RESEARCH CENTRE

Inria Centre at Université de Lorraine

IN PARTNERSHIP WITH:

CNRS, Université de Lorraine

2024

ACTIVITY REPORT

Project-Team

MULTISPEECH

Multimodal Speech in Interaction

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

DOMAIN

Perception, Cognition and Interaction

THEME

Language, Speech and Audio

Inria

Contents

Project-Team MULTISPEECH

Creation of the Project-Team: 2024 October 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

- Yves Laprie [CNRS, Senior Researcher, HDR]
- Paul Magron [INRIA, Researcher]
- Mostafa Sadeghi [INRIA, ISFP]
- Natalia Tomashenko [INRIA, Advanced Research Position, from Feb 2024]
- Emmanuel Vincent [INRIA, Senior Researcher, HDR]

Faculty Members

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

Post-Doctoral Fellows

- Tom Bourgeade [UL, Post-Doctoral Fellow, from Sep 2024]
- Constance Douwes [INRIA, Post-Doctoral Fellow]

PhD Students

- Louis Abel [UL]
- Jean Eudes Ayilo [INRIA]
- Sofiane Azzouz [UL]
- Raphael Bagat [CNRS]
- Zahra Hafida Benslimane [CEA, from Oct 2024]
- Can Cui [INRIA, until Oct 2024]
- Stephane Dilungana [INRIA, until Sep 2024]
- Aine Drelingyte [UL, from Nov 2024]
- Orane Dufour [UL, from Oct 2024]
- Guilhem Faure [INRIA, from Oct 2024]
- Imed Eddine Ghebriout [CNRS, from Apr 2024]
- Mickaella Grondin-Verdon [UL, ATER, from Nov 2024]
- Mickaella Grondin-Verdon [UNIV MONTPELLIER, until Oct 2024]
- Taous Iatariene [ORANGE]
- Mayank Mishra [UL, from Dec 2024]
- Nasser-Eddine Monir [UL]
- Sewade Ogun [INRIA, until Oct 2024]
- Robin San Roman [Meta]

Technical Staff

- Theo Biasutto-Lervat [INRIA, Engineer]
- Sam Bigeard [INRIA, Engineer]
- Louis Delebecque [CNRS, Engineer, until Apr 2024]
- Malek Yaich [INRIA, Engineer, from Nov 2024]

Interns and Apprentices

- Aine Drelingyte [UL, Intern, from Mar 2024 until Aug 2024]
- Azim Fairooz [CNRS, Intern, from Jun 2024 until Sep 2024]
- Imane Ghanem [UL, Intern, from Jun 2024 until Sep 2024]
- Kira Grudinina [CNRS, Intern, from Jul 2024 until Sep 2024]
- Stéphane Loppinet [UL, Intern, from Feb 2024 until May 2024]
- Victor Menestrel [INRIA, Intern, from Jun 2024 until Aug 2024]
- Mayank Mishra [UL, Intern, from Apr 2024 until Aug 2024]
- Nathan Olejniczak [INRIA, Intern, from Apr 2024 until Jul 2024]
- Alexandre Perreux [CESI, Intern, from Sep 2024 until Nov 2024]
- Alexandre Perrot [INRIA, Intern, from Apr 2024 until Aug 2024]
- Panagiotis Tsolakis [INRIA, Intern, from Feb 2024 until Aug 2024]

Administrative Assistant

- Emmanuelle Deschamps [INRIA]

Visiting Scientists

- Ilyass Moummad [IMT ATLANTIQUE, until Mar 2024]
- Mohammad Hassan Vali [UNIV AALTO, from Sep 2024 until Nov 2024]

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.), this is the articulatory facet of speech. The shape of the vocal tract is partly visible on the face, this 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.

3.1.4 Axis 1.4 — Reducing computational footprint

This axis includes proposing reliable methods to quantify fine-grained energy consumption, computational footprint (in terms of operations), and memory footprint, so as to identify potential bottlenecks in

the network at training and test time before applying compression methods.

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 datasets, models and software for accented French and for regional, overseas and non-territorial languages of France.

The new Axis 1.4 “Reducing computational footprint” has been added to our research program this year.

Following previous years studies on the relation between energy consumption and performance, C. Douwes et al. studied the relationship between the energy consumed at training and during test for sound event detection systems [32]. They also proposed a study on the relationship between the energy consumption of different models architectures on several GPU models based on the models complexity and number of floating point operation [33].

6 New software, platforms, open data

6.1 New software

6.1.1 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

6.1.2 MRI

Name: Magnetic Resonance Imaging

Keywords: Health, Medical imaging

Functional Description: Magnetic Resonance Imaging (MRI) takes an increasing place in the investigation of speech production because it provides a complete geometrical information of the vocal tract. We thus initiated a cooperation with the IADI laboratory (Imagerie Adaptive Diagnostique et Interventionnelle) at Nancy Hospital, which studies in particular magnetic resonance imaging. This year, we acquired static MRI data for two speakers (approximately 90 blocked articulations corresponding to vowels and consonants followed by a vowel) and we carried out preliminary experiments intended to acquire dynamic data.

Contact: Yves Laprie

6.2 New platforms

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

6.2.1 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 is mainly driven by large technology companies (mainly American), raising questions about European digital sovereignty.

To address this problem, we are currently developing an open-source platform for the creation of virtual assistants, aiming to ease the integration of such technologies in any project. 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, and dialog management.

This year, a complete redesign of the platform has occured, due to the recent predominence of LLMs in human-machine interaction, natural language processing, and ultimetaly in voice-assitant technologies. These models and their high computing requirements make a fully embedded vision near impossible in the current state of technologies, hence a full rewrite of VAC as a client/server platform.

6.2.2 Record ECOS

We have developed a simple low-cost platform to record two persons during a face-to-face conversation.

On the hardware side, deploying one acquisition platform requires two webcams with tripods, a microphones stereo set with stands, an audio interface and a low-end laptop on Linux.

On the software side, we have delopped an engine based on `ffmpeg` to synchronously record from both cameras and the microphone stereo set, while associating the audio canals to their respective videos. The engine has a simple GUI to easily start and stop recording, as well as encrypt the data on the fly. We also developped several post-processing scripts to remove the speech signal residues of speaker A from the recording of speaker B, to generate the transcription for a speaker, and to merge the two individual transcriptions into a dialog transcription

7 New results

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

Participants: Vincent Colotte, Irina Illina, Paul Magron, Mostafa Sadeghi, Romain Serizel, Emmanuel Vincent, Jean Eudes Ayilo, Sam Bigeard, Constance Douwes, Orane Dufour, Mohamed Imed Eddine Ghebriout, Sewade Olaolu Ogun, Robin San Roman, Natalia Tomashenko, Malek Yaich.

7.1.1 Axis 1.1 — Integrating domain knowledge

Integration of symbolic knowledge. We pursued our study on linguistic ambiguities arising from changes in entities in videos, focusing on instructional cooking videos as a challenging use case. We released MMAR, a multilingual version of the multimodal anaphora resolution released last year, and presented updated experimental results of a novel end-to-end joint multitask learning framework for these two tasks [24].

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 [22] 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 [16], which bridges the gap between the performance of supervised and unsupervised speech enhancement approaches.

7.1.2 Axis 1.2 - Learning from little/no labeled data

ASR for regional languages. We started collecting speech data covering some of the regional, overseas, and non-territorial languages of France from data archives, media, associations, and individuals. This data will support the training of automatic speech recognition (ASR) and text-to-speech (TTS) models for these languages. A first effort in this direction was made for Alsatian ASR by fine-tuning Whisper [30]. The results enabled us to posit that for ASR fine-tuning a few hours of speech data are required but the amount and variety of text data are key.

Learning from noisy data. Training of multi-speaker 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 ASR datasets using a non-intrusive perceptual mean opinion score estimator. We validated this method for English TTS with African accents [23]. We also used the unaccented English TTS system developed last year for model-driven data augmentation, and demonstrated a WER reduction up to 35% relative when training ASR systems on the augmented data. Sewade Ogun defended his PhD on this topic [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 [13].

7.1.3 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. In particular, we measured and reduced the amount of speaker information carried by phoneme durations [28].

We published a detailed analysis of the results of the 2nd Voice Privacy Challenge which we co-organized in 2022 [9]. We organized the 3rd Voice Privacy Challenge [41], which allows the use of external datasets and pretrained models. The challenge evaluates utility using ASR and emotion classification models trained on original (unprocessed) data to ensure that linguistic and emotional content remains undistorted. Additionally, it significantly simplifies the evaluation protocol and reduces the running time of the evaluation scripts, which are now primarily in python.

We organized the 1st VoicePrivacy Attacker Challenge [42], which focuses on developing speaker re-identification attacks against three baseline anonymization systems and four anonymization systems developed by the Voice Privacy 2024 Challenge participants. The best attacker systems reduced the equal error rate (EER) by 25–44% relative w.r.t. the semi-informed attack used in the VoicePrivacy 2024 Challenge.

7.1.4 Axis 1.4 — Reducing computational footprint

In order to be able to compare the energy footprint of a system trained over a different site, we have been studying the relationship between the energy consumption of different model architectures on several

GPU models based on the model's complexity and number of floating point operations [33]. Some of the outcome of this study has been used within the DCASE challenge in order to normalize the energy consumption of the systems submitted by the participants. This allowed to study the evolution of the energy consumption of state-of-the-art sound event systems over the past years and its relation to the performance [26, 38]. During her PhD thesis work, Zahra Benslimane studied in detail the computational load in terms of floating point operations of three different multichannel speech enhancement systems. She then proposed an analysis of the relationship between algorithm latency, computational footprint and performance in terms of speech enhancement. The end goal is to identify the bottleneck of the current system in order to convert them to low latency and low complexity speech enhancement algorithms targeting implementation in hearing aids.

7.2 Axis 2 — Extracting information from speech signals

Participants: Anne Bonneau, Irina Illina, Paul Magron, Romain Serizel, Emmanuel Vincent, Raphaël Bagat, Can Cui, Louis Delebecque, Stéphane Dilungana, Aine Drelingyte, Taous Iatariene, Mayank Mishra, Nasser-Eddine Monir.

7.2.1 Axis 2.1 — Linguistic speech content

Speaker-attributed ASR. We improved speaker assignment in speaker-attributed ASR (SA-ASR) systems [17]. First, we proposed a pipeline tailored to real-life applications involving voice activity detection (VAD), speaker diarization (SD), and SA-ASR. Second, we advocated using VAD output segments to finetune the SA-ASR model, and showed a relative reduction of Speaker Error Rate (SER) up to 28%. Finally, we showed that extracting the reference speaker embeddings used as inputs by the SA-ASR system from SD output rather than annotated speaker segments results in a relative SER reduction up to 16%. Can Cui has defended her PhD on this topic [36].

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. We work on the training strategy, which allows flexible modeling of the relative proximity of neighbors retrieved from the resource-rich corpus to learn the amount of transferts.

Large language model compression. Current LLM compression approaches typically involve two steps: (1) compressing using calibration data and (2) continued pretraining on billions of tokens to recover lost performance. This costly second step is only necessary when the first step severely impacts performance. Based on the observation that activations are low-rank, we introduced a new oneshot compression method that locally distills low-rank weights. We accelerated convergence by initializing the low-rank weights with SVD and using a joint loss that combines teacher and student activations, and reduce memory requirements by applying local gradient updates only. We have shown that our approach can compress Mixtral-8x7B within minutes on a single A100 GPU, removing 10 billion parameters while maintaining over 95Phi-2 3B can be compressed by 40% using only 13 million calibration tokens into a small model that competes with recent models of similar size. Our approach further works on non-transformer architectures.

7.2.2 Axis 2.2 — Speaker identity and states

Speaker recognition. Following our previous work on far-field speaker recognition, we introduced a new far-field speaker recognition benchmark called RoboVox recorded by a mobile robot [34]. We

also proposed a multi-channel speaker embedding models system based on a pre-trained single channel model for each channel and cross-channel information exchange, eventually fusing channels into one [20].

7.2.3 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) 2024. This evaluation method can provide a more complete picture of the system's 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 compared energy consumption of the challenge submissions and the relation for their sound event detection performance [26, 38].

Speech enhancement. Targeting speech enhancement for hearing aids, we started investigating 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). To that end, we conducted an extensive evaluation of state-of-the-art speech enhancement algorithms at the phoneme level, rather than at the commonly-considered utterance level. These investigations demonstrate important improvements in phoneme clarity in noisy conditions, with insights that could drive the development of more personalized and phoneme-aware hearing aid technologies [8].

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 [27]. Our experiments demonstrated the effectiveness of the proposed methods, narrowing the performance gap between supervised and unsupervised approaches in speech enhancement.

7.3 Axis 3 — Multimodal Speech: generation and interaction

Participants: Théo Biasutto-Lervat, Domitille Caillat, Vincent Colotte, Yves Laprie, Slim Ouni, Mostafa Sadeghi, Emmanuel Vincent, Louis Abel, Sofiane Azzouz, Tom Bourgeade, Guilhem Faure, Mickaella Grondin-Verdon.

7.3.1 Axis 3.1 — Multimodality modeling and analysis

Unsupervised performance analysis of face frontalization for audio-visual speech processing. In a related work [12], 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.

7.3.2 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 continued the acquisition of 3 stuttering subjects and 2 reference subjects, and began the acquisition of a German corpus. As with the corpus recorded for French, we have chosen a large corpus (around 1700 sentences), which has the advantage of making deep learning possible. These recordings have begun and will continue in early 2025.

Exploitation of RT-MRI data recorded for beatboxing. MRI data on beatboxing are interesting for studying speech production because beatboxing involves production mechanisms that differ from those involved in French, but which are also found in click languages. A particular feature of beatboxing is the ability to use several constrictor muscles together to increase, resp. decrease, the pressure inside the pharynx in order to produce ejective, resp. injective, consonants [31].

Acoustic to articulatory inversion. Acoustic to articulatory inversion is a major processing challenge, with a wide range of applications from speech synthesis to feedback systems for language learning and rehabilitation. We used the large corpus that served as the basis for the vocal tract shape generation work Vinicius Ribeiro had developed in his thesis [44] to train an approach to articulatory acoustic inversion. The data are, on the one hand, articulator contours obtained by automatic tracking and, on the other, the denoised speech signal. This year we focused on the tongue, the most important and mobile articulator. Several architectures relying on a Bi-LSTM including or not an autoencoder to reduce the dimensionality of the latent space, using or not the phonetic segmentation have been explored. The results show that the tongue contour can be recovered with a median accuracy of 2.21 mm (or 1.37 pixel) taking a context of 1 MFCC frame (static, delta and double-delta cepstral features). Unlike other inversion approaches that use EMA data limited to the oral cavity, our approach covers the entire vocal tract.

Sign language. Sign languages are rich visual languages with unique linguistic and grammatical structures that do not directly map to spoken languages. Computational research in this area is limited by small-scale datasets focused on narrow domains (e.g., weather forecasts) and issues like low inter- and intra-signer variation, limited vocabulary, and poor visual quality. To address these gaps, we collected and processed over 300 hours of signing News video footage from a German broadcaster. The data includes skeletal features for the face, hands, and body, along with textual transcriptions. We conducted comprehensive analyses, including signer labeling, outlier detection, undersigning quality assessment, and landmark error rate measurement. Additionally, we proposed a multimodal Transformer-based framework to annotate the corpus using mDGS dataset glossaries.

Furthermore, current sign language generation research faces two key challenges: reliance on gloss annotations, which act as a bottleneck and omit critical text or speech information, and poor alignment between text or speech and sign videos. To address these, we have started a PhD project aiming to explore gloss-free approaches and develop alternative representations that preserve more linguistic information while improving performance, among other objectives. We also plan to develop an efficient framework for automatic alignment of text or speech with sign videos, enabling training of advanced neural models and more natural sign language generation.

Evaluation of multimodal speech generation in language therapy support. As part of our work on multimodal speech generation, we collaborated with INSEI and the Corsica Regional Health Agency (ARS) to evaluate talking avatars with children and young people, both with and without speech and language disorders. The aim was to explore the potential integration of this technology into home settings as a complement to speech therapy. Our findings highlight the importance of lip-reading in oral sentence comprehension for deaf children, aligning with previous research in the field. The overall articulation quality of the avatars was deemed very well but also depends on the precision of the controls applied to the avatars [10, 35].

7.3.3 Axis 3.3 — Interaction

Co-speech 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.

This year, our research has focused on advancing co-speech gesture synthesis and enhancing methodologies for gesture data management. A key contribution is the development of STARGATE, a novel model for synthesizing gestures from audio-text embeddings [29, 15]. STARGATE overcomes a common limitation in existing systems: the use of complex and resource-intensive architectures. By using autoregression for fast gesture generation and incorporating graph convolutions along with attention mechanisms, STARGATE efficiently handles spatial and temporal aspects of gestures. This makes it suitable for integration into Embodied Conversational Agents (ECAs) and supports linguistic research, where understanding the relationship between speech and gestures has traditionally been difficult. Both subjective and objective evaluations show that STARGATE produces highly credible and coherent gestures, matching or slightly surpassing state-of-the-art models in gesture realism. We also demonstrated that our model is capable of generating convincing gestures and conducted an in-depth analysis to show how the model produces gestures from its input [14].

We have examined data management and corpus enrichment practices in gesture research, focusing on the challenges of gesture annotation, including segmentation, labeling, and identifying lexical affiliates. Our analysis revealed discrepancies in gesture data due to inconsistencies in these annotation processes. Based on these findings, we proposed strategies to improve annotation practices, enhance methodological transparency, and increase the reliability of enriched corpora. These strategies aim to refine the management and interpretation of gesture datasets, ensuring their utility for detailed analysis and promoting a standardized approach to annotation across the field [19].

8 Bilateral contracts and grants with industry

8.1 Bilateral grants with industry

8.1.1 Vivoka

- Company: Vivoka (France)
- Duration: Oct 2021 – Apr 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.

8.1.2 Meta AI

- Company: Meta AI (France)
- Duration: May 2022 – Apr 2025
- Participants: Robin San Roman, 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.

8.1.3 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.

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Participation in other International Programs

ANR-JST CONFLUENCE

Title: Semantic Segmentation of CompLexe Sound Scenes on Edge Devices

Duration: Nov 2023 - Nov 2027

Coordinator: Nicolas Turpault, Sonaide

Partners:
- Sonaide (France)
- CEA List (France)
- LORIA (France)
- Nippon Telegraph and Telephone (Japan)
- Tokyo Metropolitan University (Japan)

Participants: Paul Magron, Mayank Mishra, Romain Serizel

Abstract: The CONFLUENCE project aims to propose audio semantic segmentation algorithms for use on embedded systems (home support and 3GPP codec).

9.2 International research visitors

9.2.1 Visits of international scientists

Other international visits to the team

Mohammad Hassan Vali

Status PhD

Institution of origin: Aalto University

Country: Finland

Dates: Sep 9 until Nov 29

Context of the visit: collaboration starting soon with his supervisor Tom Backström in the MSCA PSST project

Mobility program/type of mobility: internship

9.3 European initiatives

9.3.1 H2020 projects

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 (Italy)
- Commpla (Italy)
- 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 that aims to organize cross-community workshops.

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.

9.4 National initiatives

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 - Dec 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 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, Paul Magron, Tom Sprunck

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.

ANR Full3DTalkingHead

Title: Synthèse articulatoire phonétique

Duration: Apr 2021 - Mar 2025

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 school 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-fundings 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 ReNAR

Title: Reducing Noise with Augmented Reality

Duration: Feb 2024 - Jan 2028

Coordinator: CEA List (Saclay)

Partners: Ircam (Paris), Laboratoire des Sciences du Numérique de Nantes (Nantes), LORIA (Nancy)

Participants: Romain Serizel, Aine Drelingyte

Abstract: The aim of the ReNAR project is to design a solution that can attenaute the impact of noise in office working scenarios (in particular in open spaces). We will target two aspects: generating noise maskers that results in sound scenes that are pleasent to hear for workers and generating signals that can obfuscate surrounding speech.

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.

LANGU:IA

Title: Pôle de référence transversal et pluridisciplinaire du Traitement automatique du Français et des Langues de France

Duration: Jan 2024 - Sep 2024

Coordinator: Communauté de communes de Retz-en-Valois

Partners: Communauté de communes de Retz-en-Valois, APIL, ATALA, CMN – Cité internationale de la langue française, Ministère de la Culture – Délégation générale à la langue française et aux langues de France, ELDA, LISN (Saclay), Le VoiceLab

Participants: Emmanuel Vincent

Abstract: LANGU:IA aims to create a national cluster for language technologies that will bring together all stakeholders (industrial, academic, institutional) in order to develop technologies and innovation for the automatic processing of French and the languages of France.

ANR SPEECHPRIVACY

Title: Multiple-attribute disentanglement and semantic privacy

Duration: Feb 2024 - Jan 2028

Coordinator: Vincent Colotte

Partners: LORIA (Nancy), EURECOM (Sophia Antipolis), LIA (Avignon)

Participants: Vincent Colotte, Emmanuel Vincent

Abstract: SpeechPrivacy will deliver a flexible solution to privacy preservation based on isolated/disentangled representations and the selective obfuscation/modification of individual attributes beyond the usual voice identity/sex and sensitive keywords.

ANSES IPIAMA

Title: Reducing Noise with Augmented Reality

Duration: Dec 2023 - Dec 2026

Coordinator: Jean-Pierre Arz, INRS (Nancy)

Partners: INRS (Nancy), Laboratoire Énergies et Mécanique Théorique et Appliquée (Nancy), LORIA (Nancy)

Participants: Romain Serizel

Abstract: The IPIAMA project aims to propose binaural speech intelligibility measurements (with both ears) for people equipped with hearing aids. The project will rely jointly on classic listening tests (reliable but expensive) and models based on data collected in realistic conditions.

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair

- Co-chair, 4th Inria-DFKI European Summer School on Artificial Intelligence, Saarbrücken, Sep 2024 (E. Vincent)

- Main organizer, UDICE-U15 workshop on AI: Stronger together – How to train and retain the next generation of talent in Europe and develop efficient and competitive French-German ecosystems?, Nancy, Mar 2025 (E. Vincent)

- Co-chair, DCASE Challenge 2024 (R. Serizel)

Member of the organizing committees

- Organizer, 3rd VoicePrivacy Challenge (N. Tomashenko, E. Vincent)

- Organizer, 1st VoicePrivacy Attacker Challenge (N. Tomashenko, E. Vincent)

10.1.2 Scientific events: selection

Chair of conference program committees

- Area chair, Interspeech 2025 (E. Vincent)

Member of the conference program committees

- JEP-TALN-RECITAL 2024 (V. Colotte, Y. Laprie, S. Ouni)

- Member of the external committee of the special issue of the TAL journal on hate speech (I. Illina)

Reviewer

- ICASSP 2025 – IEEE International Conference on Acoustics, Speech, and Signal Processing (N. Tomashenko, E. Vincent, I. Illina, R. Serizel)

- INTERSPEECH 2024 (E. Vincent, I. Illina, V. Colotte, R. Serizel)

- EUSIPCO 2024 (V. Colotte, R. Serizel)

- SPSC 2024 – 4th Symposium on Security and Privacy in Speech Communication (N. Tomashenko, E. Vincent)

- SLT 2024 - Spoken Language Technology (I. Illina)

- ARR ACL Rolling Review several editions 2024 (I. Illina)

- Journal Speech Communication 2024 (I. Illina)

- Journal IEEE Transactions on Speech and Audio Processing 2024 (I. Illina)

- Journal Traitement Automatique des Langues, spec. issue Discours de haine (I. Illina)

10.1.3 Invited talks

- Keynote "Observer l'humain pour générer un avatar multimodal", WACAI 2024, Bordeaux, Jun 2024 (S. Ouni)

- Invited speaker at Computational Analysis and Simulation of the Human Voice Dagstuhl Seminar, Jun 2024 (Y. Laprie)

10.1.4 Leadership within the scientific community

- Vice-president of AFCP - Association Francophone de la Communication Parlée (S. Ouni)
- Secretary/Treasurer, executive member of AVISA (Auditory-VIsual Speech Association), an ISCA Special Interest Group (S. Ouni)
- 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)
- Chair of the Steering Committe of DCASE (R. Serizel)

10.1.5 Scientific expertise

- Expert for the Data Governance Working Group of the Global Partnership on AI (GPAI) (E. Vincent)
- Expertise for the Czech Science Foundation (I. Illina)
- Expertise for the CIFRE (I. Illina)
- Expertise for the « Collaborative Research in Computational Neuroscience » (CRCNS), NSF-ANR (R. Serizel)

10.1.6 Research administration

- Coordinator of the AI Cluster ENACT (European Centre for Artificial Intelligence through Innovation, E. Vincent)
- Head of Science of the Inria Research Center at Université de Lorraine (E. Vincent, until Feb)
- Head of pole scientifique Automatique, Mathématiques, Informatique et leurs interactions (AM2I) (Y. Laprie)
- Member of the selection committee for the position of assistant professor at Université de Strasbourg (S. Ouni)
- Member of the selection committee for the position of assistant professor at Université de Paris-Saclay (S. Ouni)
- Member of the selection committee for the position of assistant professor at IUT Lannion (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)
- Inria representative on the Lorraine Steering Committee for Open Science (E. Vincent)
- Vice-chair of the hiring committee for Junior Research Scientists, Inria Research Center at Université de Lorraine (E. Vincent)
- Member of Comité Espace Transfert, Inria Research Center at Université de Lorraine (E. Vincent)
- Member of Inria's Evaluation Committee (E. Vincent, until Feb)
- Member of Conseil de Laboratoie du LORIA (V. Colotte)
- Member of the bureau du pole scientifique Automatique, Mathématiques, Informatique et leurs interactions (AM2I) (I. Illina)
- Member of the Comité du pole scientifique Automatique, Mathématiques, Informatique et leurs interactions (AM2I) (I. Illina)

- Member of the board, Université de Lorraine (Y. Laprie)
- Member of the Comité Utilisateurs des Moyens de Calculs, Inria Research Center at Université de Lorraine (T. Biasutto–Lervat)
- Referent Plateformes-Outils, Inria Research Center at Université de Lorraine (T. Biasutto–Lervat)

10.2 Teaching - Supervision - Juries

10.2.1 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: 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
- Licence: V. Colotte, Introduction to speech processing (20 hours), 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: E. Vincent, P. Magron, and N. Tomashenko, Speech recognition and synthesis (20 hours), M2, Université de Lorraine
- Master: T. Biasutto-Lervat, Deep learning (labs, 16 hours), M2, Université de Lorraine
- Master: E. Vincent and P. Magron, Neural networks (32 hours + 11 hours), M2, Université de Lorraine
- 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

10.2.2 Supervision

- PhD: Can Cui, “Joint speech separation, diarization and recognition for automatic meeting transcription”, Oct 2024, M. Sadeghi and E. Vincent [36].
- PhD: Sewade Olaolu Ogun, “Generating diverse synthetic data for ASR training data augmentation”, Oct 2024, V. Colotte and E. Vincent [37].
- PhD in progress: Louis Abel, "Expressive audio-visual speech synthesis in an interaction context", Oct 2021, V. Colotte and S. Ouni.
- PhD in progress: Mickaëlla Grondin, "Modeling gestures and speech in interactions", Nov 2021, S. Ouni and F. Hirsch (Praxiling).
- 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: Yaya Sy "Efficient Continued pre-training of Large Language Models", Oct 2023, I. Illina and C. Cerisara.
- PhD in progress: Jean-Eudes Ayilo "Audio-visual Speech Enhancement: Bridging the Gap between Supervised and Unsupervised Approaches", Oct 2023, M. Sadeghi and R. Serizel.
- PhD in progress: Mohamed Imed Eddine Ghebriout, “LLM adaptation and exploitation for medical emergency call triage”, Apr 2024, G. Guibon (LIPN) and E. Vincent.
- PhD in progress: Orane Dufour, “Towards a comprehensive speech anonymization framework”, Oct 2024, M. Rouvier (LIA) and E. Vincent.
- PhD in progress: Guilhem Faure, "End-to-end speech-to-sign language generation", Oct 2024, M. Sadeghi and S. Ouni.

10.2.3 Juries

- Participation in the HDR jury of Lina Rojas (Université de Lorraine, Jun 2024), S. Ouni, examiner
- Participation in the PhD jury of Théo Mariotte (University of Le Mans, Jan 2024), E. Vincent, reviewer
- Participation in the PhD jury of Samir Sadok (CentraleSupélec Rennes, Mar 2024), S. Ouni, reviewer
- Participation in the PhD jury of Salah Zaiem (Télécom ParisTech, Mar 2024), E. Vincent, chair
- Participation in the PhD jury of Francisco Teixeira (University of Lisbon, Jul 2024), E. Vincent, reviewer
- Participation in the PhD jury of Sanjana Sankar (Université Grenoble-Alpes, Sep 2024), S. Ouni, reviewer
- Participation in the PhD jury of Hamza Bayd (CentraleSupélec Rennes, Dec 2024), S. Ouni, chair
- Participation in the PhD jury of Thomas Lebrun (INSA de Lyon, Dec 2024), E. Vincent, reviewer
- Participation in the PhD jury of Léa-Marie Lam-Yee-Mui (Université Paris Saclay, Nov 2024), I. Illina, reviewer
- Participation in the PhD jury of Etienne Labbé (Université Paul Sabatier, Feb 2024), R. Serizel, reviewer
- Participation in the PhD jury of Philippe Gonzales (Technical University of Denmark, Apr 2024), R. Serizel, reviewer
- Participation in the PhD jury of Kevin Wilkinghoff (Universität Bonn, May 2024), R. Serizel, reviewer

10.3 Popularization

10.3.1 Participation in Live events

- Panel discussion at Viva Technology in Paris, Jun 2024 (S. Ouni)
- Panel discussion at the VoiceLab event on Ethics of AI systems for speech and language processing and generation, Jul 2024 (E. Vincent)
- Talk on automatic lipsync technology at Rencontres Animation Développement Innovation (RADI) in Angoulême, Nov 2024 (S. Ouni)

11 Scientific production

11.1 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] N.-E. E. Monir, P. Magron and R. Serizel. 'A Phoneme-Scale Assessment of Multichannel Speech Enhancement Algorithms'. In: *Trends in Hearing* 28 (12th Dec. 2024). DOI: 10.1177/23312165241292205. URL: https://hal.science/hal-04854449.

[3] 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.

[4] 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.

[5] 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.

11.2 Publications of the year

International journals

[6] F. Effa, J.-P. Arz, R. Serizel and N. Grimault. 'Evaluating and predicting the audibility of acoustic alarms in the workplace using experimental methods and deep learning'. In: *Applied Acoustics* 219 (Mar. 2024), p. 109955. DOI: 10.1016/j.apacoust.2024.109955. URL: https://hal.science/hal-04645900.

[7] S. Leglaive, M. Fraticelli, H. ElGhazaly, L. Borne, M. Sadeghi, S. Wisdom, M. Pariente, J. R. Hershey, D. Pressnitzer and J. P. Barker. 'Objective and subjective evaluation of speech enhancement methods in the UDASE task of the 7th CHiME challenge'. In: *Computer Speech and Language* 89 (Jan. 2025). DOI: 10.1016/j.csl.2024.101685. URL: https://hal.science/hal-04430786.

[8] N.-E. E. Monir, P. Magron and R. Serizel. 'A Phoneme-Scale Assessment of Multichannel Speech Enhancement Algorithms'. In: *Trends in Hearing* 28 (12th Dec. 2024). DOI: 10.1177/23312165241292205. URL: https://hal.science/hal-04854449 (cit. on p. 12).

[9] M. Panariello, N. Tomashenko, X. Wang, X. Miao, P. Champion, H. Nourtel, M. Todisco, N. Evans, E. Vincent and J. Yamagishi. 'The VoicePrivacy 2022 Challenge: Progress and perspectives in voice anonymisation'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* (2024). DOI: 10.1109/TASLP.2024.3430530. URL: https://inria.hal.science/hal-04667625. In press (cit. on p. 10).

[10] A. Piquard-Kipffer, K. Martinelli, L. Dussere, A. Sancier, J. Zytnicki, C. Barbot-Bouzit and S. Ouni. 'AVI-Corse : méthodologie et enjeux d'un projet participatif. Des avatars numériques au service du langage et de la communication.: AVI-Corse: methodology and challenges of a participatory project. Digital avatars, new tools for language and communication needs.' In: *La Nouvelle revue – Éducation et société inclusives* 98 (2024), pp. 237–250. URL: https://hal.science/hal-04314092 (cit. on p. 13).

[11] 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), p. 107907. DOI: 10.1016/j.cmpb.2023.107907. URL: https://inria.hal.science/hal-04376938. In press.

[12] 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), pp. 1–16. DOI: 10.1016/j.neucom.2023.126941. URL: https://hal.science/hal-04265797 (cit. on p. 12).

[13] 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 (cit. on p. 10).

International peer-reviewed conferences

[14] L. Abel, V. Colotte and S. Ouni. 'Towards interpretable co-speech gestures synthesis using STARGATE'. In: International Conference on Multimodal Interaction (ICMI Companion '24: GENEA Workshop). San José, Costa Rica, 5th Nov. 2024. DOI: 10.1145/3686215.3688819. URL: https://hal.science/hal-04678537 (cit. on p. 14).

[15] L. Abel, V. Colotte and S. Ouni. 'Towards realtime co-speech gestures synthesis using STARGATE'. In: 25th Interspeech Conference (INTERSPEECH 2024). Kos Island, Greece, 1st Sept. 2024. URL: https://hal.science/hal-04667107 (cit. on p. 14).

[16] 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. DOI: 10.48550/arXiv.2309.10457. URL: https://hal.science/hal-04210729 (cit. on p. 10).

[17] C. Cui, I. A. Sheikh, M. Sadeghi and E. Vincent. 'Improving Speaker Assignment in Speaker-Attributed ASR for Real Meeting Applications'. In: The Speaker and Language Recognition Workshop Odyssey 2024. Quebec, Canada, 18th June 2024. URL: https://hal.science/hal-04495886 (cit. on p. 11).

[18] A. Golmakani, M. Sadeghi, X. Alameda-Pineda and R. Serizel. 'A weighted-variance variational autoencoder model for speech enhancement'. In: ICASSP 2024 - International Conference on Acoustics Speech and Signal Processing. Seoul (Korea), South Korea, 2024, pp. 1–5. URL: https://inria.hal.science/hal-03833827.

[19] M. Grondin-Verdon, D. Caillat and S. Ouni. 'Qualitative study of gesture annotation corpus : Challenges and perspectives'. In: ICMI Companion '24: Companion Proceedings of the 26th International Conference on Multimodal Interaction. San Jose, Costa Rica: ACM, 2024, pp. 147–155. DOI: 10.1145/3686215.3688820. URL: https://hal.science/hal-04767088 (cit. on p. 14).

[20] L. Mošner, R. Serizel, L. Burget, O. Plchot, E. Vincent, J. Peng and J. Černocký. 'Multi-channel extension of pre-trained models for speaker verification'. In: Interspeech. Kos, Greece, 1st Sept. 2024. URL: https://inria.hal.science/hal-04667593 (cit. on p. 12).

[21] I. Moummad, N. Farrugia, R. Serizel, J. Froidevaux and V. Lostanlen. 'Mixture of Mixups for Multi-label Classification of Rare Anuran Sounds'. In: EUSIPCO 2024. Lyon, France, 26th Aug. 2024. URL: https://imt.hal.science/hal-04620733.

[22] 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. DOI: 10.48550/arXiv.2309.10450. URL: https://hal.science/hal-04210707 (cit. on p. 10).

[23] S. Ogun, A. T. Owodunni, T. Olatunji, E. Alese, B. Oladimeji, T. Afonja, K. Olaleye, N. A. Etori and T. Adewumi. '1000 African Voices: Advancing inclusive multi-speaker multi-accent speech synthesis'. In: Interspeech 2024. Kos Island, Greece, 1st Sept. 2024. URL: https://hal.science/hal-04663033 (cit. on p. 10).

[24] C. Oguz, P. Denis, S. Ostermann, N. Skachkova, E. Vincent and J. van Genabith. 'MMAR: Multilingual and multimodal anaphora resolution in instructional videos'. In: Findings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, United States, 12th Nov. 2024. URL: https://inria.hal.science/hal-04733760 (cit. on p. 9).

[25] R. S. Roman, P. Fernandez, H. Elsahar, A. Défossez, T. Furon and T. Tran. 'Proactive Detection of Voice Cloning with Localized Watermarking'. In: *Proceedings of the 41st International Conference on Machine Learning*. ICML 2024 - 41st International Conference on Machine Learning. Vol. 235. Vienna, Austria, July 2024, pp. 1–17. URL: https://hal.science/hal-04610152.

[26] F. Ronchini and R. Serizel. 'Performance and Energy Balance: A Comprehensive Study of State-of-the-Art Sound Event Detection Systems'. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, South Korea: IEEE, 14th Apr. 2024, pp. 1096–1100. DOI: 10.1109/ICASSP48485.2024.10445834. URL: https://inria.hal.science/hal-04892368 (cit. on pp. 11, 12).

[27] 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. DOI: 10.48550/arXiv.2309.10439. URL: https://hal.science/hal-04210679 (cit. on p. 12).

[28] N. Tomashenko, E. Vincent and M. Tommasi. 'Analysis of Speech Temporal Dynamics in the Context of Speaker Verification and Voice Anonymization'. In: 2025 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2025). Hyderabad, India, 6th Apr. 2025. URL: https://hal.science/hal-04853872 (cit. on p. 10).

National peer-reviewed Conferences

[29] L. Abel, V. Colotte and S. Ouni. 'Synthèse de gestes communicatifs via STARGATE'. In: *35èmes Journées d'Études sur la Parole (JEP 2024)*. 35èmes Journées d'Études sur la Parole (JEP 2024) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024) 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024). Toulouse, France: ATALA & AFPC, 2024, pp. 181–190. URL: https://inria.hal.science/hal-04623071 (cit. on p. 14).

Conferences without proceedings

[30] S. Bigeard, P. Tsolakis, E. Vincent, V. Colotte, P. Erhart and S. Ouni. 'Retour d'expérience : Whisper pour les langues régionales'. In: LIFT 2: Journées scientifiques du GdR Linguistique Informatique, Formelle et de Terrain. Orléans, France, 14th Nov. 2024. URL: https://hal.science/hal-04787239 (cit. on p. 10).

[31] A. Dehais-Underdown, L. Crevier-Buchman, D. Demolin, P.-A. Vuissoz, M. Fauvel, J. Felblinger and Y. Laprie. 'Are glottalic mechanisms in Human Beatboxing really glottalic ?' In: 13th International Seminar of Speech Production. Autrans, France, 17th May 2024. URL: https://hal.science/hal-04579667 (cit. on p. 13).

[32] C. Douwes and R. Serizel. 'From Computation to Consumption: Exploring the Compute-Energy Link for Training and Testing Neural Networks for SED Systems'. In: Detection and Classification of Acoustic Scenes and Events 2024. Tokyo, Japan, 23rd Oct. 2024. URL: https://hal.science/hal-04697137 (cit. on p. 8).

[33] C. Douwes and R. Serizel. 'Normalizing Energy Consumption for Hardware-Independent Evaluation'. In: 2024 IEEE International Workshop on Machine Learning for Signal Processing. London, United Kingdom, 22nd Sept. 2024. URL: https://hal.science/hal-04697122 (cit. on pp. 8, 11).

[34] M. Mohammadamini, D. Matrouf, M. Rouvier, J.-F. Bonastre, R. Serizel and T. Gonos. 'RoboVox: A Single/Multi-channel Far-field Speaker Recognition Benchmark for a Mobile Robot'. In: LREC_COLING. Turino, Italy, 8th Apr. 2024. URL: https://hal.science/hal-04536499 (cit. on p. 11).

[35] A. Piquard-Kipffer, A. Krilanovic, J. Zytnicki, K. Martinelli, L. Dussere, A. Sancier and S. Ouni. 'Assessment of avatar lip-reading technology (AVI-Corse project). Perspectives of young people with and without hearing loss'. In: 36th WCA 2024, World Congress of Audiology. Paris (CNIT - La Défense), France, 19th Sept. 2024. URL: https://hal.science/hal-04839280 (cit. on p. 13).

Doctoral dissertations and habilitation theses

[36] C. Cui. 'Joint speech separation, diarization, and recognition for automatic meeting transcription'. Université de Lorraine, 1st Oct. 2024. URL: https://hal.univ-lorraine.fr/tel-04813660 (cit. on pp. 11, 24).

[37] S. Ogun. 'Generating diverse synthetic data for ASR training data augmentation'. Université de Lorraine, 10th Oct. 2024. URL: https://hal.univ-lorraine.fr/tel-04847952 (cit. on pp. 10, 24).

Reports & preprints

[38] C. Douwes and R. Serizel. *Energy Consumption Trends in Sound Event Detection Systems*. 13th Sept. 2024. URL: https://hal.science/hal-04697014 (cit. on pp. 11, 12).

[39] I. Moummad, R. Serizel, E. Benetos and N. Farrugia. *Domain-Invariant Representation Learning of Bird Sounds*. 13th Sept. 2024. URL: https://hal.science/hal-04696391.

[40] R. S. Roman, P. Fernandez, A. Deleforge, Y. Adi and R. Serizel. *Latent Watermarking of Audio Generative Models*. 2024. DOI: 10.48550/arXiv.2409.02915. URL: https://hal.science/hal-04716743.

[41] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi and M. Todisco. *The Voice Privacy 2024 Challenge Evaluation Plan*. Inria; Eurecom; NII, 8th Mar. 2024. URL: https://inria.hal.science/hal-04531444 (cit. on p. 10).

[42] N. Tomashenko, X. Miao, E. Vincent and J. Yamagishi. *The First VoicePrivacy Attacker Challenge Evaluation Plan*. 10th Oct. 2024. URL: https://hal.science/hal-04730990 (cit. on p. 10).

Scientific popularization

[43] G. Coiffier, S. Ogun, L. Valque and P. Trivedi. 'STATE OF THE ART'. In: *THINK BEFORE LOADING*. 2024. URL: https://hal.univ-lorraine.fr/hal-04509255.

11.3 Cited publications

[44] V. Ribeiro. 'Deep Supervision of the Vocal Tract Shape for Articulatory Synthesis of Speech'. Theses. Université de Lorraine, Dec. 2023. URL: https://hal.univ-lorraine.fr/tel-04602247 (cit. on p. 13).