstliche Intelligenz GmbH (Germany)
- National University of Ireland Galway (Ireland)
- AI, Data and Robotics Association (Belgium)
- Hrvatska udruga za umjetnu inteligenciju (Croatia)

Coordinator: Jozef Geurts (Inria)

Participant: Emmanuel Vincent

Summary: In tight liaison with the AI, Data and Robotics Association (ADRA) and the AI, Data and Robotics Partnership, ADRA-E aim to support convergence and cross-fertilization between the three communities so as to bootstrap an effective and sustainable European AI, Data and Robotics (ADR) ecosystem. Together with Marc Schoenauer (Inria's Deputy Director in charge of AI), Emmanuel Vincent is the scientific representative of Inria. He is involved in WP1 which aims to organize cross-community workshops.

HumanE-AI-Net

Title: Making artificial intelligence human-centric

Duration: Sep 2020 - Aug 2023

Partners: 53 institutions and companies all across Europe

Coordinator: Paul Lukowicz (DFKI/TU Kaiserslautern, Germany)

Participant: Slim Ouni

Summary: The objective of the EU HumanE AI Net project is to create a network that will exploit the synergies between the involved centers of excellence to develop the scientific foundations and technological advances to guide AI to benefit humans, both individually and societally, and that respects European ethical, cultural, legal and political values. The main challenge is to develop robust and reliable AI systems that can "understand" humans, adapt to complex real-world environments, and interact appropriately in complex social contexts. The goal is to facilitate the implementation of AI systems that enhance human capabilities and empower individuals and society as a whole. Slim Ouni represents LORIA/CNRS within the WP2 & WP3.

TAILOR

Title: Foundations of Trustworthy AI - Integrating Reasoning, Learning and Optimization

Duration: Sep 2020 – Aug 2024

Partners: 53 institutions and companies all across Europe

Coordinator: Fredrik Heintz (Linköpings Universitet)

Participant: Emmanuel Vincent

Summary: TAILOR aims to bring European research groups together in a single scientific network on the Foundations of Trustworthy AI. The four main instruments are a strategic roadmap, a basic research programme to address grand challenges, a connectivity fund for active dissemination, and network collaboration activities. Emmanuel Vincent is involved in privacy preservation research in WP3.

VISION

Title: Value and Impact through Synergy, Interaction and coOperation of Networks of AI Excellence Centres

Duration: Sep 2020 – Aug 2024

Partners:

- České Vysoké Učení Technické v Praze (Czech Republic)
- Deutsche Forschungszentrum für Künstliche Intelligenz GmbH (Germany)
- Fondazione Bruno Kessler (Italy)
- Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek (Netherlands)
- Intellera Consulting SRL (Italy)
- Thales SIX GTS France (France)
- Universiteit Leiden (Netherlands)
- University College Cork – National University of Ireland, Cork (Ireland)

Coordinator: Holger Hoos (Universiteit Leiden)

Participant: Emmanuel Vincent

Summary: VISION aims to connect and strengthen AI research centres across Europe and support the development of AI applications in key sectors. Together with Marc Schoenauer (Inria's Deputy Director in charge of AI), Emmanuel Vincent is the scientific representative of Inria. He is involved in WP2 which aims to produce a roadmap aimed at higher level policy makers and non-AI experts which outlines the high-level strategic ambitions of the European AI community.

10.2.2 Other european programs/initiatives

IMPRESS

Title: Improving Embeddings with Semantic Knowledge

Duration: Sep 2020 – Aug 2023

Partners:

- Inria MAGNET (Lille) and SEMAGRAMME (Nancy)
- Deutsche Forschungszentrum für Künstliche Intelligenz GmbH (Germany)

Coordinators: Pascal Denis (Inria MAGNET) and Ivana Kruijff-Korbayová (DFKI)

Participant: Emmanuel Vincent

Summary: The goals of IMPRESS are to investigate the integration of semantic and common sense knowledge into linguistic and multimodal word embeddings and the impact on selected downstream tasks. IMPRESS also develops open source software and lexical resources, focusing on video activity recognition as a practical testbed.

10.3 National initiatives

ANR JCJC DiSCogs

Title: Distant speech communication with heterogeneous unconstrained microphone arrays

Duration: Sep 2018 – Aug 2022

Coordinator: Romain Serizel (LORIA, Nancy)

Participants: Louis Delebecque, Nicolas Furnon, Irina Illina, Romain Serizel, Emmanuel Vincent

Collaborators: Télécom ParisTech, 7sensing

Abstract: The objective is to solve fundamental sound processing issues in order to exploit the many devices equipped with microphones that populate our everyday life. The solution proposed is to apply deep learning approaches to recast the problem of synchronizing devices at the signal level as a multi-view learning problem.

ANR DEEP-PRIVACY

Title: Distributed, Personalized, Privacy-Preserving Learning for Speech Processing

Duration: Jan 2019 - Jun 2023

Coordinator: Denis Jouvet until Aug 2022; Emmanuel Vincent from Sep 2022

Partners: LIUM (Le Mans), Inria MAGNET (Lille), LIA (Avignon)

Participants: Pierre Champion, Denis Jouvet, Hubert Nourtel, Emmanuel Vincent

Abstract: The objective of the DEEP-PRIVACY project is to elaborate a speech transformation that hides the speaker identity for an easier sharing of speech data for training speech recognition models; and to investigate speaker adaptation and distributed training.

ANR ROBOVOX

Title: Robust Vocal Identification for Mobile Security Robots

Duration: Mar 2019 – Apr 2024

Coordinator: Laboratoire d'informatique d'Avignon (LIA)

Partners: Inria (Nancy), LIA (Avignon), A.I. Mergence (Paris)

Participants: Antoine Deleforge, Sandipana Dowerah, Denis Jouvet, Romain Serizel

Abstract: The aim of ROBOVOX project is to improve speaker recognition robustness for a security robot in real environment. Several aspects will be particularly considered such as ambiant noise, reverberation and short speech utterances.

ANR BENEPHIDIRE

Title: Stuttering: Neurology, Phonetics, Computer Science for Diagnosis and Rehabilitation

Duration: Mar 2019 - Dec 2023

Coordinator: Praxiling (Montpellier)

Partners: Praxiling (Montpellier), LORIA (Nancy), INM (Montpellier), LiLPa (Strasbourg).

Participants: Yves Laprie, Slim Ouni, Shakeel Ahmad Sheikh

Abstract: The BENEPHIDIRE project brings together neurologists, speech-language pathologists, phoneticians, and computer scientists specializing in speech processing to investigate stuttering as a speech impairment and to develop techniques for diagnosis and rehabilitation.

ANR LEAUDS

Title: Learning to understand audio scenes

Duration: Apr 2019 - Mar 2023

Coordinator: Université de Rouen Normandie

Partners: Université de Rouen Normandie, Inria (Nancy), Netatmo (Paris)

Participants: Félix Gontier, Mauricio Michel Olvera Zambrano, Romain Serizel, Emmanuel Vincent

Abstract: LEAUDS aims to make a leap towards developing machines that understand audio input through breakthroughs in the detection of audio events from little annotated data, the robustness to "out-of-the lab" conditions, and language-based description of audio scenes. MULTISPEECH is responsible for research on robustness and for bringing expertise on natural language generation.

Inria Project Lab HyAIAI

Title: Hybrid Approaches for Interpretable AI

Duration: Sep 2019 - Aug 2023

Coordinator: Inria LACODAM (Rennes)

Partners: Inria TAU (Saclay), SEQUEL, MAGNET (Lille), MULTISPEECH, ORPAILLEUR (Nancy)

Participants: Irina Illina, Emmanuel Vincent, Georgios Zervakis

Abstract: HyAIAI is about the design of novel, interpretable artificial intelligence methods based on hybrid approaches that combine state of the art numeric models with explainable symbolic models.

ANR HAIKUS

Title: Artificial Intelligence applied to augmented acoustic Scenes

Duration: Dec 2019 - Nov 2023

Coordinator: Ircam (Paris)

Partners: Ircam (Paris), Inria (Nancy), IJLRA (Paris)

Participants: Antoine Deleforge, Prerak Srivastava, Emmanuel Vincent

Abstract: HAIKUS aims to achieve seamless integration of computer-generated immersive audio content into augmented reality (AR) systems. One of the main challenges is the rendering of virtual auditory objects in the presence of source movements, listener movements and/or changing acoustic conditions.

ANR JCJC DENISE

Title: Tackling hard problems in audio using Data-Efficient Non-linear InverSe mEthods

Duration: Oct 2020 – Sep 2024

Coordinator: Antoine Deleforge

Participants: Antoine Deleforge, Tom Sprunck, Marina Krémé

Collaborators: UMR AE, Institut de Recherche Mathématiques Avancées de Strasbourg, Institut de Mathématiques de Bordeaux

Abstract: DENISE aims to explore the applicability of recent breakthroughs in the field of nonlinear inverse problems to audio signal reparation and to room acoustics, and to combine them with compact machine learning models to yield data-efficient techniques.

Action Exploratoire Inria Acoust.IA

Title: Acoust.IA: *l'Intelligence Artificielle au Service de l'Acoustique du Bâtiment*

Duration: Oct 2020 - Sep 2023

Coordinator: Antoine Deleforge

Collaborators: UMR AE (Cerema Est, Strasbourg).

Participants: Antoine Deleforge, Stéphane Dilungana, and Cédric Foy (CEREMA)

Abstract: This project aims at radically simplifying and improving the acoustic diagnosis of rooms and buildings using new techniques combining machine learning, signal processing and physics-based modeling.

InriaHub ADT PEGASUS

Title: PEGASUS: *rehaussement de la ParolE Généralisé par Apprentissage SUperviSé*

Duration: Nov 2020 - Oct 2022

Coordinator: Antoine Deleforge

Participants: Joris Cosentino, Antoine Deleforge, Manuel Pariente, Emmanuel Vincent

Abstract: This engineering project aims at further developing, expanding and transfering the Asteroid speech enhancement and separation toolkit recently released by the team [80].

ANR Full3DTalkingHead

Title: Synthèse articulatoire phonétique

Duration: Apr 2021 - Sep 2024

Coordinator: Yves Laprie

Partners: Gipsa-Lab (Grenoble), IADI (Nancy), LPP (Paris)

Participants: Slim Ouni, Vinicius Ribeiro, Yves Laprie

Abstract: The objective is to realize a complete three-dimensional digital talking head including the vocal tract from the vocal folds to the lips, the face and integrating the digital simulation of the aero-acoustic phenomena.

PEPR Cybersécurité, projet iPOP

Title: Protection des données personnelles

Duration: Oct 2022 – Sep 2028

Coordinator: Vincent Roca (Inria PRIVATICS)

Partners: Inria PRIVATICS (Lyon), COMETE, PETRUS (Saclay), MAGNET, SPIRALS (Lille), IRISA (Rennes), LIFO (Bourges), DCS (Nantes), CESICE (Grenoble), EDHEC (Lille), CNIL (Paris)

Participant: Emmanuel Vincent

Summary: The objectives of iPOP are to study the threats on privacy introduced by new digital technologies, and to design privacy-preserving solutions compatible with French and European regulations. Within this scope, Multispeech focuses on speech data.

ANR DFG M-PHASIS

Title: Migration and Patterns of Hate Speech in Social Media - A Cross-cultural Perspective

Duration: March 2018 – August 2022

Coordinator: Angeliki Monnier and Christian Schemer

Partners: CREM Université de Lorraine, LORIA, JGUM Johannes Gutenberg-Universität Mainz, SAAR Saarland University Saarbrücken

Participant: Irina Illina, Dominique Fohr, Ashwin Geet D'sa

Summary: Focusing on the social dimension of hate speech, the project M-PHASIS seeks to study the patterns of hate speech related to migrants in user-generated content.

ANR REFINED

Title: Real-Time Artificial Intelligence for Hearing Aids

Duration: Mar 2022 - Mar 2026

Coordinator: CEA List (Saclay)

Partners: CEA List (Saclay), Institut de l'audition (Paris), LORIA (Nancy)

Participants: Paul Magron, Nasser-Eddine Monir, Romain Serizel

Abstract: The Refined project brings together audiologists, computer scientists and specialists about hardware implementation to design new speech enhancement algorithms that fit both the needs of patients suffering of hearing losses and the computational constraints of hearing aid devices.

10.4 Regional initiatives

PIA2 ISITE LUE

Title: *Lorraine Université d'Excellence*

Duration: Sep 2019 - Jan 2023

Coordinator: Université de Lorraine

Participants: Tulika Bose, Dominique Fohr, Irina Illina

Abstract: LUE (*Lorraine Université d'Excellence*) was designed as an "engine" for the development of excellence, by stimulating an original dialogue between knowledge fields. The IMPACT initiative OLKI (Open Language and Knowledge for Citizens) funds the PhD thesis of Tulika Bose on the detection and classification of hate speech.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- Co-chair, Dagstuhl Seminar on Privacy in Speech and Language Technology, Wadern, Aug 2022 (E. Vincent)
- Co-chair, 2nd Inria-DFKI European Summer School on Artificial Intelligence, Saarbrücken, Aug 2022 (E. Vincent)
- Co-chair, 3rd Inria-DFKI Workshop on Artificial Intelligence, Bordeaux, Oct 2022 (E. Vincent)
- Co-chair, 7th Workshop on Detection and Classification of Acoustic Scenes and Events, Nancy, Nov 2022 (R. Serizel)

Member of the organizing committees

- Organizer, 2nd VoicePrivacy Challenge (E. Vincent)
- Organizer, 8th Challenge on Detection and Classification of Acoustic Scenes and Events (R. Serizel)

11.1.2 Scientific events: selection

Chair of conference program committees

- Area chair, 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (A. Deleforge, R. Serizel, E. Vincent)

Member of the conference program committees

- Member of program committee, 24th International Conference on Speech and Computer (SPECOM 2022) (D. Jouvet)
- Member of program committee, 25th International Conference on Text, Speech and Dialogue (TSD 2022) (D. Jouvet)
- Member of program committee, 34e Journées d'Études sur la Parole (JEP2022) (V. Colotte, Y. Laprie, S. Ouni)

Reviewer

- AACL-IJCNLP 2022 - 2nd Conference of the Asia-Pacific Chapter of the ACL and 12th International Joint Conference on Natural Language Processing (I. Illina)
- COLING 2022 - 29th International Conference on Computational Linguistics (I. Illina)
- EUSIPCO 2022 - European Signal Processing Conference (D. Jouvet, R. Serizel, A. Deleforge)
- GRETSI 2022 – XXVIIIème Colloque Francophone de Traitement du Signal et des Images (E. Vincent, A. Deleforge)
- ICASSP 2022 – IEEE International Conference on Acoustics, Speech, and Signal Processing (A. Bonneau, D. Jouvet, P. Magron, I. Illina, R. Serizel, A. Deleforge)
- ICML 2022 – 39th International Conference on Machine Learning (A. Deleforge)
- INTERSPEECH 2022 (A. Bonneau, D. Jouvet, P. Magron, E. Vincent)

- JEP 2022 – Journées d'Etudes sur la Parole (A. Bonneau, V. Colotte, I. Illina, D. Jouvet, Y. Laprie, P. Magron, S. Ouni)
- IWAENC 2022 – 17th International Workshop on Acoustic Signal Enhancement (A. Deleforge)
- LREC 2022 – 13th Conference on Language Resources and Evaluation (E. Vincent, I. Illina)
- NeurIPS 2022 – 36th Conference on Neural Information Processing Systems (A. Deleforge)
- SLT 2022 – IEEE Spoken Language Technology Workshop (D. Jouvet, E. Vincent, I. Illina, R. Serizel)
- SP 2022 – 11th international conference on Speech Prosody (A.Bonneau, D. Jouvet)
- SPECOM 2022 - International Conference on Speech and Computer (D. Jouvet)
- SPSC 2022 – 2nd Symposium on Security and Privacy in Speech Communication (P. Champion, D. Jouvet, E. Vincent)
- TSD 2022 - International Conference on Text, Speech and Dialogue (D. Jouvet)

11.1.3 Journal

Member of the editorial boards

- Guest Editor of Neural Networks, special issue on Advances in Deep Learning Based Speech Processing [22] (E. Vincent)
- Speech Communication (D. Jouvet)
- Associate Editor of IEEE/ACM Transactions on Audio, Speech and Language Processing (R. Serizel)
- Associate Editor of IEEE Open Journal of Signal Processing (R. Serizel)
- Associate Editor of EURASIP Journal on Audio, Speech, and Music Processing (A. Deleforge, Y. Laprie)

Reviewer - reviewing activities

- IEEE Transactions on Audio, Speech, and Language Processing (A. Deleforge, P. Magron, V. Colotte, R. Serizel)
- IEEE Signal Processing Letters (P. Magron)
- IEEE Access (P. Magron)
- EURASIP Journal on Audio, Speech, and Music Processing (P. Magron)
- Revue TAL, Special issue "inter-/multi-modal" (S. Ouni)
- Speech Communication (Y. Laprie)
- Journal of the Acoustical Society of America (Y. Laprie)

11.1.4 Invited talks

- Expert session “Speech anonymization”, 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (E. Vincent)
- Panel session “Beyond Words: Recognition, Spoofing, and Anonymization of Individual Traits in Speech”, 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (E. Vincent)
- Keynote “Speech anonymization”, 2022 Annual Conference of the UKRI CDT in Speech and Language Technologies and their Applications (E. Vincent)
- Seminar “Phase recovery for audio demixing: contributions and perspectives”, Neural DSP Technologies (Helsinki, Finland), Aug 2022 (P. Magron)
- Seminar "Prédiction de la forme géométrique du conduit vocal à partir de la suite de phonèmes à articuler, GDR ISIS seminar "Voix", Oct 2022 (Y. Laprie)
- Seminar "Hearing the Shape of a Room: When Signal Processing Meets Acoustics" Institut de Recherche Mathématique Avancée, Univ. Strasbourg, Oct 2022 (A. Deleforge)

11.1.5 Leadership within the scientific community

- Member of the Steering Committee of ISCA’s Special Interest Group on Security and Privacy in Speech Communication (E. Vincent)
- Member of the Steering Committee for the CHiME Challenge series (E. Vincent)
- Secretary/Treasurer, executive member of AVISA (Auditory-VIsual Speech Association), an ISCA Special Interest Group (S. Ouni)
- Vice-president of AFCP - Association Francophone de la Communication Parlée (S. Ouni)
- Board member of AFCP - Association Francophone de la Communication Parlée (V. Colotte, Y. Laprie)
- Member of the Steering Group on Detection and Classification of Acoustic Scenes and Events (R. Serizel)
- Chair of the "Datasets and Challenges" subcommittee of the IEEE Technical Committee for Audio and Acoustics Signal Processing (A. Deleforge)

11.1.6 Scientific expertise

- Member of the Advisory Board of H2020 FVLLMONTI (E. Vincent)
- Coordinator of a task-force on large language models for the French National AI Research Plan (E. Vincent)
- Member of ANR Evaluation Committee 23 on Artificial Intelligence (D. Jouvet)
- Reviewer of an ANR CIFRE proposal (D. Jouvet)
- Reviewer of a project for the Czech Science Foundation (D. Jouvet, I. Illina)
- Reviewer of ANR generic call proposals (A. Deleforge, Y. Laprie, S. Ouni)
- Reviewer of a project for the Swiss National Science Foundation (S. Ouni)
- Reviewer of a project for the Canadian program MITACS (P. Magron)

11.1.7 Research administration

- Head of Science of Inria Nancy – Grand Est (E. Vincent)
- Scientific Director for the partnership between Inria and DFKI (E. Vincent, until Aug 2022)
- Member of Inria's Evaluation Committee (E. Vincent)
- Member of the Comité Espace Transfert of Inria Nancy – Grand Est (E. Vincent)
- Member of the hiring committee for Inria Senior Research Scientists (E. Vincent)
- Member of the hiring committee for Junior Research Scientists, Inria Lille – Nord Europe (E. Vincent)
- Member of the hiring committee for a Junior Professor Chair, Inria Nancy – Grand Est (E. Vincent)
- Member of the admission committee for Inria Starting Faculty Positions, Inria Nancy – Grand Est (E. Vincent)
- Member of the working group "Research visibility and attractiveness" of Métropole du Grand Nancy (E. Vincent)
- Member of the working group "Stimulate the production and use of health data" at City Healthcare 2022 in Nancy (E. Vincent)
- Member of pole scientifique automatique, mathématiques, informatique et leurs interactions (AM2I) of Université de Lorraine (I. Illina)
- Member of the CNU 27 (Conseil National des Universités) - Computer Science (S. Ouni)
- Member of the Commission du personnel scientifique (COMIPERS), Inria Nancy – Grand Est (R. Serizel)
- Member of the Commission de développement technologique (CDT), Inria Nancy – Grand Est (R. Serizel)
- Head of pole scientifique automatique, mathématiques, informatique et leurs interactions (AM2I) of Université de Lorraine (Y. Laprie)
- Member of Commission paritaire of Université de Lorraine (Y. Laprie)
- Member of the board, Université de Lorraine (Y. Laprie)
- President of the Commission pour l'Action et la Responsabilité Ecologique (CARE) Inria/Loria (A. Deleforge).
- Member of the Information & Edition Scientifique (IES) commision of INRIA (A. Deleforge, P. Magron)
- Participation in the hiring committee for the professor of general phonetics at Université de la Sorbonne Nouvelle (Y. Laprie)

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- BUT: I. Illina, Java programming (100 hours), Linux programming (58 hours), and Advanced Java programming (40 hours), L1, Université de Lorraine, France
- BUT: I. Illina, Supervision of student projects and internships (50 hours), L2, Université de Lorraine, France
- BUT: R. Serizel, Introduction to office tools (108 hours), Multimedia and web (20 hours), Documents and databases (20 hours), L1, Université de Lorraine, France

- BUT: R. Serizel, Multimedia content and indexing (14 hours), Content indexing and retrieval software (20 hours), L2, Université de Lorraine, France
- BUT: S. Ouni, Programming in Java (24 hours), Web Programming (24 hours), Graphical User Interface (96 hours), L1, Université de Lorraine, France
- BUT: S. Ouni, Advanced Algorihms (24 hours), L2, Université de Lorraine, France
- Licence: A. Bonneau, Phonetics (16 hours), L2, *École d'audioprothèse*, Université de Lorraine, France
- Licence: V. Colotte, Digital literacy and tools (hybrid courses, 50 hours), L1, Université de Lorraine, France
- Licence: V. Colotte, System (80 hours), L2-L3, Université de Lorraine, France
- Master: V. Colotte, Integration project: multimodal interaction with Pepper Robot (17 hours), M2, Université de Lorraine, France
- Master: V. Colotte and S. Ouni, Multimodal oral comunication (24 hours), M2, Université de Lorraine, France
- Master: V. Colotte, Introduction to speech processing (24 hours), M1, Université de Lorraine, France
- Master: Y. Laprie, Speech corpora (30 hours), M1, Université de Lorraine, France
- Master: S. Ouni, Multimedia in Distributed Information Systems (31 hours), M2, Université de Lorraine, France
- Master: R. Serizel, S. Ouni, P. Magron and V. Ribeiro, Oral speech processing (24 hours), M2, Université de Lorraine
- Master: E. Vincent and P. Magron, Neural networks (40 hours), M2, Université de Lorraine
- Master: A. Deleforge, Initiation to Machine Learning (24 hours), M1-M2, Télécom Physique Strasbourg
- Other: V. Colotte, Co-Responsible for NUMOC (Digital literacy by hybrid courses) for Université de Lorraine, France (for 7,000 students)
- Other: S. Ouni, Responsible of *Année Spéciale* DUT, Université de Lorraine, France

11.2.2 Supervision

- PhD: Adrien Dufraux, "Exploitation of noisy transcriptions for automatic speech recognition", Apr 2022, E. Vincent, A. Brun (LORIA) and M. Douze (Meta AI) [64].
- PhD: Ashwin Geet D'Sa, "Expanding the training data for neural network based hate speech classification", May 2022, I. Illina, D. Fohr [65]
- PhD: Ajinkya Kulkarni, "Expressivity transfer in deep learning based text-to-speech synthesis", Jul 2022, V. Colotte and D. Jouvet [66].
- PhD: Mauricio Michel Olvera Zambrano, "Robust audio event detection", Dec 2022, E. Vincent and G. Gasso (LITIS).
- PhD in progress: Tulika Bose, "Trandfert learning for abusive language detection", Sept 2019, I. Illina, D. Fohr
- PhD in progress: Pierre Champion, "Privacy preserving and personalized transformations for speech recognition", Oct 2019, D. Jouvet and A. Larcher (LIUM).

- PhD in progress: Sandipana Dowerah, “Robust speaker verification from far-field speech”, Oct 2019, D. Jouvet and R. Serizel.
- PhD in progress: Shakeel Ahmad Sheikh, “Identifying disfluency in speakers with stuttering, and its rehabilitation, using DNN”, Oct 2019, S. Ouni
- PhD in progress: Georgios Zervakis, “Integration of symbolic knowledge into deep learning”, Nov 2019, M. Couceiro (LORIA) and E. Vincent.
- PhD in progress: Nicolas Zampieri, "classification des discours haineux dans les réseaux sociaux", Nov 2019, I. Illina, D. Fohr
- PhD in progress: Prerak Srivastava, “Hearing the walls of a room: machine learning for audio augmented reality”, Oct 2020, A. Deleforge and E. Vincent.
- PhD in progress: Stéphane Dilungana, "ACOUST.IA: artificial intelligence for building acoustics", Oct 2020, A. Deleforge, C. Foy, S. Faisan
- PhD in progress: Seyed Ahmad Hosseini, “3D sign language generation”, Feb 2021, S. Ouni and M. Sadeghi
- PhD in progress: Can Cui, “Joint and embedded automatic speech separation, diarization and recognition for the generation of meeting minutes”, Oct 2021, M. Sadeghi and E. Vincent
- PhD in progress: Sewade Olaolu Ogun, “Multi-factor data augmentation and transfer learning for embedded automatic speech recognition”, Oct 2021, V. Colotte and E. Vincent
- PhD in progress: Louis Abel, “Expressive audio-visual speech synthesis in an interaction context”, Oct 2021, S. Ouni and V. Colotte
- PhD in progress: Mickaëlla Grondin, “Modeling gestures and speech in interactions”, Nov 2021, S. Ouni and F. Hirsch (Praxiling)
- PhD in progress: Tom Sprunck, "Hearing the Shape of a Room: Towards Acoustic Super-Resolution", Nov 2021, A. Deleforge, C. Foy, Y. Privat
- PhD in progress: Robin San Roman, "Self supervised disentangled representation learning of audio data for compression and generation", Jun 2022, R. Serizel, A. Deleforge, Y. M. Adi, G. Synnaeve
- PhD in progress: Nasser-Eddine Monir, “Multichannel speech enhancement for patients with auditory neuropathy spectrum disorders”, Dec 2022, R. Serizel and P. Magron

11.2.3 Juries

- Participation in the HDR jury of Antoine Laurent (Le Mans Université, Jan 2022), D. Jouvet, reviewer
- Participation in the HDR jury of Jean-Luc Rouas (Université de Bordeaux, Mar 2022), E. Vincent, member
- Participation in the PhD jury of Adrien Llave (Centrale Supélec, Mar 2022), R. Serizel, member.
- Participation in the PhD jury of Nicolas Olivier (Univ. de Rennes, Mar 2022), S. Ouni, reviewer
- Participation in the PhD jury of Marie-Anne Lacroix (Université de Rennes 1, Apr 2022), R. Serizel, reviewer
- Participation in the PhD jury of Jiri Martinek (University of West Bohemia, Apr 2022), D. Jouvet, reviewer
- Participation in the PhD jury of Manon Macary (Le Mans Université, Jun 2022), D. Jouvet, reviewer
- Participation in the PhD jury of Yingming Gao (Technische Universität Dresden, Jun 2022), Y. Laprie, reviewer.

- Participation in the PhD jury of Pierre-Hugo Vial (IRIT, Toulouse, Nov 2022), P. Magron, member
- Participation in the PhD jury of Thomas BARGUIL, (Univ de Bretagne sud, Nov 2022), I. Illina, reviewer

11.3 Popularization

11.3.1 Articles and contents

- The race to hide your voice, Wired, June 1, 2022 (E. Vincent)
- Interview for "Dynalips, solution de lipsync de pointe, anime MetaHuman !", Loria Actualité, and Facutel (université de Lorraine), July 2022 (S. Ouni)
- Les assistants vocaux haussent le ton, 01net, August 24, 2022 (E. Vincent)
- La reconnaissance vocale : vous allez tout comprendre, Monde Numérique, August 22, 2022 (E. Vincent)
- Interview for "Comment le numérique peut aider au diagnostic et à la prise en charge du bégaiement par les orthophonistes ?", Inria Actualité, Dec 2022 (S. Ouni)
- Interview for "REFINED : vers des prothèses auditives plus efficaces grâce à l'IA", Inria Actualité, Dec 2022 (R. Serizel)

11.3.2 Interventions

- Talk on speech anonymization at the Cycle de la Voix #3 seminar organized by Le VoiceLab, Feb 2022 (E. Vincent)
- Talk on speech neural synthesis at the Cycle de la Voix #9 seminar organized by Le VoiceLab, Nov 2022 (V. Colotte)
- Panel discussion at BpiFrance's Deeptech Tour in Metz, Mar 2022 (E. Vincent)
- Chiche, Lycée Saint-Exupéry, Fameck (2 classes), Mar 2022 (E. Vincent)
- Panel discussion at Viva Technology in Paris, Jun 2022 (E. Vincent)
- Chiche, Lycée Kastler, Stenay (3 classe), Mar 2022 (R. Serizel)
- Representations of the theater piece "Drone Control" with "Les sens des mots" at Reine Blanche Theater, Paris, Mar 2022, and Jardin Botanique, Villers-lès-Nancy, Sep 2022 (A. Deleforge)
- Representation of the theater piece "Le Procès du Robot" with "Crache Texte" at Faculté des Sciences et Technologies, Villers-lès-Nancy, Oct 2022 (A. Deleforge)

12 Scientific production

12.1 Major publications

[1] T. Bose, N. Aletras, I. Illina and D. Fohr. 'Dynamically Refined Regularization for Improving Cross-corpora Hate Speech Detection'. In: ACL 2022 - 60th meeting Association for Computational Linguistics Findings. Dublin, Ireland, 22nd May 2022. DOI: 10.18653/v1/2022.findings-acl.32. URL: https://hal.inria.fr/hal-03690174.

[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* 141 (2021), pp. 315–329. DOI: 10.1016/j.neunet.2021.04.021. URL: https://hal.inria.fr/hal-03204193.

[3] N. Furnon, R. Serizel, S. Essid and I. Illina. 'DNN-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 29 (2021), pp. 2310–2323. DOI: 10.1109/TASLP.2021.3092838. URL: https://hal.archives-ouvertes.fr/hal-02985867.

[4] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge and E. Vincent. 'Asteroid: the PyTorch-based audio source separation toolkit for researchers'. In: Interspeech 2020. Shanghai, China, 25th Oct. 2020. URL: https://hal.inria.fr/hal-02962964.

[5] V. Ribeiro, K. Isaieva, J. Leclere, P.-A. Vuissoz and Y. Laprie. 'Automatic generation of the complete vocal tract shape from the sequence of phonemes to be articulated'. In: *Speech Communication* 141 (22nd Apr. 2022), pp. 1–13. DOI: 10.1016/j.specom.2022.04.004. URL: https://hal.univ-lorraine.fr/hal-03650212.

[6] S. Sheikh, M. Sahidullah, F. Hirsch and S. Ouni. 'Machine Learning for Stuttering Identification: Review, Challenges & Future Directions'. In: *Neurocomputing* 514.2022 (12th Oct. 2022), p. 17. DOI: 10.1016/j.neucom.2022.10.015. URL: https://hal.archives-ouvertes.fr/hal-03634072.

[7] 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* (15th June 2022). URL: https://hal.inria.fr/hal-03197376.

12.2 Publications of the year

International journals

[8] S. Amini, M. Soltanian, M. Sadeghi and S. Ghaemmaghami. 'Non-Smooth Regularization: Improvement to Learning Framework through Extrapolation'. In: *IEEE Transactions on Signal Processing* 70 (2022), pp. 1213–1223. DOI: 10.1109/TSP.2022.3154969. URL: https://hal.inria.fr/hal-03586153.

[9] S. Dey, M. Sahidullah and G. Saha. 'Cross-corpora spoken language identification with domain diversification and generalization'. In: *Computer Speech and Language* (13th Feb. 2023). URL: https://hal.inria.fr/hal-03984643.

[10] I. K. Douros, Y. Xie, C. Dourou, K. Isaieva, P.-A. Vussoz, J. Felblinger and Y. Laprie. '3D dynamic spatiotemporal atlas of the vocal tract during consonant-vowel production from 2D real time MRI'. In: *Journal of Imaging*. Special Issue Spatio-Temporal Biomedical Image Analysis 8.9 (25th Aug. 2022), p. 227. DOI: 10.3390/jimaging8090227. URL: https://hal.inria.fr/hal-03808325.

[11] Z. Kang, M. Sadeghi, R. Horaud and X. Alameda-Pineda. 'Expression-preserving face frontalization improves visually assisted speech processing'. In: *International Journal of Computer Vision* (12th Jan. 2023). URL: https://hal.science/hal-03902610.

[12] P. Magron and C. Févotte. 'A majorization-minimization algorithm for nonnegative binary matrix factorization'. In: *IEEE Signal Processing Letters* (June 2022). URL: https://hal.inria.fr/hal-03647772.

[13] P. Magron and C. Févotte. 'Neural content-aware collaborative filtering for cold-start music recommendation'. In: *Data Mining and Knowledge Discovery* (2022). URL: https://hal.inria.fr/hal-03152158.

[14] O. Mokrý, P. Magron, T. Oberlin and C. Févotte. 'Algorithms for audio inpainting based on probabilistic nonnegative matrix factorization'. In: *Signal Processing* (1st May 2023). URL: https://hal.inria.fr/hal-03708613.

[15] V. Ribeiro, K. Isaieva, J. Leclere, P.-A. Vuissoz and Y. Laprie. 'Automatic generation of the complete vocal tract shape from the sequence of phonemes to be articulated'. In: *Speech Communication* 141 (22nd Apr. 2022), pp. 1–13. DOI: 10.1016/j.specom.2022.04.004. URL: https://hal.univ-lorraine.fr/hal-03650212.

[16] A. S. Shamsabadi, B. M. L. Srivastava, A. Bellet, N. Vauquier, E. Vincent, M. Maouche, M. Tommasi and N. Papernot. 'Differentially private speaker anonymization'. In: *Proceedings on Privacy Enhancing Technologies* 2023.1 (1st Jan. 2023). URL: https://hal.inria.fr/hal-03588932.

[17] S. Sheikh, M. Sahidullah, F. Hirsch and S. Ouni. 'Machine Learning for Stuttering Identification: Review, Challenges & Future Directions'. In: *Neurocomputing* 514.2022 (12th Oct. 2022), p. 17. DOI: 10.1016/j.neucom.2022.10.015. URL: https://hal.science/hal-03634072.

[18] T. Sprunck, A. Deleforge, Y. Privat and C. Foy. 'Gridless 3D Recovery of Image Sources from Room Impulse Responses'. In: *IEEE Signal Processing Letters* (25th Nov. 2022). DOI: 10.1109/LSP.2022.3224682. URL: https://hal.inria.fr/hal-03763838.

[19] 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* (15th June 2022). URL: https://hal.inria.fr/hal-03197376.

[20] 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'. In: *Computer Speech and Language* 74 (July 2022), p. 101362. DOI: 10.1016/j.csl.2022.101362. URL: https://hal.science/hal-03332224.

[21] P.-H. Vial, P. Magron, T. Oberlin and C. Fevotte. 'Learning the Proximity Operator in Unfolded ADMM for Phase Retrieval'. In: *IEEE Signal Processing Letters* 29 (2022), pp. 1619–1623. DOI: 10.1109/LSP.2022.3189275. URL: https://hal.science/hal-03790178.

[22] X. Zhang, L. Xie, E. Fosler-Lussier and E. Vincent. 'Guest editorial: Special issue on advances in deep learning based speech processing'. In: *Neural Networks* 158 (1st Jan. 2023). DOI: 10.1016/j.neunet.2022.11.033. URL: https://hal.inria.fr/hal-03883292.

International peer-reviewed conferences

[23] M. Abdollahi, R. Serizel, A. Rakotomamonjy and G. Gasso. 'Integrating isolated examples with weakly-supervised sound event detection: a direct approach'. In: 7th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE). Nancy, France, 3rd Nov. 2022. URL: https://hal.science/hal-03894965.

[24] T. Bose, N. Aletras, I. Illina and D. Fohr. 'Domain Classification-based Source-specific Term Penalization for Domain Adaptation in Hate-speech Detection'. In: COLING 2022 - Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, South Korea, 12th Oct. 2022. URL: https://hal.inria.fr/hal-03815708.

[25] T. Bose, N. Aletras, I. Illina and D. Fohr. 'Dynamically Refined Regularization for Improving Cross-corpora Hate Speech Detection'. In: ACL 2022 - 60th meeting Association for Computational Linguistics Findings. Dublin, Ireland, 22nd May 2022. DOI: 10.18653/v1/2022.findings-acl.32. URL: https://hal.inria.fr/hal-03690174.

[26] T. Bose, I. Illina and D. Fohr. 'Transferring Knowledge via Neighborhood-Aware Optimal Transport for Low-Resource Hate Speech Detection'. In: Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP). Online, Taiwan, 21st Nov. 2022. URL: https://hal.inria.fr/hal-03846693.

[27] P. Champion, D. Jouvet and A. Larcher. 'Are disentangled representations all you need to build speaker anonymization systems?' In: *Proc. Interspeech 2022*. INTERSPEECH 2022 - Human and Humanizing Speech Technology. incheon, South Korea, 18th Sept. 2022. URL: https://hal.science/hal-03753746.

[28] P. Champion, D. Jouvet and A. Larcher. 'Privacy-Preserving Speech Representation Learning using Vector Quantization'. In: JEP 2022 - Journées d'Études sur la Parole. Journées d'Études sur la Parole. Île de Noirmoutier, France, 13th June 2022. URL: https://hal.science/hal-03609205.

[29] S. Dilungana, A. Deleforge, C. Foy and S. Faisan. 'Geometry-Informed Estimation of Surface Absorption Profiles from Room Impulse Responses'. In: 30th European Signal Processing Conference (EUSIPCO). Belgrade, Serbia: IEEE, 29th Aug. 2022, pp. 867–871. DOI: 10.23919/EUSIPCO55093.2022.9909667. URL: https://hal.science/hal-03636502.

[30] S. Dowerah, R. Serizel, D. Jouvet, M. Mohammadamini and D. Matrouf. 'Joint optimization of diffusion probabilistic-based multichannel speech enhancement with far-field speaker verification'. In: IEEE SLT 2022. Doha, Qatar, 9th Jan. 2023. URL: https://hal.science/hal-03671583.

[31] S. Dowerah, R. Serizel, D. Jouvet, M. Mohammadamini and D. Matrouf. 'How to Leverage DNN-based speech enhancement for multi-channel speaker verification?' In: 4th International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI' 2022). Corfu, Greece, 19th Oct. 2022. URL: https://hal.science/hal-03619903.

[32] J. Ebbers, R. Haeb-Umbach and R. Serizel. 'Threshold independent evaluation of sound event detection scores'. In: ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, Singapore, 22nd May 2022. DOI: 10.1109/ICASSP43922.2022.9747556. URL: https://hal.inria.fr/hal-03562763.

[33] A. Geet d'Sa, I. Illina, D. Fohr and A. Akbar. 'Exploration of Multi-Corpus Learning for Hate Speech Classification in Low Resource Scenarios'. In: TSD 2022 - 25th International Conference on Text, Speech and Dialogue. proceedings of TSD 2022. Brno, Czech Republic, 6th Sept. 2022. URL: https://hal.science/hal-03712918.

[34] I. Illina and D. Fohr. 'BERT Semantic Context Model for Efficient Speech Recognition'. In: ICCAS 2022 - International Conference on Cognitive Aircraft Systems. proceedings of ICCAS 2022. Toulouse, France, 1st June 2022. URL: https://hal.science/hal-03712987.

[35] I. Illina and D. Fohr. 'Semantic Information Investigation for Transformer-based Rescoring of N-best Speech Recognition'. In: LTC 2023. Proceedings of Language and Technology 2023. Poznan, Poland, 21st Apr. 2023. URL: https://hal.science/hal-03965397.

[36] Z. Kang, M. Sadeghi, R. Horaud, X. Alameda-Pineda, J. Donley and A. Kumar. 'The Impact of Removing Head Movements on Audio-visual Speech Enhancement'. In: ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, Singapore: IEEE, 1st Feb. 2022, pp. 1–5. DOI: 10.1109/ICASSP43922.2022.9746401. URL: https://hal.inria.fr/hal-03551610.

[37] R. Karpinski, V. Ribeiro and Y. Laprie. 'Accelerating the Centerline Processing of Vocal Tract Shapes for Articulatory Synthesis'. In: ICA 2022- 24th International Congress on Acoustics. Gyeongyu, South Korea, 22nd Oct. 2022. URL: https://hal.inria.fr/hal-03798827.

[38] A. M. Kreme, B. Torrésani and A. Deleforge. 'Local time-frequency fading'. In: *Proceedings of the 24th International Congress on Acoustics*. ICA 22 - International Congress on Acoustics 2022. Gyeongju, South Korea, 2022. URL: https://hal.science/hal-03923596.

[39] A. Kulkarni, V. Colotte and D. Jouvet. 'Analysis of expressivity transfer in non-autoregressive end-to-end multispeaker TTS systems'. In: INTERSPEECH 2022. Incheon, South Korea, 18th Sept. 2022. URL: https://hal.inria.fr/hal-03832870.

[40] A. Kulkarni, V. Colotte and D. Jouvet. 'Multi-stage attention for fine-grained expressivity transfer in multispeaker text-to-speech system'. In: EUSIPCO 2022. Belgrade, Serbia, 29th Aug. 2022. URL: https://hal.science/hal-03615773.

[41] M.-A. Lacroix, N. Bertin, R. Rocher and P. Scalart. 'Vers un système embarqué de classification d'événements sonores : étude de l'impact de la quantification des descripteurs'. In: GRETSI 2022 XXVIIIème Colloque Francophone de Traitement du Signal et des Images. GRETSI 2022 XXVIIIème Colloque Francophone de Traitement du Signal et des Images. Nancy, France, 6th Sept. 2022. URL: https://hal.science/hal-03741340.

[42] M. Maouche, B. M. L. Srivastava, N. Vauquier, A. Bellet, M. Tommasi and E. Vincent. 'Enhancing speech privacy with slicing'. In: Interspeech 2022 - Human and Humanizing Speech Technology. Incheon, South Korea, 18th Sept. 2022. URL: https://hal.inria.fr/hal-03369137.

[43] M. Mohammadamini, D. Matrouf, J.-F. Bonastre, S. Dowerah, R. Serizel and D. Jouvet. 'A Comprehensive Exploration of Noise Robustness and Noise Compensation in ResNet and TDNN-based Speaker Recognition Systems'. In: EUSIPCO 2022 - 30th European Signal Processing Conference. Belgrade, Serbia, 29th Aug. 2022. URL: https://hal.science/hal-03669919.

[44] M. Mohammadamini, D. Matrouf, J.-F. A. Bonastre, S. Dowerah, R. Serizel and D. Jouvet. 'Barlow Twins self-supervised learning for robust speaker recognition'. In: Interspeech 2022 - Human and Humanizing Speech Technology. Incheon, South Korea, 18th Sept. 2022. DOI: 10.21437/Interspeech.2022-11301. URL: https://hal.science/hal-03710445.

[45] H. Nourtel, P. Champion, D. Jouvet, A. Larcher and M. Tahon. 'Evaluation of Speaker Anonymization on Emotional Speech'. In: JEP 2022 - Journées d'Études sur la Parole. Actes des 34e Journées d'Études sur la Parole — JEP2022 « Parole, Geste, Musique : des unités à leur organisation ». Île de Noirmoutier, France, 13th June 2022. URL: https://hal.science/hal-03636737.

[46] S. Ogun, V. Colotte and E. Vincent. 'Can we use Common Voice to train a Multi-Speaker TTS system?' In: The 2022 IEEE Spoken Language Technology Workshop (SLT 2022). Doha, Qatar, 9th Jan. 2023. URL: https://hal.science/hal-03812715.

[47] C. Oguz, I. Kruijff-Korbayová, P. Denis, E. Vincent and J. van Genabith. 'Chop and change: Anaphora resolution in instructional cooking videos'. In: Findings of AACL-IJCNLP 2022 - 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics - 12th International Joint Conference on Natural Language Processing. Taipeh, Taiwan, 20th Nov. 2022. URL: https://hal.inria.fr/hal-03807530.

[48] M. Olvera, E. Vincent and G. Gasso. 'On the impact of normalization strategies in unsupervised adversarial domain adaptation for acoustic scene classification'. In: ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, Singapore, 22nd May 2022. DOI: 10.1109/ICASSP43922.2022.9747540. URL: https://hal.inria.fr/hal-03668251.

[49] V. Ribeiro and Y. Laprie. 'Autoencoder-Based Tongue Shape Estimation During Continuous Speech'. In: 23rd INTERSPEECH Conference on "Human and Humanizing Speech Technology". Incheon, South Korea, 18th Sept. 2022. URL: https://hal.inria.fr/hal-03798790.

[50] F. Ronchini and R. Serizel. 'A benchmark of state-of-the-art sound event detection systems evaluated on synthetic soundscapes'. In: ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore/Virtual, Singapore, 22nd May 2022. DOI: 10.1109/ICASSP43922.2022.9747577. URL: https://hal.inria.fr/hal-03554305.

[51] D. Ruiter, L. Reiners, A. Geet d'Sa, T. Kleinbauer, D. Fohr, I. Illina, D. Klakow, C. Schemer and A. Monnier. 'Placing M-Phasis on the Plurality of Hate: A Feature-Based Corpus of Hate Online'. In: LREC 2022 – 13th Language Resources and Evaluation Conference. LREC 2022 Proceedings. Marseille, France, 19th June 2022, pp. 791–804. URL: https://hal.science/hal-03712978.

[52] M. Sadeghi and P. Magron. 'A Sparsity-promoting Dictionary Model for Variational Autoencoders'. In: INTERSPEECH 2022. Incheon, South Korea, 18th Sept. 2022. URL: https://hal.inria.fr/hal-03623769.

[53] I. A. Sheikh, E. Vincent and I. Illina. 'Transformer versus LSTM Language Models Trained on Uncertain ASR Hypotheses in Limited Data Scenarios'. In: LREC 2022 - 13th Language Resources and Evaluation Conference. Marseille, France, 20th June 2022. URL: https://hal.inria.fr/hal-03362828.

[54] S. Sheikh, M. Sahidullah, F. Hirsch and S. Ouni. 'Robust Stuttering Detection via Multi-task and Adversarial Learning'. In: EUSIPCO 2022 - 30th European Signal Processing Conference. Belgrade, Serbia, 29th Aug. 2022. URL: https://hal.inria.fr/hal-03629785.

[55] S. A. Sheikh, M. Sahidullah, F. Hirsch and S. Ouni. 'End-to-End and Self-Supervised Learning for ComParE 2022 Stuttering Sub-Challenge'. In: ACM Multimedia 2022 Computational Paralinguistics Challenge (ComParE). Lisbon, Portugal, 14th Oct. 2022. URL: https://hal.inria.fr/hal-03728331.

[56] P. Srivastava, A. Deleforge and E. Vincent. 'Realistic sources, receivers and walls improve the generalisability of virtually-supervised blind acoustic parameter estimators'. In: 17th International Workshop on Acoustic Signal Enhancement (IWAENC). 17th International Workshop on Acoustic Signal Enhancement (IWAENC). Bamberg, Germany, 5th Sept. 2022. URL: https://hal.science/hal-03727423.

[57] M. A. T. Turan, D. Klakow, E. Vincent and D. Jouvet. 'Adapting Language Models When Training on Privacy-Transformed Data'. In: LREC 2022 - 13th Language Resources and Evaluation Conference. Marseille, France, 20th June 2022. URL: https://hal.inria.fr/hal-03189354.

[58] N. Zampieri, C. Ramisch, I. Illina and D. Fohr. 'Identification des Expressions Polylexicales dans les Tweets'. In: RECITAL 2022- Traitement Automatique des Langues Naturelles (TALN). actes de TALN-RECITAL 2022. Avignon, France, 27th June 2022. URL: https://hal.science/hal-03676506.

[59] N. Zampieri, C. Ramisch, I. Illina and D. Fohr. 'Identification of Multiword Expressions in Tweets for Hate Speech Detection'. In: LREC 2022 - 13th Edition of its Language Resources and Evaluation Conference. LREC 2022 Proceedings. Marseille, France, 20th June 2022. URL: https://hal.science/hal-03676508.

[60] G. Zervakis, E. Vincent, M. Couceiro, M. Schoenauer and E. Marquer. 'An analogy based approach for solving target sense verification'. In: NLPIR 2022 - 6th International Conference on Natural Language Processing and Information Retrieval. Bangkok, Thailand, 16th Dec. 2022. URL: https://hal.inria.fr/hal-03792071.

Conferences without proceedings

[61] D. Caillat, L. Marin, C. Dodane, F. Hirsch, S. Ouni, P. Slangen, P. Guyot, V. Colotte, A. Morgenstern, L. Abel, M. Grondin-Verdon and J. L. Goupil. 'Synchronization of speech and gestures in an interactional context (SyncoGest Project)'. In: ISGS 2022 - 9th Conference of the International Society for Gesture Studies. Chicago, United States, 2022. URL: https://hal.mines-ales.fr/hal-03875218.

[62] S. Deckert, A. Piquard-Kipffer and A. Bonneau. 'Perception of German fricatives by French dyslexic subjects'. In: *New Sounds 2022 Book of abstracts*. New Sounds 2022, 10th International Symposium on the Acquisition of Second Language Speech. Barcelone, Spain, 20th Apr. 2022. URL: https://hal.inria.fr/hal-03852125.

Edition (books, proceedings, special issue of a journal)

[63] M. Lagrange, A. Mesaros, T. Pellegrini, G. Richard, R. Serizel and D. Stowell, eds. *Proceedings of the 7th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2022)*. Tampere University, Nov. 2022, pp. 1–225. URL: https://hal.inria.fr/hal-03850901.

Doctoral dissertations and habilitation theses

[64] A. Dufraux. 'Leveraging noisy transcriptions for automatic speech recognition'. Université de Lorraine, 14th Apr. 2022. URL: https://hal.univ-lorraine.fr/tel-03669875.

[65] A. Geet d'Sa. 'Expanding the training data for neural network based hate speech classification'. Université de Lorraine, 6th May 2022. URL: https://hal.univ-lorraine.fr/tel-03833821.

[66] A. Kulkarni. 'Expressivity transfer in deep learning based text-to-speech synthesis'. Université de Lorraine, 7th July 2022. URL: https://hal.univ-lorraine.fr/tel-03844914.

[67] R. Serizel. 'Contributions to speech processing and ambient sound analysis'. Université de Lorraine, 16th Mar. 2022. URL: https://hal.inria.fr/tel-03612609.

Reports & preprints

[68] L. Bahrman, M. Krémé, P. Magron and A. Deleforge. *Signal Inpainting from Fourier Magnitudes*. 27th Oct. 2022. URL: https://hal.science/hal-03832480.

[69] L. Delebecque, R. Serizel and N. Furnon. *Towards an efficient computation of masks for multichannel speech enhancement*. 10th Mar. 2022. URL: https://hal.science/hal-03604983.

[70] A. Golmakani, M. Sadeghi, X. Alameda-Pineda and R. Serizel. *Weighted variance variational autoencoder for speech enhancement*. 28th Oct. 2022. URL: https://hal.inria.fr/hal-03833827.

[71] A. Golmakani, M. Sadeghi and R. Serizel. *AUDIO-VISUAL SPEECH ENHANCEMENT WITH A DEEP KALMAN FILTER GENERATIVE MODEL*. 28th Oct. 2022. URL: https://hal.inria.fr/hal-03833814.

[72] F. Gontier, R. Serizel and C. Cerisara. *SPICE+: EVALUATION OF AUTOMATIC AUDIO CAPTIONING SYSTEMS WITH PRE-TRAINED LANGUAGE MODELS*. 2022. URL: https://hal.inria.fr/hal-03933981.

[73] M. Krémé and B. Torrésani. *Étude d'un algorithme d'optimisation pour le fading temps-fréquence*. 22nd June 2022. URL: https://hal.science/hal-03701278.

[74] M. Sadeghi and R. Serizel. *FAST AND EFFICIENT SPEECH ENHANCEMENT WITH VARIATIONAL AUTOENCODERS*. 28th Oct. 2022. URL: https://hal.inria.fr/hal-03833836.

[75] R. Serizel, S. Cornell and N. Turpault. *Performance above all ? energy consumption vs. performance for machine listening, a study on dcase task 4 baseline*. 14th Nov. 2022. URL: https://hal.inria.fr/hal-03850797.

[76] P. Srivastava, A. Deleforge, A. Politis and E. Vincent. *How to (Virtually) Train Your Sound Source Localizer*. 30th Nov. 2022. URL: https://hal.science/hal-03855912.

[77] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi and J.-F. Bonastre. *The VoicePrivacy 2022 Challenge Evaluation Plan*. 26th Sept. 2022. URL: https://hal.science/hal-03623516.

[78] 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. *Supplementary material to the paper The VoicePrivacy 2020 Challenge: Results and findings*. 26th Sept. 2022. URL: https://hal.science/hal-03335126.

12.3 Cited publications

[79] J. Barker, R. Marxer, E. Vincent and S. Watanabe. 'The third 'CHIME' speech separation and recognition challenge: Analysis and outcomes'. In: *Computer Speech and Language* 46 (July 2017), pp. 605–626. DOI: 10.1016/j.csl.2016.10.005. URL: https://hal.inria.fr/hal-01382108.

[80] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martin-Doñas, D. Ditter, A. Frank, A. Deleforge and E. Vincent. 'Asteroid: the PyTorch-based audio source separation toolkit for researchers'. In: *Interspeech 2020*. Fully Virtual Conference. Shanghai, China, Oct. 2020. URL: https://hal.inria.fr/hal-02962964.

[81] F. Ronchini, R. Serizel, N. Turpault and S. Cornell. 'The impact of non-target events in synthetic soundscapes for sound event detection'. In: *DCASE 2021 - Detection and Classification of Acoustic Scenes and Events*. Barcelona/Virtual, Spain, Nov. 2021. URL: https://hal.inria.fr/hal-03355184.