RESEARCH CENTRE

Inria Centre at Université de Lorraine

IN PARTNERSHIP WITH:

CNRS, Université de Lorraine

2023

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 1

1 Team members, visitors, external collaborators 3

2 Overall objectives 5

3 Research program 5
3.1 Axis 1 — Data-efficient and privacy-preserving learning . . . 5
3.1.1 Axis 1.1 — Integrating domain knowledge . . . 5
3.1.2 Axis 1.2 — Learning from little/no labeled data . . . 6
3.1.3 Axis 1.3 — Preserving privacy . . . 6
3.2 Axis 2 — Extracting information from speech signals . . . 6
3.2.1 Axis 2.1 — Linguistic speech content . . . 6
3.2.2 Axis 2.2 — Speaker identity and states . . . 6
3.2.3 Axis 2.3 — Speech environment information . . . 6
3.3 Axis 3 — Multimodal Speech: generation and interaction . . . 6
3.3.1 Axis 3.1 - Multimodality modeling and analysis . . . 7
3.3.2 Axis 3.2 - Multimodal speech generation . . . 7
3.3.3 Axis 3.3 — Interaction . . . 7
3.4 Software platform: Multimodal Voice assistant . . . 7

4 Application domains 7
4.1 Language Learning . . . 7
4.2 Health Assistance . . . 8

5 Social and environmental responsibility 8

6 Highlights of the year 8
6.1 Awards . . . 8

7 New software, platforms, open data 9
7.1 New software . . . 9
7.1.1 ASTALI . . . 9
7.1.2 Voice Transformer 2 . . . 9
7.1.3 Web Multimodal Annotation . . . 9
7.1.4 VAC . . . 10
7.1.5 PercEval . . . 10
7.2 New platforms . . . 11
7.2.1 Virtual Assistant Creator . . . 11
7.3 Open data . . . 11
7.3.1 MRI Dataset . . . 11

8 New results 12
8.1 Axis 1 — Data-efficient and privacy-preserving learning . . . 12
8.1.1 Axis 1.1 — Integrating domain knowledge . . . 12
8.1.2 Axis 1.2 - Learning from little/no labeled data . . . 12
8.1.3 Axis 1.3 - Preserving privacy . . . 13
8.2 Axis 2 — Extracting information from speech signals . . . 13
8.2.1 Axis 2.1 — Linguistic speech content . . . 13
8.2.2 Axis 2.2 — Speaker identity and states . . . 14
8.2.3 Axis 2.3 — Speech in its environment . . . 15
8.3 Axis 3 — Multimodal Speech: generation and interaction . . . 16
8.3.1 Axis 3.1 — Multimodality modeling and analysis . . . 16
8.3.2 Axis 3.2 — Multimodal speech generation . . . 16
8.3.3 Axis 3.3 — Interaction . . . 17

9 Bilateral contracts and grants with industry **18**
9.1 Bilateral grants with industry . . . 18
9.1.1 Vivoka . . . 18
9.1.2 Meta AI . . . 18
9.1.3 Orange Labs . . . 18

10 Partnerships and cooperations **18**
10.1 International initiatives . . . 18
10.1.1 Participation in other International Programs . . . 18
10.2 International research visitors . . . 19
10.2.1 Visits of international scientists . . . 19
10.3 European initiatives . . . 19
10.3.1 Horizon Europe . . . 19
10.3.2 Other european programs/initiatives . . . 21
10.4 National initiatives . . . 22

11 Dissemination **25**
11.1 Promoting scientific activities . . . 25
11.1.1 Scientific events: selection . . . 25
11.1.2 Journal . . . 26
11.1.3 Invited talks . . . 26
11.1.4 Leadership within the scientific community . . . 27
11.1.5 Scientific expertise . . . 27
11.1.6 Research administration . . . 27
11.2 Teaching - Supervision - Juries . . . 28
11.2.1 Teaching . . . 28
11.2.2 Supervision . . . 29
11.2.3 Juries . . . 30
11.3 Popularization . . . 31
11.3.1 Education . . . 31
11.3.2 Interventions . . . 31

12 Scientific production **31**
12.1 Major publications . . . 31
12.2 Publications of the year . . . 32
12.3 Cited publications . . . 35

Project-Team MULTISPEECH

Creation of the Project-Team: 2015 July 01

Keywords

Computer sciences and digital sciences

A3.4. – Machine learning and statistics
A3.4.1. – Supervised learning
A3.4.2. – Unsupervised learning
A3.4.3. – Reinforcement learning
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
A5.6.3. – Avatar simulation and embodiment
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
A5.10.5. – Robot interaction (with the environment, humans, other robots)
A6.2.4. – Statistical methods
A6.3.1. – Inverse problems
A6.3.4. – Model reduction
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, until Sep 2023]
- Dominique Fohr [CNRS, Researcher, until Feb 2023]
- Yves Laprie [CNRS, Senior Researcher, HDR]
- Paul Magron [INRIA, Researcher]
- Mostafa Sadeghi [INRIA, ISFP]
- Emmanuel Vincent [INRIA, Senior Researcher, HDR]

Faculty Members

- Slim Ouni [Team leader, UL, Professor, HDR]
- Domitille Caillat [UNIV MONTPELLIER III, Associate Professor Delegation, from Sep 2023]
- Vincent Colotte [UL, Associate Professor]
- Irina Illina [UL, Associate Professor, HDR]
- Romain Serizel [UL, Associate Professor, HDR]

Post-Doctoral Fellows

- Constance Douwes [INRIA, Post-Doctoral Fellow, from Sep 2023]
- Felix Gontier [INRIA, Post-Doctoral Fellow, until Jan 2023]
- Ama Marina Kreme [INRIA, Post-Doctoral Fellow, until Sep 2023]
- Marie-Anne Lacroix [INRIA, Post-Doctoral Fellow, until May 2023]

PhD Students

- Louis Abel [UL]
- Jean Eudes Ayilo [INRIA, from Oct 2023]
- Sofiane Azzouz [UL, from Oct 2023]
- Raphael Bagat [CNRS, from Oct 2023]
- Tulika Bose [UL, until Jan 2023]
- Pierre Champion [INRIA, until Jun 2023]
- Can Cui [INRIA]
- Stephane Dilungana [INRIA, until Nov 2023]
- Sandipana Dowerah [INRIA, until May 2023]
- François Effa [UNIVERSITE LYON 1]
- Mickaella Grondin-Verdon [CNRS]

- Seyed Ahmad Hosseini [UL, until Mar 2023]
- Taous Iatariene [ORANGE, from Mar 2023]
- Nasser-Eddine Monir [UL]
- Sewade Ogun [INRIA]
- Robin San Roman [Meta AI]
- Shakeel Sheikh [UL, until Mar 2023]
- Vinicius Souza Ribeiro [UL]
- Prerak Srivastava [INRIA, until Oct 2023]
- Nicolas Zampieri [UL, ATER, until Aug 2023]
- Georgios Zervakis [INRIA, until Mar 2023]

Technical Staff

- Sofiane Azzouz [CNRS, Engineer, until Oct 2023]
- Théo Biasutto-Lervat [Inria, Engineer, SED]
- Sam Bigeard [INRIA, Engineer, from Dec 2023]
- Louis Delebecque [CNRS, Engineer]

Interns and Apprentices

- Jean Eudes Ayilo [INRIA, Intern, from Mar 2023 until Aug 2023]
- Raphael Bagat [UL, Intern, from Mar 2023 until Aug 2023]
- Hugo Bergerat [INRIA, Intern, from Jul 2023]
- Antoine Bruez [INRIA, Intern, from Apr 2023 until Aug 2023]
- Aine Drelingyte [UL, Intern, from Jul 2023 until Sep 2023]
- Lucas Duchene [UL, Intern, from Apr 2023 until Jun 2023]
- Solene Faure [UL, Intern, from Feb 2023 until Apr 2023]
- Valentin Gerard [UL, Intern, until Aug 2023]
- Soklong Him [INRIA, Intern, from Mar 2023 until Aug 2023]
- Louis Lalay [ENS PARIS-SACLAY, Intern, from Mar 2023 until Sep 2023]
- Berne Nortier [INRIA, Intern, from Apr 2023 until Aug 2023]
- Alexandre Perrot [UL, Intern, from Feb 2023 until Apr 2023]
- Emilien Visentini [INRIA, Intern, from Mar 2023 until Aug 2023]

Administrative Assistant

- Emmanuelle Deschamps [INRIA]

Visiting Scientists

- Antonio Almudevar [UNIV SARAGOSSE, from Sep 2023]
- Ladislav Mosner [UNIV TECH BRNO, from Sep 2023 until Oct 2023]

2 Overall objectives

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 and 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 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 the 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 the temporal, spatial and logical structure of ambient sound scenes over a long duration is also considered.

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 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 expressions 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

This research program aims to develop a unified software platform for embodied voice assistants, fueled by our research outcomes. The platform will not only aid our research but also facilitate other researchers in the field of human-computer interaction. It will also help in creating practical applications for human interactions, with a primary focus on 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 into 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

The Défi Inria COLaF co-led by S. Ouni aims to increase the inclusiveness of speech technologies by releasing open data, models and software for accented French and for regional, overseas and non-territorial languages of France.

A. Deleforge co-chaired 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 worked on the compression of large Wav2vec 2.0 audio models for embedded devices. Her work was 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. [43] 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 energy consumption and should be normalized for a fair comparison across systems. Additionally, they proposed an analysis of the relationship between energy consumption and the sound event detection performance that calls into question the current way to evaluate systems. Following this study, C. Douwes et al. [46] proposed a tutorial on how to effectively measure the energy consumption of machine listening systems at DCASE workshop.

6 Highlights of the year

6.1 Awards

E. Vincent received the 2023 IEEE SPS Sustained Impact Paper Award for the paper [56].

The startup Nijta co-founded by E. Vincent was awarded the i-Lab Prize of the national innovation challenge organized by the French Ministry of Higher Education, Research and Innovation in partnership with Bpifrance.

L. Abel is the recipient of the national Pépite award (for the startup Dynalips co-founded by S. Ouni, L. Abel and T. Biasutto–Lervat)

7 New software, platforms, open data

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: We migrated the software to a more robust server and fixed a few bugs. We also reorganized the code hierarchy under Gitlab to facilitate easier updates in the future.

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

Contact: Theo Biasutto-Lervat

7.1.2 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-vector 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.

News of the Year: A transfer contract was signed with the startup Nijta.

Contact: Nathalie Vauquier

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

7.1.3 Web Multimodal Annotation

Name: Web-based Multimodal Data Annotation

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

Scientific Description: Web Multimodal Annotation offers the reading, segmentation, and annotation of audio and video files, while providing a visualization of the associated waveform and spectrogram. Annotations can be manual or automatic on audio or video type data. In addition, the tool allows for manual correction of annotations, whether they have been added automatically or manually by annotators. To ensure extensive compatibility, the application uses the TextGrid (Praat) format.

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.

News of the Year: During this year, we implemented an automatic alignment system for the automatic labeling of audio or video data. In addition, we simplified the deployment of our Web Multimodal Annotation solution on servers, particularly by developing Docker images. We also worked on improving the performance and speed of this tool, significantly reducing the loading times required for the visualization of audio and video data. To overcome challenges related to the size of audio and video files, we integrated a streaming system. To ensure compatibility, we added the TextGrid format, which is widely used in the fields of linguistics, phonetics, and automatic language processing research.

Contact: Slim Ouni

Participants: Slim Ouni, Sofiane Azzouz

Partners: CNRS, Université de Lorraine

7.1.4 VAC

Name: Virtual Assistant Creator

Keywords: Artificial intelligence, Audio signal processing, Speech processing

Functional Description: VAC is a framework for creating interconnectable components for speech and natural language processing. It also offers a set of standard components for creating virtual assistants such as noise reduction, speech recognition and synthesis, and natural language understanding.

Release Contributions: Middleware for communication between components - Speech processing components - Access to microphone and speakers (via python-sounddevice and libportaudio2) - Denoising (ConvTasNet model) - Activity detection (via webrtcvad) - Speech recognition (K2/Sherpa zipformer model, NeMo ConformerTransducer model) - Speech synthesis (BalacoonTTS, FastPitch and HifiGAN from NeMo) - Natural language processing components - Text completion (via llama-cpp) - Chat (via llama-cpp) - Video processing components - Face detection (via dlib) - Facial landmark detection (via dlib) - Body detection (via opencv2) - Body pose estimation (via opencv2)

URL: https://gitlab.inria.fr/multispeech/vac

Contact: Theo Biasutto-Lervat

7.1.5 PercEval

Name: Perceptive Evaluation Platform

Keywords: Evaluation, Speech processing

Functional Description: PercEval offers users the ability to create surveys with multimedia content, publish them, and analyze their results.

Release Contributions: - survey administration webpage - survey creation webpage - survey response webpage - survey analytics webpage

URL: https://perceval.inria.fr

Contact: Theo Biasutto-Lervat

7.2 New platforms

Participants: Théo Biasutto-Lervat, Emmanuel Vincent, Slim Ouni, Vincent Colotte.

7.2.1 Virtual Assistant Creator

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 can 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 the second year of development, we mainly focused on the final design of the middleware API, relying on Protobuf and ZeroMQ, and on the implementation of the cross-platform python framework.

Moreover, we also add several key components into the python library, such as speech processing components (microphone and speaker access, voice activity detection with webrtcvad, speech enhancement with ConvTasNet, speech recognition with ZipFormer and ConformerTransducer, speech synthesis with BalacoonTTS, FastPitch and HifiGAN), video processing components (Face bounding-box and landmarks detection with dlib, Body bounding-box and pose estimation with opencv2) and text processing components (text completion and chat with llama-cpp).

7.3 Open data

7.3.1 MRI Dataset

The study of articulatory gestures has a wide range of applications, particularly in the study of speech production and automatic speech recognition. Real-time MRI data is particularly interesting because it offers good temporal resolution and complete coverage of the medio-sagittal section of the vocal tract. Existing MRI databases focus mainly on English, although the articulation of sounds depends on the language concerned. We therefore acquired MRI data for 10 native French speakers with no speech production or perception problems. A corpus consisting of sentences was used to ensure good phonetic coverage of French. Real-time MRI technology with a temporal resolution of 20 ms was used to acquire images of the vocal tract of the participants during speech production. The Sound was recorded simultaneously, denoised and temporally aligned with the images. The Speech was transcribed to obtain the phonetic segmentation. We also acquired static 3D MRI images of the French phonemes. In addition, we included annotations on spontaneous swallowing. Data are available together with the presentation of the database https://www.nature.com/articles/s41597-021-01041-3.

8 New results

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

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

8.1.1 Axis 1.1 — Integrating domain knowledge

Integration of symbolic knowledge. Transformer-based language models embed linguistic and commonsense knowledge and reasoning capabilities which are not always correct. We believe that integrating symbolic knowledge and reasoning into these models is a necessary step towards making them more trustworthy. Georgios Zervakis successfully defended his PhD on this topic [52]. 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 released Find2Find, a joint anaphora resolution and object localization dataset consisting of 500 anaphora-annotated recipes with corresponding videos, and we presented experimental results of a novel end-to-end joint multitask learning framework for these two tasks [39].

Optimization-based signal processing. Optimization approaches for signal 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. Besides, in such a framework, light neural networks can be used as a proxy to obtain appropriate prior information and/or initialization. We derived optimization-based algorithms for data-efficient speech signal restoration [20]. We also combined such algorithms with factorization models for music signal restoration [11]. Finally, we derived novel optimization algorithms for audio source separation and speech enhancement [34]. These have the potential to be further unfolded into neural networks in order to perform these tasks in an end-to-end fashion.

Generative-based speech enhancement. A widely used approach for speech enhancement consists in directly learning a deep neural network (DNN) to estimate clean speech from input noisy speech. Despite its promising performance, it comes with two main challenges. First, it requires very large DNNs to learn over a huge dataset, covering many noise types, noise levels, etc. Second, its generalisation is usually limited to seen environments. Unsupervised speech enhancement tries to address these challenges by proposing to learn only clean speech distribution and model noise at inference. Along with this line of research, we proposed a novel diffusion-based speech enhancement method [36] that leverages the power of diffusion-based generative models, currently showing great performance in computer vision. Furthermore, we developed a new training loss for diffusion-based supervised speech enhancement [19], which bridges the gap between the performance of supervised and unsupervised speech enhancement approaches.

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 method enhances the quality of training on a curated dataset and paves the way for automated TTS dataset curation across a broader spectrum of languages. [37].

Domain-specific ASR systems are usually trained or adapted on a suitable amount of transcribed speech data. By contrast, we studied the training and the adaptation of recurrent neural network (RNN) ASR language models from a small amount of untranscribed speech data using multiple ASR hypotheses

embedded in ASR confusion networks. Our sampling-based method achieved up to 12% relative reduction in perplexity on a meeting dataset as compared to training on ASR 1-best hypotheses [15].

Self-supervised learning. We proposed a subband diffusion based sound signal reconstruction from discrete compressed representations [40]. While diffusion-based reconstruction approaches were mainly targetting speech signals we proposed a high-fidelity multi-band diffusion-based framework that generates any type of audio (e.g., speech, music, environmental sounds) from low-bitrate discrete representations.

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 the speaker (x-vector), phonetic and pitch features and resynthesizing the signal with a different target x-vector. To reduce the amount of residual speaker information in the phonetic and pitch features, we explored the use of Laplacian noise [14] inspired from differential privacy. Pierre Champion defended his PhD [48], and we released the report [8] of the interdisciplinary Dagstuhl Seminar organized in 2022 on this topic.

8.2 Axis 2 — Extracting information from speech signals

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

8.2.1 Axis 2.1 — Linguistic speech content

Speaker-attributed ASR End-to-end ASR has enabled the transcription of overlapping speech utterances using speaker-attributed ASR (SA-ASR) systems. We presented an end-to-end multichannel SA-ASR system that combines a Conformer-based encoder with multi-frame crosschannel attention and a speaker-attributed Transformer-based decoder. To the best of our knowledge, this is the first model that efficiently integrates ASR and speaker identification modules in a multichannel setting [21, 55].

Rich transcriptions We proposed two approaches to joint rich and normalized ASR, that produces transcriptions both with and without punctuation and capitalization. The first approach, which uses a language model to generate pseudo-rich transcriptions of normalized training data, performs better on out-of-domain data, with up to 9% relative error reduction. The second approach, which uses a single decoder conditioned on the type of output, demonstrates the feasibility of joint rich and normalized ASR using as little as 5% rich training data with moderate (2.4% absolute) error increase [53].

Emotion representation During the master internship of Raphaël Bagat, we investigated the representation of emotion in latent space. We combined several acoustic representations from melspectrum, extracted features or Wav2Vec2 encoding, with linguistic representation based on SBERT of the utterance. We evaluated the latent representation in several places in an emotion recognition system (concatenation after encodings or during steps of decoding). We showed that the kind of combination is very polarizing and one of both modalities can be under-leveraged. The system with Wav2Vec2 and with a secondary system with only-acoustic emotion detection gives 71% of recognition. A constractive loss have been also used but without giving a significant gain. The latent representations have been used to evaluate a set of emotional acoustic features (eGeMAPS) in order to evaluate the coded information contained in the latent space. This work could be continued to drive an expressive TTS system with this kind of latent representation to express an emotion.

Dyslexia and non-native speech In the framework of our study concerning the perception of German fricatives by French dyslexic subjects, we analyzed the homogeneity of the answers inside the groups of dyslexic people and average readers. 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. Previous results have shown that people with dyslexia exhibited a slight deficit in the categorization of all the sounds and a relatively poor discrimination of the new sound /ç/, which invalidated Serniclaes' hypothesis about a better sensitivity to universal contrasts by dyslexic people. At first sight, this result seems to corroborate the relatively common view that dyslexic people have impoverished phonological representations. A new analysis of the results for each individual showed, in agreement with those of Hazan for L1, that a substantial number of dyslexic individuals (approximately half of people with dyslexia, in our study) behave like average readers (a very homogeneous group). Thus for a large number of individuals with dyslexia, it appears that phonological representations are in fact intact. With respect to L2, this last group is not disadvantaged by dyslexia (as would suggest the hypothesis of impoverished phonological representations), at least at the phonological level and for the sounds present in this study.

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 considered transferring knowledge from a resource-rich source to a low-resource target with fewer labeled instances, across different online platforms. A novel training strategy is proposed, which allows flexible modeling of the relative proximity of neighbors retrieved from the resource-rich corpus to learn the amount of transfer. We incorporate neighborhood information with Optimal Transport that permits exploiting the embedding space geometry. By aligning the joint embedding and label distributions of neighbors, substantial improvements are obtained in low-resource hate speech corpora [47]. Moreover, in [47], we proposed two DA approaches using feature attributions, which are post-hoc model explanations. Particularly, the problem of spurious corpusspecific correlations is studied that restricts the generalizability of classifiers for detecting hate speech, a sub-category of abusive language. While the prior approaches rely on a manually curated list of terms, we automatically extracted and penalized the terms causing spurious correlations. Our dynamic approaches improved the cross-corpus performance over previous works both independently and in combination with pre-defined dictionaries.

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 [45], we studied the impact of the self-attention mechanism and the multi-task learning for hate speech detection. The two tasks that we want to achieve are the MWE identification task and the hate speech detection task. We carried out our experiments on four corpora and using two contextual embeddings. We observed that multi-task systems significantly outperform the baseline single-task system. The best performance is obtained using the multi-task system with two attention heads.

8.2.2 Axis 2.2 — Speaker identity and states

Speaker recognition. We proposed diffusion probabilistic models 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 [25]. This approach was further extended to replace the fully supervised joint training stage by a self supervised joint training stage [24].

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. 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. The stuttered speech is usually available in limited amounts and is highly imbalanced. This year, we addressed the class imbalance problem via a multibranching scheme and by

weighting the contribution of classes in the overall loss function, resulting in a huge improvement in stuttering classes on the SEP-28k dataset over the baseline (StutterNet) [17]. We have also applied speech embeddings from pre-trained deep learning models, specifically ECAPA-TDNN and Wav2Vec2.0, for various tasks. When benchmarked with traditional classifiers on Speaker Diarization tasks, our method outperforms standard systems trained on the limited SEP-28k dataset, with further improvements observed when combining embeddings and concatenating multiple layers of Wav2Vec2.0 [16]. Shakeel Sheikh defended his PhD thesis on February 24th, 2023 [50].

8.2.3 Axis 2.3 — Speech in its environment

Ambient sound recognition. 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) 2023. For this new edition we have been investigating new evaluation metrics that are pontential more independent to post processing tuning [26]. This evaluation method can provide a more complete picture of the systems behaviour under different working conditions. 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 [43].

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 [30].

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 [23]. 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. Targeting speech enhancement for hearing aids, we also started investing the performance of speech enhancement at a fine grained phonetic level. The goal here is to link the results obtained with objective metrics to the outcome of listening tests conducted at our partner site (Institut de l'audition).

Acoustic Parameter Estimation. To estimate acoustic quantities of interest from speech signals, state-of-the-art methods rely on supervised learning on simulated data. However, few studies carefully examine the impact of acoustic simulation realism. We contributed such a study on speech direction-of-arrival estimation [44] and revealed that improving the realism of source, microphone and wall responses at training time consistently and significantly improves generalization to real data. Prerak Srivastava defended his PhD on this topic [51].

Fast and efficient unsupervised speech enhancement. A new direction in speech enhancement involves an unsupervised framework. Unlike the common supervised method, which trains models using both clean and noisy speech data, the unsupervised method trains solely with clean speech. This training often employs variational autoencoders (VAEs) to create a data-driven speech model. This unsupervised approach could significantly enhance generalisation performance while keeping the model less complex than supervised alternatives. However, unsupervised methods typically require more computational resources during the inference (enhancement) phase. To address this issue, we have introduced several fast and efficient inference techniques tailored for speech enhancement, using posterior sampling strategies [41, 42]. Initially, we applied these techniques to a basic, non-sequential VAE model [41]. Later, we adapted them for more advanced dynamical VAE models and introduced additional sampling-based methods [42]. Our experiments demonstrated the effectiveness of the proposed methods, narrowing the performance gap between supervised and unsupervised approaches in speech enhancement.

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, Domitille Caillat, 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 [10] 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. 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 audio-visual speech enhancement tasks. The results confirmed the benefits of the proposed framework over the previous works.

Unsupervised performance analysis of face frontalization for audio-visual speech processing. In a related work [13], we addressed the problem of analyzing the performance of 3D face alignment (3DFA) methods, which is a necessary preprocessing step for audio-visual speech processing. While typically reliant on supervised learning from annotated datasets, 3DFA faces annotation errors, which could strongly bias the results. We explored unsupervised performance analysis (UPA), centering on estimating the rigid transformation between predicted and model landmarks. This approach is resilient to non-rigid facial changes and landmark errors. UPA involves extracting 3D landmarks from a 2D face, mapping them onto a canonical pose, and computing a robust confidence score for each landmark to determine their accuracy. The methodology is tested using public datasets and various 3DFA software, demonstrating consistency with supervised metrics and effectiveness in error detection and correction in 3DFA datasets.

Audio-visual speech enhancement with dynamical VAE. Variational autoencoder (VAE) based generative models have shown potential in unsupervised audiovisual speech enhancement (AVSE), but current models do not fully leverage the sequential nature of speech and visual data. In a recent work [29], we introduced an audio-visual deep Kalman filter (AV-DKF) generative model, which combines audio-visual data more effectively using a first-order Markov chain for latent variables and an efficient inference method for speech signal estimation. Experiments comparing various generative models highlighted the AV-DKF's superiority over audio-only and non-sequential VAE-based models in speech enhancement.

8.3.2 Axis 3.2 — Multimodal speech generation

Acquisition of rt-MRI (real-time Magnetic Resonance Imaging) data for French. This year, in collaboration with the IADI laboratory (P.-A. Vuissoz and K. Isaieva), we recorded rt-MRI data for 3 stuttering subjects plus a normally fluent control subject within the framework of the BENEPHIDIRE ANR project. We also recorded beatboxing data for an expert in the field. Those data enabled the investigation of the preparation of beatboxing patterns and the corresponding places of articulation [22]. Finally, the manual correction of the phonetic segmentation of the large database (2100 sentences) recorded for one female speaker has been completed this year. This database [32] was exploited by Vinicius Ribeiro to achieve his evaluation of the prediction of the vocal tract shape from a sequence of phonemes to be articulated.

Processing real-time Magnetic Resonance Imaging The exploitation of real-time MRI data requires the capability to process a very large number of images automatically. We have continued our work on segmenting MRI images so that we can move easily from one speaker to another and identify all the articulators. A Mask R-CNN network was trained to detect and segment the vocal tract articulator

contours in two real-time MRI (rt-MRI) datasets with speech recordings of multiple speakers [12]. Two post-processing algorithms were then proposed to convert the network's outputs into geometrical curves. Nine articulators were considered: the two lips, tongue, soft palate, pharynx, arytenoid cartilage, epiglottis, thyroid cartilage, and vocal folds. Rt-MRI of the vocal tract is often performed in 2D because, despite its interest, 3D rt-MRI does not offer sufficient quality. The goal of this study was to test the applicability of super-resolution algorithms for dynamic vocal tract MRI [9]. In total, 25 sagittal 2D slices of 8 mm with an in-plane resolution of 1.6 × 1.6 mm2 were acquired consecutively. The slices were aligned using the simultaneously recorded speech signal. The super-resolution strategy was used to reconstruct 1.6 × 1.6 × 1.6 mm3 isotropic volumes. The resulting images were less sharp than the native 2D images but demonstrated a higher signal-to-noise ratio. Super-resolution also allows for eliminating inconsistencies leading to regular transitions between the slices. The proposed method allows for the reconstruction of high-quality dynamic 3D volumes of the vocal tract during natural speech.

Evaluating articulatory synthesis of speech through phoneme recognition. Our recent work on the generation of the temporal vocal tract shape from a sequence of phonemes to be articulated exploited a large real time MRI database. However, the assessment of the shape quality still needs to be included in the process. Ranking generative models is tricky since the acoustic simulation alone is not good enough to guarantee that it does not introduce strong perceptive biases. A purely geometric assessment is therefore generally used, which is itself insufficient to deal with articulatory speaker variability.

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 enable 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 Flow-based generative models are widely used in text-to-speech (TTS) systems to learn the distribution of audio features given the input tokens. Yet the generated utterances lack diversity and naturalness. We proposed to improve the diversity of utterances by explicitly learning the distribution of pitch contours of each speaker during training using a stochastic flow-based pitch predictor, then conditioning the model on generated pitch contours during inference. The experimental results demonstrate that the proposed method yields a significant improvement in the naturalness and diversity of generated speech [38].

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 concurrent generation of speech and gestures by the speaker, taking into account both non-verbal and verbal gestures. In the context of Louis Abel's Ph.D., we focus on non-verbal gesture generation (upper body and arms) derived from the acoustic signal and the text. The first step is to autoencode the motion representation from positions to obtain a latent representation. The model is based on Graphical Neural Networks to leverage the skeleton's constraints. The motion is then predicted using a flow-based architecture from the motion (latent) representation, audio speech feature context, and the text of the utterance. The model is based on a short context window for recurrent architecture. At present, we are conducting perceptive experiments to assess the benefits of using GNN in this type of generation. In Mickaëlla Grondin-Verdon's PhD, our focus shifts more towards

dyadic gestures by analyzing specific gesture components, the strokes (duration, intensity, alignment, etc.). The aim is to feed models with this data to predict and generate gestures in a dialogue context. This year, we used the BEAT corpus, a comprehensive and varied multimodal corpus, to examine gesture categories, temporal references, durations, and comparisons. This analysis helps us in gaining a deeper understanding of the corpus structure and exploit the data in subsequent project stages. Since joining Multispeech, Domitille Caillat has been working on the role of multimodal annotations in the generation and automatic recognition of gestures. She plans to manually annotate a sample of the BEAT corpus, which is used by several Multispeech members, with the goal to objectively evaluate the results of gesture generation through a measured comparison of authentic data and generated data (frequency of gestures, favored places of occurrence, relationship between gestures and lexical affiliates, etc.).

9 Bilateral contracts and grants with industry

9.1 Bilateral grants with industry

9.1.1 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.

9.1.2 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.

9.1.3 Orange Labs

- Company: Orange Labs (France)
- Duration: March 2023 – Feb 2026
- Participants: Taous Iatariene, Romain Serizel
- Abstract: This CIFRE grant funds the PhD of Taous Iatariene on sound source tracking.

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Participation in other International Programs

ANR-JST Confluence

Title: semantiC segmentatiON oF compLex soUnd scEnes oN edge deviCEs

Duration: Dec 2023 – Nov 2027

Partners:

- CEA-List (France)
- Sonaide (France)
- NTT Communication Science Laboratories (Japan)
- Tokyo Metropolitan University (Japan)

Coordinator: Nicolas Turpault (Sonaide), Noboru Harada (NTT)

Participant: Romain Serizel

Summary: The CONFLUENCE project aims to develop artificial intelligence (AI) technologies for sound semantic segmentation of acoustic signals that can recognize sound events and separate/isolate the signals of the sound sources forming semantic entities. The objective is to design an embedded AI system that can implement these technologies for two telecom application services: immersive communication and home monitoring/assistance devices.

10.2 International research visitors

10.2.1 Visits of international scientists

Antonio Almudévar

Status PhD

Institution of origin: Universidad de Zaragoza

Country: Spain

Dates: 18/09/2023 – 21/12/2023

Context of the visit: The main objective of this rsearch stay was to work on the tasks of Classification, Few Shot Classification and Unsupervised Anomaly Detection.

Mobility program/type of mobility: research stay

Ladislav Mosner

Status PhD

Institution of origin: Brno University of Technology

Country: Czech Republic

Dates: 23/09/2023 – 23/10/2023

Context of the visit: The main objective of this research stay was to work on multichannel speaker representation

Mobility program/type of mobility: research stay

10.3 European initiatives

10.3.1 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.

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.

HumanE-AI-Net

Title: Making artificial intelligence human-centric

Duration: Sep 2020 – Aug 2024

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.

10.3.2 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.

10.4 National initiatives

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 ambient noise, reverberation and short speech utterances.

ANR BENEPHIDIRE

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

Duration: Mar 2019 - Jun 2024

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.

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 and the face, and integrating the digital simulation of the aero-acoustic phenomena.

ANR Lorraine Artificicial Intelligence – LOR-AI LOR-AI

Title: Lorraine Artificicial Intelligence Cofinancement de thèses en IA

Duration: Sep 2020- Dec 2025

Coordinator: Yves Laprie

Partners: CNRS, Inria, Regional University Hospital Centre (CHRU)

Participants: Doctoral schools of Université de Lorraine

Abstract: This project about Artificial Intelligence, led by the Université de Lorraine (UL), has a double objective by providing 12 co-funding for doctoral theses: on the one hand, to strengthen UL areas of excellence in AI and domains tightly connected to IA, i.e. particularly Health, and on the other hand, to open other research areas to AI with the objective of leading to scientific breakthroughs.

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 both fit the needs of patients suffering of hearing losses and the computational constraints of hearing aid devices.

ANR LLM4all

Title: Large Language Models for All

Duration: Oct 2023 - Mars 2027

Coordinator: Synalp Loria (Nancy)

Partners: LORIA-Synalp, LORIA-Multispeech, LIX, Linagora, Ap-HP, HuggingFace

Participants: Irina Illina, Emmanuel Vincent

Abstract: Large Language Models (LLM) of sufficient size exhibit outstanding emergent abilities, such as learning from their input context and decomposing a complex problem into a chain of simpler steps. The LLM4all project will thus focus on such large models, or on models at the same level of generic performances, and will propose methods to solve two related fundamental issues: how to update these LLMs automatically, and how to reduce their computing requirements in order to facilitate their deployment.

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.

DGA DEEP MAUVES

Title: Deep automatic aircraft speech recognition for non native speakers

Duration: Dec 2022 – Dec 2026

Coordinator: Irina Illina

Participant: Irina Illina, Raphaël Bagat, Emmanuel Vincent

Summary: This project proposes methods and tools that increase the usability of ASR systems for non-native speakers in noisy conditions in the aeronautical domain.

Défi Inria COLaF

Title: Corpus et Outils pour les Langues de France

Duration: Aug 2023 – Jul 2027

Coordinator: Slim Ouni and Benoît Sagot (Inria ALMANACH)

Partners: Inria ALMANACH (Paris)

Participant: Slim Ouni, Sam Bigeard, Vincent Colotte, Emmanuel Vincent

Summary: This project aims to increase the inclusiveness of speech technologies by releasing open data, models and software for accented French and for regional, overseas and non-territorial languages of France.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: selection

Chair of conference program committees

- Area chair, 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (E. Vincent)
- Technical program co-chair, 2023 Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) (R. Serizel)

Member of the conference program committees

- Area co-chair, 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (A. Deleforge, R. Serizel)

Reviewer

- CHiME 2023 – 7th International Workshop on Speech Processing in Everyday Environments (E. Vincent)
- INTERSPEECH 2023 (E. Vincent, R. Serizel, I. Illina, Y. Laprie)
- ICPhS 2023 (A. Bonneau)
- PaPE 2023 (A. Bonneau)
- ICML 2023 (A. Deleforge)
- ICASSP 2023 (A. Deleforge, R. Serizel, I. Illina)
- ASRU 2023 (I. Illina)
- LTC 2023 (I. Illina)
- ISCA SPSC (I. Illina)
- EUSIPCO 2023 (V. Colotte, A. Deleforge)
- WASPAA 2023 (A. Deleforge)
- Neurips 2023 (A. Deleforge)

11.1.2 Journal

Member of the editorial boards

- Guest Editor of Neural Networks, special issue on Advances in Deep Learning Based Speech Processing [18] (E. Vincent)
- Associate Editor of EURASIP Journal on Audio, Speech and Music Processing (A. Deleforge)
- Associate Editor IEEE/ACM Transactions on Audio Speech and Language Processing (R. Serizel)
- Associate Editor IEEE Open Journal in Signal Processing (R. Serizel)

Reviewer - reviewing activities

- Journal of the Acoustical Society of America (A. Deleforge, Y. Laprie)
- IEEE Signal Processing Letters (A. Deleforge)
- IEEE Transactions in Audio, Speech and Language Processing (A. Deleforge, R. Serizel, Y. Laprie)
- IEEE Transactions on Signal Processing (A. Deleforge)

11.1.3 Invited talks

- A brief introduction to multichannel noise reduction with deep neural networks, International Symposium on Auditory and Audiological Research (ISAAR), Denmark, August 2023 (R. Serizel)

11.1.4 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)
- Board member of Le VoiceLab, the association of French voice tech players (E. Vincent)
- Member of the IEEE Technical Committee on Audio and Acoustic Signal Processing (A. Deleforge)
- 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)
- Chair of the Steering Group of DCASE (R. Serizel)

11.1.5 Scientific expertise

- Expert for the Data Governance Working Group of the Global Partnership on AI (GPAI) (E. Vincent)
- Expert for HORIZON Europe 2023 (S. Ouni, Y. Laprie)
- Expertise for the Cognition Institute Carnot (A. Bonneau)
- Expertise for Collaborative Research in Computational Neuroscience (CRCNS) (R. Serizel)
- Expertise for the the Czech Science Foundation (I. Illina)

11.1.6 Research administration

- Head of Science of Inria Nancy – Grand Est (E. Vincent)
- Member of Inria's Evaluation Committee (E. Vincent)
- Member of the Comité Espace Transfert of Inria Nancy – Grand Est (E. Vincent)
- Vice-chair of the hiring committee for Junior Research Scientists, Inria Nancy – Grand Est (E. Vincent)
- Member of the admission committee for Inria Starting Faculty Positions, Inria Nancy – Grand Est (E. Vincent)
- Co-Chair of the local commission for sustainable development, Inria Nancy – Grand Est (A. Deleforge)
- Referent for research data, Inria Nancy – Grand Est (A. Deleforge)
- Member of the Commission de Développement Technologique, Inria Nancy – Grand Est (R. Serizel)
- Member of the Comipers, Inria Nancy – Grand Est (R. Serizel)
- Member of the Comité Utilisateurs des Moyens de Calculs, Inria Nancy – Grand Est (T. Biasutto–Lervat)
- Referent Plateformes-Outils, Inria Nancy – Grand Est (T. Biasutto–Lervat)
- Head of pole scientifique Automatique, Mathematiques, Informatique et leurs interactions (AM2I) (Y. Laprie)
- Member of the board, Université de Lorraine (Y. Laprie)

- Member of the bureau du pole scientifique Automatique, Mathematiques, Informatique et leurs interactions (AM2I) (I. Illina)
- Member of the Comité du pole scientifique Automatique, Mathematiques, Informatique et leurs interactions (AM2I) (I. Illina)
- Member of the evaluation committee of Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur (HCERES) for STIH (I. Illina)
- Member of the RIPEC jury, UL (I. Illina)
- Member of the admission committee for Nantes University (I. Illina)
- Member of the selection committee for the position of senior lecturer in phonetics at the University Paul Valéry of Montpellier (Y. Laprie)
- Academic representative of Université de Lorraine at AIDA (Artificial Intelligence Doctoral Academy) (S. Ouni)

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), Data Structures (50h), 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 Algorithms (24 hours), L2, Université de Lorraine, France
- Licence: A. Bonneau, Phonetics (18 hours), L1, *École d'orthophonie*, Université de Lorraine, France
- Licence: Y. Laprie, 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, Multimodal oral communication (24 hours), M2, Université de Lorraine, France
- Master: V. Colotte, AI introduction (6 hours), M2 - intellectual property rights, 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
- Master: A. Deleforge, Artificial Intelligence, Machine Learning, Deep Learning (24 hours), M1, 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, Co-Responsible of *parcours Ingénierie Logiciel*, BUT, Université de Lorraine, France

11.2.2 Supervision

- PhD: Tulika Bose, "Transfer Learning for Abusive Language Detection", Jan 2023, I. Illina and D. Fohr. [47]
- PhD: Shakeel Sheikh, "Deep learning for stuttering detection", Feb 2023, S. Ouni [50].
- PhD: Georgios Zervakis, "Enriching large language models with semantic lexicons and analogies", Mar 2023, M. Couceiro (LORIA) and E. Vincent [52].
- PhD: Sandipana Dowerah, "Deep Learning-based Speaker Verification In Real Conditions", May 2023, D. Jouvet and R. Serizel [49].
- PhD: Prerak Srivastava, "Realism in virtually supervised learning for acoustic room characterization and sound source localization", Nov 2023, A. Deleforge and E. Vincent [51].
- PhD: Nicolas Zampieri, "Détection des discours haineux dans les réseaux sociaux : apport des expressions polylexicales", Dec 2023, I. Illina and D. Fohr.
- PhD: Vinicius Ribeiro, "Deep Supervision of the Vocal Tract Shape for Articulatory Synthesis of Speech", Dec 2023, Y. Laprie
- PhD in progress: Stéphane Dilungana, "New Algorithms for Automated Room Acoustic Diagnosis", Oct 2020, A. Deleforge and C. Foy and S. Faisan.
- 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: Raphaël Bagat, "Automatic speech recognition for non-native speakers in a noisy environment", Oct 2023, I. Illina 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: François Effa, "Caractérisation de la détectabilité d'alarmes dans le bruit", Dec 2020, J.-P. Arz (Institut National de Recherche et de Sécurité), Nicolas Grimault (Centre de Recherche en Neurosciences de Lyon) and R. Serizel

- PhD in progress: Robin San Roman, “Self supervised disentangled representation learning of audio data for compression and generation”, Jun 2022, Y. M. Adi (Meta AI), A. Deleforge, R. Serizel and G. Synnaeve (Meta AI)
- PhD in progress: Nasser-Eddine Monir, “Multichannel speech enhancement for patients with auditory neuropathy spectrum disorders”, Dec 2022, P. Magron and R. Serizel
- PhD in progress: Taous Iatariene, “Spatial tracking of multiple moving sound sources”, Mar 2023, A. Guérin (Orange Labs) and R. Serizel
- PhD in progress: Jean-Eudes Ayilo, “Audio-visual Speech Enhancement: Bridging the Gap between Supervised & Unsupervised Approaches”, Oct 2023, M. Sadeghi and R. Serizel
- PhD Sofiane Azzouz, November 2023, “Acoustic to Articulatory Inversion by using dynamic MRI images”, Y. Laprie and P-A. Vuissoz
- PhD in progress: Zahra Benslimane, “Neural Networks For Improving Speech Understanding In Real-Time Embedded Systems”, Dec 2023, T. Allenet (CEA-List), F. Auzanneau (CEA-List) and R. Serizel

11.2.3 Juries

- Participation in the PhD jury of Mohamed Nabih Ali Mohamed Nawar (University of Trento, Jan 2023), E. Vincent, reviewer
- Participation in the PhD jury of Ranya Aloufi (Imperial College London, Mar 2023), E. Vincent, opponent
- Participation in the PhD jury of Yann Teytaut (Sorbonne Université, Jul 2023), E. Vincent, chair
- Participation in the PhD jury of Xiaoyu Bie (Université de Grenoble, Oct 2023), A. Deleforge, examiner
- Participation in the PhD jury of Vladimir Iashin, (Tampere University, Apr 2023), R. Serizel, reviewer
- Participation in the PhD jury of Samuele Cornell (Università Politecnica delle Marche, May 2023), R. Serizel, reviewer
- Participation in the PhD jury of Arjun Pankajakshan (Queen Mary University, Jul 2023), R. Serizel, reviewer
- Participation in the PhD jury of Marc-Antoine Georges (Université Grenoble Alpes, May 2023), Y. Laprie, reviewer
- Participation in the PhD jury of Martin Lebourdais (Université du Mans, Oct 2023), R. Serizel, reviewer
- Participation in the PhD jury of Arjun Pankajakshan (Queen Mary University, Jul 2023), R. Serizel, reviewer
- Participation in the PhD jury of Alexis Dehais-Underdown (Université de la Sorbonne Nouvelle, Sept 2023), Y. Laprie, reviewer
- Participation in the PhD jury of Pablo Zinemanas (Universitat Pompeu Fabra, Oct 2023), R. Serizel, reviewer
- Participation in the PhD jury of Mohammad Mohammadamini (Avignon University), I. Illina, examiner
- Participation in the HDR jury of Marie Tahon (Université du Mans, Jan 2023), E. Vincent, reviewer
- Participation in the HDR jury of Nicolas Obin (Sorbonne Université, Sep 2023), E. Vincent, reviewer

11.3 Popularization

11.3.1 Education

- MATh.enJEANS - Presentation to college students of thr MULTIMOD platform (for multimodal data acquision) (L. Abel, T. Biasutto–Lervat)

11.3.2 Interventions

- IA : Quels défis pour demain ?, *Quoi de Neuf Chercheur?*, Twitch, Jul 2023 (E. Vincent)
- Interview on speech anonymization for *L'Esprit Sorcier*, Dec 2023 (E. Vincent)
- Video Interview "Let's talk about research with..." for Inria, Sep 2023 (A. Deleforge)
- L'éthique et l'IA, Podcast Monde Numérique, Sept 2023 (L. Abel)
- Showcase of the DeepLipsync technology, VivaTech, Jun 2023 (T. Biasutto–Lervat, L. Abel)

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. Fischer-Hübner, D. Klakow, P. Valcke and E. Vincent. 'Privacy in Speech and Language Technology'. In: *Dagstuhl Reports* 12.8 (1st Mar. 2023), pp. 60–102. DOI: 10.4230/DagRep.12.8.60. URL: https://inria.hal.science/hal-04015119.

[9] K. Isaieva, F. Odille, Y. Laprie, G. Drouot, J. Felblinger and P.-A. Vuissoz. 'Super-Resolved Dynamic 3D Reconstruction of the Vocal Tract during Natural Speech'. In: *Journal of Imaging* 9.10 (20th Oct. 2023), p. 233. DOI: 10.3390/jimaging9100233. URL: https://hal.science/hal-04373276.

[10] 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* 131.5 (1st May 2023), pp. 1122–1140. DOI: 10.1007/s11263-022-01742-1. URL: https://hal.science/hal-03902610.

[11] 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). DOI: 10.1016/j.sigpro.2022.108905. URL: https://inria.hal.science/hal-03708613.

[12] V. Ribeiro, K. Isaieva, J. Leclere, J. Felblinger, P.-A. Vuissoz and Y. Laprie. 'Automatic segmentation of vocal tract articulators in real-time magnetic resonance imaging'. In: *Computer Methods and Programs in Biomedicine* 243.2 (Jan. 2024). DOI: 10.1016/j.cmpb.2023.107907. URL: https://inria.hal.science/hal-04376938.

[13] M. Sadeghi, X. Alameda-Pineda and R. Horaud. 'Unsupervised Performance Analysis of 3D Face Alignment with a Statistically Robust Confidence Test'. In: *Neurocomputing* 564 (Jan. 2024). DOI: 10.1016/j.neucom.2023.126941. URL: https://hal.science/hal-04265797.

[14] 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). DOI: 10.48550/arXiv.2202.11823. URL: https://inria.hal.science/hal-03588932.

[15] I. A. Sheikh, E. Vincent and I. Illina. 'Training RNN Language Models on Uncertain ASR Hypotheses in Limited Data Scenarios'. In: *Computer Speech and Language* 83 (1st Jan. 2024), p. 101555. DOI: 10.1016/j.csl.2023.101555. URL: https://inria.hal.science/hal-03327306.

[16] S. Sheikh, M. Sahidullah, F. Hirsch and S. Ouni. 'Stuttering Detection Using Speaker Representations and Self-supervised Contextual Embeddings'. In: *International Journal of Speech Technology* (2nd June 2023). DOI: 10.1007/s10772-023-10032-1. URL: https://inria.hal.science/hal-03629758.

[17] S. A. Sheikh, M. Sahidullah, F. Hirsch and S. Ouni. 'Advancing Stuttering Detection via Data Augmentation, Class-Balanced Loss and Multi-Contextual Deep Learning'. In: *IEEE Journal of Biomedical and Health Informatics* (20th Feb. 2023). DOI: 10.1109/JBHI.2023.3248281. URL: https://inria.hal.science/hal-03998392.

[18] 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://inria.hal.science/hal-03883292.

International peer-reviewed conferences

[19] J.-E. Ayilo, M. Sadeghi and R. Serizel. 'Diffusion-based speech enhancement with a weighted generative-supervised learning loss'. In: International Conference on Acoustics Speech and Signal Processing (ICASSP). Seoul (Korea), South Korea, 14th Apr. 2024. URL: https://hal.science/hal-04210729.

[20] L. Bahrman, M. Krémé, P. Magron and A. Deleforge. 'Signal Inpainting from Fourier Magnitudes'. In: *European Signal Processing Conference (EUSIPCO)*. EUSIPCO 2023. Helsinki, Finland, 22nd June 2023. URL: https://hal.science/hal-03832480.

[21] C. Cui, I. A. Sheikh, M. Sadeghi and E. Vincent. 'End-to-end Multichannel Speaker-Attributed ASR: Speaker Guided Decoder and Input Feature Analysis'. In: 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2023). Taipei, Taiwan, 16th Dec. 2023. URL: https://inria.hal.science/hal-04235774.

[22] A. Dehais-Underdown, P. Vignes, L. Crevier-Buchman, D. Demolin, P.-A. Vuissoz, K. Isaieva, M. Fauvel, Y. Laprie and J. Felblinger. 'Non-pulmonic initiation in human beatboxing: a real-time MRI study'. In: 20th International Congress of Phonetic Sciences (ICPhS 2023). Prague, Czech Republic, 7th Aug. 2023. URL: https://hal.science/hal-04180282.

[23] L. Delebecque and R. Serizel. 'BinauRec: A dataset to test the influence of the use of room impulse responses on binaural speech enhancement'. In: *European Signal Processing Conference (EUSIPCO)*. EUSIPCO23. Helsiinki, Finland, 6th Sept. 2023. URL: https://hal.science/hal-04193377.

[24] S. Dowerah, A. Kulkarni, R. Serizel and D. Jouvet. 'Self-supervised learning with diffusion-based multichannel speech enhancement for speaker verification under noisy conditions'. In: *Proceedings of Interspeech 2023*. INTERSPEECH 2023. Dublin (Ireland), Ireland: ISCA, 20th Aug. 2023, pp. 3849–3853. DOI: 10.21437/Interspeech.2023-1890. URL: https://hal.science/hal-04151411.

[25] 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.

[26] J. Ebbers, R. Haeb-Umbach and R. Serizel. 'Post-Processing Independent Evaluation of Sound Event Detection Systems'. In: DCASE 2023 - 8th Workshop on Detection and Classification of Acoustic Scenes and Events. Tampere, Finland, 20th Sept. 2023. DOI: 10.48550/arXiv.2306.15440. URL: https://inria.hal.science/hal-04385022.

[27] F. Effa, R. Serizel, J.-P. Arz and N. Grimault. 'Lightweight Annotation and Class Weight Training for Automatic Estimation of Alarm Audibility in Noise'. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece: IEEE, 4th June 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10094730. URL: https://inria.hal.science/hal-04385004.

[28] A. Golmakani, M. Sadeghi, X. Alameda-Pineda and R. Serizel. 'A weighted-variance variational autoencoder model for speech enhancement'. In: International Conference on Acoustics Speech and Signal Processing (ICASSP). Seoul (Korea), South Korea, 14th Apr. 2024. URL: https://inria.hal.science/hal-03833827.

[29] A. Golmakani, M. Sadeghi and R. Serizel. 'Audio-visual speech enhancement with a deep kalman filter generative model'. In: International Conference on Acoustics Speech and Signal Processing (ICASSP). Rhodes island, Greece, 4th June 2023. URL: https://inria.hal.science/hal-03833814.

[30] F. Gontier, R. Serizel and C. Cerisara. 'SPICE+: Evaluation of automatic audio captioning systems with pre-trained language models'. In: *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2023)*. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2023). Rhodes Island, Greece, 4th June 2023. URL: https://inria.hal.science/hal-03933981.

[31] 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.

[32] Y. Laprie, V. Ribeiro, K. Isaeva, J. Leclere, J. Felblinger and P.-A. Vuissoz. 'Modeling the temporal evolution of the vocal tract shape with deep learning'. In: 20th International Congress on Phonetic Sciences. Prague (CZ), Czech Republic, 7th Aug. 2023. URL: https://inria.hal.science/hal-04209848.

[33] S. Leglaive, L. Borne, E. Tzinis, M. Sadeghi, M. Fraticelli, S. Wisdom, M. Pariente, D. Pressnitzer and J. R. Hershey. 'The CHiME-7 UDASE task: Unsupervised domain adaptation for conversational speech enhancement'. In: 7th International Workshop on Speech Processing in Everyday Environments (CHiME). Dublin, Ireland, 7th July 2023. URL: https://hal.science/hal-04156930.

[34] P. Magron and T. Virtanen. 'Spectrogram Inversion for Audio Source Separation via Consistency, Mixing, and Magnitude Constraints'. In: *Proc. European Signal Processing Conference (EUSIPCO)*. EUSIPCO 2023. Helsinki, Finland, 4th Sept. 2023. URL: https://hal.science/hal-04015991.

[35] I. Moummad, R. Serizel and N. Farrugia. 'Pretraining Representations for Bioacoustic Few-Shot Detection using Supervised Contrastive Learning'. In: Detection and Classification of Acoustic Scenes and Events 2023. TAMPERE, Finland, 2023. URL: https://imt.hal.science/hal-04383609.

[36] B. Nortier, M. Sadeghi and R. Serizel. 'Unsupervised speech enhancement with diffusion-based generative models'. In: International Conference on Acoustics Speech and Signal Processing (ICASSP). Seoul (Korea), South Korea, 14th Apr. 2024. URL: https://hal.science/hal-04210707.

[37] 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.

[38] S. Ogun, V. Colotte and E. Vincent. 'Stochastic Pitch Prediction Improves the Diversity and Naturalness of Speech in Glow-TTS'. In: InterSpeech 2023. Dublin, Ireland, 20th Aug. 2023. URL: https://hal.univ-lorraine.fr/hal-04108825.

[39] C. Oguz, P. Denis, E. Vincent, S. Ostermann and J. van Genabith. 'Find-2-Find: Multitask Learning for Anaphora Resolution and Object Localization'. In: 2023 Conference on Empirical Methods in Natural Language Processing. Singapore, Singapore, 2023. URL: https://hal.science/hal-04259861.

[40] R. S. Roman, Y. Adi, A. Deleforge, R. Serizel, G. Synnaeve and A. Défossez. 'From Discrete Tokens to High-Fidelity Audio Using Multi-Band Diffusion'. In: NeurIPS 2023 - Conference on Neural Information Processing Systems. New Orleans, United States, 2023. DOI: 10.48550/arXiv.2308.02560. URL: https://inria.hal.science/hal-04385071.

[41] M. Sadeghi and R. Serizel. 'Fast and efficient speech enhancement with variational autoencoders'. In: International Conference on Acoustics Speech and Signal Processing (ICASSP). Rhodes island, Greece, 4th June 2023. URL: https://inria.hal.science/hal-03833836.

[42] M. Sadeghi and R. Serizel. 'Posterior sampling algorithms for unsupervised speech enhancement with recurrent variational autoencoder'. In: International Conference on Acoustics Speech and Signal Processing (ICASSP). Seoul (Korea), South Korea, 14th Apr. 2024. URL: https://hal.science/hal-04210679.

[43] R. Serizel, S. Cornell and N. Turpault. 'Performance above all ? energy consumption vs. performance for machine listening, a study on dcase task 4 baseline'. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, France: IEEE; IEEE, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095938. URL: https://inria.hal.science/hal-03850797.

[44] P. Srivastava, A. Deleforge, A. Politis and E. Vincent. 'How to (Virtually) Train Your Speaker Localizer'. In: INTERSPEECH 2023. Dublin, Ireland, 22nd Aug. 2023. URL: https://hal.science/hal-03855912.

[45] N. Zampieri, I. Illina and D. Fohr. 'Improving Hate Speech Detection with Self-Attention Mechanism and Multi-Task Learning'. In: LTC'23 - 10th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. Poznan, Poland, 21st Apr. 2023. URL: https://hal.science/hal-04017250.

Conferences without proceedings

[46] C. Douwes, F. Ronchini and R. Serizel. 'Monitoring environmental impact of DCASE systems: Why and how ?' In: Detection and Classification of Acoustic Scene and Events (DCASE) Workshop. Tampere (Finlande), Finland, Sept. 2023. URL: https://hal.sorbonne-universite.fr/hal-04286654.

Doctoral dissertations and habilitation theses

[47] T. Bose. 'Transfer learning for abusive language detection'. Université de Lorraine, 30th Jan. 2023. URL: https://hal.univ-lorraine.fr/tel-04106135.

[48] P. Champion. 'Anonymizing Speech : Evaluating and Designing Speaker Anonymization Techniques'. Université de Lorraine, 20th Apr. 2023. URL: https://hal.univ-lorraine.fr/tel-04218098.

[49] S. Dowerah. 'Deep Learning-based Speaker Verification In Real Conditions'. Université de Lorraine; CNRS, Inria, LORIA, 30th May 2023. URL: https://hal.univ-lorraine.fr/tel-04257423.

[50] S. A. Sheikh. 'Deep learning for stuttering detection'. Université de Lorraine, 24th Feb. 2023. URL: https://hal.univ-lorraine.fr/tel-04073284.

[51] P. Srivastava. 'Realism in virtually supervised learning for acoustic room characterization and sound source localization'. Université de Lorraine, 13th Nov. 2023. URL: https://theses.hal.science/tel-04313405.

[52] G. Zervakis. 'Enriching large language models with semantic lexicons and analogies'. Université de Lorraine, 8th Mar. 2023. URL: https://theses.hal.science/tel-04138899.

Reports & preprints

[53] C. Cui, I. A. Sheikh, M. Sadeghi and E. Vincent. *End-to-end Joint Rich and Normalized ASR with a limited amount of rich training data*. 14th Sept. 2023. URL: https://inria.hal.science/hal-04304642.

[54] I. Moummad, R. Serizel and N. Farrugia. *Supervised contrastive learning for pre-training bioacoustic few-shot systems*. IMT Atlantique; LORIA, 31st May 2023. URL: https://imt-atlantique.hal.science/hal-04165306.

Other scientific publications

[55] C. Cui, I. A. Sheikh, M. Sadeghi and E. Vincent. 'End-to-end Multichannel Speaker-Attributed ASR: Speaker Guided Decoder and Input Feature Analysis'. In: Rencontre des Jeunes Chercheurs en Parole 2023 - 10E Edition. Grenoble, France, 29th Nov. 2023. URL: https://hal.science/hal-04321252.

12.3 Cited publications

[56] E. Vincent, R. Gribonval and C. Févotte. 'Performance measurement in blind audio source separation'. In: *IEEE Transactions on Audio, Speech and Language Processing* 14.4 (2006), pp. 1462–1469. URL: https://inria.hal.science/inria-00544230.