RESEARCH CENTRE

Nancy - Grand Est

IN PARTNERSHIP WITH:

CNRS, Université de Lorraine

2021

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

Contents

Project-Team MULTISPEECH

Creation of the Project-Team: 2015 July 01

Keywords

Computer sciences and digital sciences

A3.4. – Machine learning and statistics

A3.4.6. – Neural networks

A3.4.8. – Deep learning

A3.5. – Social networks

A4.8. – Privacy-enhancing technologies

A5.1.5. – Body-based interfaces

A5.1.7. – Multimodal interfaces

A5.6.2. – Augmented reality

A5.7. – Audio modeling and processing

A5.7.1. – Sound

A5.7.3. – Speech

A5.7.4. – Analysis

A5.7.5. – Synthesis

A5.8. – Natural language processing

A5.9. – Signal processing

A5.9.1. – Sampling, acquisition

A5.9.2. – Estimation, modeling

A5.9.3. – Reconstruction, enhancement

A5.10.2. – Perception

A5.11.2. – Home/building control and interaction

A6.2.4. – Statistical methods

A6.3.1. – Inverse problems

A6.3.5. – Uncertainty Quantification

A9.2. – Machine learning

A9.3. – Signal analysis

A9.4. – Natural language processing

A9.5. – Robotics

Other research topics and application domains

B8.1.2. – Sensor networks for smart buildings

B8.4. – Security and personal assistance

B9.1.1. – E-learning, MOOC

B9.5.1. – Computer science

B9.5.2. – Mathematics

B9.5.6. – Data science

B9.6.8. – Linguistics

B9.6.10. – Digital humanities

B9.10. – Privacy

1 Team members, visitors, external collaborators

Research Scientists

- Denis Jouvet [Team leader, Inria, Senior Researcher, HDR]
- Anne Bonneau [CNRS, Researcher]
- Antoine Deleforge [Inria, Researcher]
- Dominique Fohr [CNRS, Researcher]
- Yves Laprie [CNRS, Senior Researcher, HDR]
- Paul Magron [Inria, Researcher, from Oct 2021]
- Mostafa Sadeghi [Inria, Starting Faculty Position]
- Md Sahidullah [Inria, Starting Research Position, until Aug 2021]
- Emmanuel Vincent [Inria, Senior Researcher, HDR]

Faculty Members

- Vincent Colotte [Univ de Lorraine, Associate Professor]
- Irène Illina [Univ de Lorraine, Associate Professor, HDR]
- Slim Ouni [Univ de Lorraine, Associate Professor, HDR]
- Agnès Piquard-Kipffer [Univ de Lorraine, Associate Professor, until Aug 2021]
- Romain Serizel [Univ de Lorraine, Associate Professor]

Post-Doctoral Fellows

- Théo Biasutto-Lervat [Univ de Lorraine, from Apr 2021 until Oct 2021]
- Felix Gontier [Inria, from Feb 2021]
- Imran Sheikh [Inria, until Aug 2021]

PhD Students

- Louis Abel [Univ de Lorraine, from Oct 2021]
- Tulika Bose [Univ de Lorraine]
- Pierre Champion [Inria]
- Can Cui [Inria, from Oct 2021]
- Ashwin Geet D'sa [Univ de Lorraine]
- Stéphane Dilungana [Inria]
- Sandipana Dowerah [Inria]
- Adrien Dufraux [Facebook, CIFRE]
- Raphaël Duroselle [Ministère des armées, until Aug 2021]
- Francois Effa [Univ de Lyon]

- Nicolas Furnon [Univ de Lorraine]
- Mickaëlla Grondin [CNRS, from Nov 2021]
- Seyed Ahmad Hosseini [Univ de Lorraine]
- Ajinkya Kulkarni [Univ de Lorraine, until Oct 2021]
- Xuechen Liu [Inria, until Aug 2021]
- Sewade Olaolu Ogun [Inria, from Oct 2021]
- Mauricio Michel Olvera Zambrano [Inria]
- Manuel Pariente [Univ de Lorraine, until Aug 2021]
- Shakeel Ahmad Sheikh [Univ de Lorraine]
- Vinicius De Paulo Souza Ribeiro [Univ de Lorraine]
- Tom Sprunck [Inria, from Nov 2021]
- Prerak Srivastava [Inria]
- Nicolas Turpault [Inria, until Mar 2021]
- Nicolas Zampieri [Inria]
- Georgios Zervakis [Inria]

Technical Staff

- Ismaël Bada [Univ de Lorraine, Engineer, until Mar 2021]
- Akira Campbell [Inria, Engineer, until Nov 2021]
- Joris Cosentino [Inria, Engineer]
- Louis Delebecque [Univ de Lorraine, Engineer]
- Hubert Nourtel [Inria, Engineer]
- Francesca Ronchini [Inria, Engineer]
- Mehmet Ali Tugtekin Turan [Inria, Engineer, until Feb 2021]

Interns and Apprentices

- Louis Abel [Univ de Lorraine, from Mar 2021 until Aug 2021]
- Khalig Aghakarimov [Inria, from Mar 2021 until Jul 2021]
- Awais Akbar [CNRS, from Mar 2021 until Jul 2021]
- Colleen Beaumard [Inria, from Jul 2021 until Sep 2021]
- Rémi Bouteiller [École normale supérieure Paris-Saclay, from May 2021 until Jul 2021]
- Khaoula Chahdi [Inria, from Apr 2021 until Aug 2021]
- Saurav Jha [Inria, from Mar 2021 until Jul 2021]
- Pavithra Poornachandran [CNRS, from Mar 2021 until Jul 2021]
- Chanoudom Prach [Inria, from Mar 2021 until Aug 2021]
- Ali Rida Sahili [Inria, from Apr 2021 until Sep 2021]
- Taha Toufik [Inria, from Apr 2021 until Jul 2021]
- Emilien Visentini [Univ de Lorraine, from Apr 2021 until Jul 2021]

Administrative Assistants

- Helene Cavallini [Inria]
- Delphine Hubert [Univ de Lorraine]
- Anne-Marie Messaoudi [CNRS]

External Collaborators

- Xuechen Liu [Univ de l'est de la Finlande, from Sep 2021]
- Md Sahidullah [Independent Researcher, from Sep 2021]
- Brij Mohan Lal Srivastava [Univ de Lille, until Sep 2021]

2 Overall objectives

The goal of the project is the modeling of speech for facilitating oral-based communication. The name MULTISPEECH comes from the following aspects that are particularly considered.

- **Multisource aspects** - which means dealing with speech signals originating from several sources, such as speaker plus noise, or overlapping speech signals resulting from multiple speakers; sounds captured from several microphones are also considered.
- **Multilingual aspects** - which means dealing with speech in a multilingual context, as for example for computer assisted language learning, where the pronunciation of words in a foreign language (i.e., non-native speech) is strongly influenced by the mother tongue.
- **Multimodal aspects** - which means considering simultaneously the various modalities of speech signals, acoustic and visual, in particular for the expressive synthesis of audio-visual speech.

Our objectives are structured in three research axes, which have evolved compared to the original project proposal in 2014. Indeed, due to the ubiquitous use of deep learning, the distinction between 'explicit modeling' and 'statistical modeling' is not relevant anymore and the fundamental issues raised by deep learning have grown into a new research axis 'beyond black-box supervised learning'. The three research axes are now the following.

- **Beyond black-box supervised learning** This research axis focuses on fundamental, domain-agnostic challenges relating to deep learning, such as the integration of domain knowledge, data efficiency, or privacy preservation. The results of this axis naturally apply in the various domains studied in the two other research axes.
- **Speech production and perception** This research axis covers the topics of the research axis on 'Explicit modeling of speech production and perception' of the project proposal, but now includes a wide use of deep learning approaches. It also includes topics around prosody that were previously in the research axis on 'Uncertainty estimation and exploitation in speech processing' in the project proposal.
- **Speech in its environment** The themes covered by this research axis mainly correspond to those of the axis on 'Statistical modeling of speech' in the project proposal, plus the acoustic modeling topic that was previously in the research axis on 'Uncertainty estimation and exploitation in speech processing' in the project proposal.

A large part of the research is conducted on French and English speech data; German and Arabic languages are also considered either in speech recognition experiments or in language learning. Adaptation to other languages of the machine learning based approaches is possible, depending on the availability of speech corpora.

3 Research program

3.1 Beyond black-box supervised learning

This research axis focuses on fundamental, domain-agnostic challenges relating to deep learning, such as the integration of domain knowledge, data efficiency, or privacy preservation. The results of this axis naturally apply in the domains studied in the two other research axes.

3.1.1 Integrating domain knowledge

State-of-the-art methods in speech and audio are based on neural networks trained for the targeted task. This paradigm faces major limitations: lack of interpretability and of guarantees, large data requirements, and inability to generalize to unseen classes or tasks. We research **deep generative models** as a way to learn task-agnostic probabilistic models of audio signals and design inference methods to combine and reuse them for a variety of tasks. We pursue our investigation of hybrid methods that combine the representational power of deep learning with **statistical signal processing** expertise by leveraging recent optimization techniques for non-convex, non-linear inverse problems. We also explore the integration of deep learning and **symbolic reasoning** to increase the generalization ability of deep models and to empower researchers/engineers to improve them.

3.1.2 Learning from little/no labeled data

While fully labeled data are costly, unlabeled data are cheap but provide intrinsically less information. **Weakly supervised learning** based on not-so-expensive incomplete and/or noisy labels is a promising middle ground. This entails modeling label noise and leveraging it for unbiased training. Models may depend on the labeler, the spoken context (voice command), or the temporal structure (ambient sound analysis). We also study **transfer learning** to adapt an expressive (audiovisual) speech synthesizer trained on a given speaker to another speaker for which only neutral voice data has been collected.

3.1.3 Preserving privacy

Some voice technology companies process users' voices in the cloud and store them for training purposes, which raises privacy concerns. We aim to **hide speaker identity** and (some) speaker states and traits from the speech signal, and evaluate the resulting automatic speech/speaker recognition accuracy and subjective quality/intelligibility/identifiability, possibly after removing private words from the training data. We also explore **semi-decentralized learning** methods for model personalization, and seek to obtain statistical guarantees.

3.2 Speech production and perception

This research axis covers topics related to the production of speech through articulatory modeling and multimodal expressive speech synthesis, and topics related to the perception of speech through the categorization of sounds and prosody in native and in non-native speech.

3.2.1 Articulatory modeling

Articulatory speech synthesis relies on 2D and 3D modeling of the **dynamics of the vocal tract** from real-time MRI data. The prediction of glottis opening is also considered so as to produce better quality acoustic events for consonants. The **coarticulation model** developed to handle the animation of the visible articulators will be extended to control the face and the tongue. This helps characterize links between the vocal tract and the face, and illustrate inner mouth articulation to learners. The suspension of articulatory movements in stuttering speech is also studied.

3.2.2 Multimodal expressive speech

The dynamic realism of the animation of the talking head, which has a direct impact on audiovisual intelligibility, continues to be our goal. Both the **animation** of the lower part of the face relating to speech and of the upper part relating to the facial expression are considered, and development continues towards a multilingual talking head. We investigate further the modeling of **expressivity** both for audio-only and for audiovisual speech synthesis. We also evaluate the benefit of the talking head in various use cases, including children with language and learning disabilities or deaf people.

3.2.3 Categorization of sounds and prosody

Reading and speaking are basic skills that need to be mastered. Further analysis of schooling experience will allow a better understanding of reading acquisition, especially for children with some language impairment. With respect to L1/L2 language interference[1], a special focus is set on the impact of L2 prosody on segmental realizations. Prosody is also considered for its implication on the structuration of speech communication, including on discourse particles. Moreover, we experiment the usage of speech technologies for computer assisted language learning in middle and high schools, and, hopefully, also for helping children learning to read.

3.3 Speech in its environment

The themes covered by this research axis correspond to the acoustic environment analysis, to speech enhancement and noise robustness, and to linguistic and semantic processing.

3.3.1 Acoustic environment analysis

Audio scene analysis is key to characterize the environment in which spoken communication may take place. We investigate audio event detection methods that exploit both strongly/weakly labeled and unlabeled data, operate in real-world conditions, can discover novel events, and provide a semantic interpretation. We keep working on source localization in the presence of nearby acoustic reflectors. We also pursue our effort at the interface of **room acoustics** to blindly estimate room properties and develop acoustics-aware signal processing methods. Beyond spoken communication, this has many applications to surveillance, robot audition, building acoustics, and augmented reality.

3.3.2 Speech enhancement and noise robustness

We pursue **speech enhancement** methods targeting several distortions (echo, reverberation, noise, overlapping speech) for both speech and speaker recognition applications, and extend them to ad-hoc arrays made of the microphones available in our daily life using multi-view learning. We also continue to explore statistical signal models **beyond the usual zero-mean complex Gaussian model** in the time-frequency domain, e.g., deep generative models of the signal phase. **Robust acoustic modeling** will be achieved by learning domain-invariant representations or performing unsupervised domain adaptation on the one hand, and by extending our uncertainty-aware approach to more advanced (e.g., nongaussian) uncertainty models and accounting for the additional uncertainty due to short utterances on the other hand, with application to speaker and language recognition “in the wild”.

3.3.3 Linguistic and semantic processing

We seek to address robust speech recognition by exploiting word/sentence embeddings carrying **semantic information** and combining them with acoustical uncertainty to rescore the recognizer outputs. We also combine semantic content analysis with text obfuscation models (similar to the label noise models to be investigated for weakly supervised training of speech recognition) for the task of detecting and classifying (hateful, aggressive, insulting, ironic, neutral, etc.) **hate speech** in social media.

[1] L1 refers to the speaker’s native language, and L2 to a speaker’s second language, usually learned later as a foreign language

4 Application domains

Approaches and models developed in the MULTISPEECH team are intended to be used for facilitating oral communication in various situations through enhancements of communication channels, either directly via automatic speech recognition or speech production technologies, or indirectly, thanks to computer assisted language learning. Applications also include the usage of speech technologies for helping people in handicapped situations or for improving their autonomy. Related application domains include multimodal computer interaction, private-by-design robust speech recognition, health and autonomy (more precisely aided communication and monitoring), and computer assisted learning.

4.1 Multimodal Computer Interaction

Speech synthesis has tremendous applications in facilitating communication in a human-machine interaction context to make machines more accessible. For example, it started to be widely common to use acoustic speech synthesis in smartphones to make possible the uttering of all the information. This is valuable in particular in the case of handicap, as for blind people. Audiovisual speech synthesis, when used in an application such as a talking head, i.e., virtual 3D animated face synchronized with acoustic speech, is beneficial in particular for hard-of-hearing individuals. This requires an audiovisual synthesis that is intelligible, both acoustically and visually. A talking head could be an interface between two persons communicating remotely when their video information is not available, and can also be used in language learning applications as vocabulary tutoring or pronunciation training tool. Expressive acoustic synthesis is of interest for the reading of a story, such as an audiobook, as well as for better human-machine interaction.

4.2 Private-by-design robust speech recognition

Many speech-based applications process speech signals on centralized servers. However speech signals exhibit a lot of private information. Processing them directly on the user's terminal helps keeping such information private. It is nevertheless necessary to share large amounts of data collected in actual application conditions to improve the modeling and thus the quality of the resulting services. This can be achieved by anonymizing speech signals before sharing them. With respect to robustness to noise and environment, the speech recognition technology is combined with speech enhancement approaches that aims at extracting the target clean speech signal from a noisy mixture (environment noises, background speakers, reverberation, ...).

4.3 Aided Communication and Monitoring

Source separation techniques should help locate and monitor people through the detection of sound events inside apartments, and speech enhancement is mandatory for hands-free vocal interaction. A foreseen application is to improve the autonomy of elderly or disabled people, e.g., in smart home scenarios. In the longer term, adapting speech recognition technologies to the voice of elderly people should also be useful for such applications, but this requires the recording of suitable data. Sound monitoring in other application fields (security, environmental monitoring) can also be envisaged.

4.4 Computer Assisted Learning

Although speaking seems quite natural, learning foreign languages, or one's mother tongue for people with language deficiencies, represents critical cognitive stages. Hence, many scientific activities have been devoted to these issues either from a production or a perception point of view. The general guiding principle with respect to computer assisted mother or foreign language learning is to combine modalities or to augment speech to make learning easier. Based upon an analysis of the learner's production, automatic diagnoses can be considered. However, making a reliable diagnosis on each individual utterance is still a challenge, which is dependent on the accuracy of the segmentation of the speech utterance into phones, and of the computed prosodic parameters.

5 Social and environmental responsibility

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

6 Highlights of the year

Emmanuel Vincent was elevated as IEEE Fellow for his contributions to audio source separation and challenge series methodology. He also received the ISCA Award for the best paper published in *Computer Speech and Language* (2016-2020) [93].

Arie Nugraha, a former PhD student of the team, received the 6th IEEE Signal Processing Society (SPS) Japan Young Author Best Paper Award for an article published during his PhD [91].

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

The theater play *Binôme* inspired by Antoine Deleforge's research was premiered at the Avignon Festival.

7 New software and platforms

7.1 New software

7.1.1 COMPRISE Voice Transformer

Name: COMPRISE Voice Transformer

Keywords: Speech, Privacy

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

Release Contributions: This version gives access to the 2 generations of tools that have been used to transform the voice, as part of the COMPRISE project (https://www.compriseh2020.eu/). The first one is a python library that implements 2 basic voice conversion methods, both using VLTN. The second one implements an anonymization method using x-vectors and neural waveform models.

News of the Year: We modified the x-vector based transformer by fixing the percentile-based pitch conversion method, using conda in Docker to fix issues with the Python version, and adding data from the speaker pool to simplify quick start.

URL: https://gitlab.inria.fr/comprise/voice_transformation

Contact: Marc Tommasi

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

7.1.2 COMPRISE Weakly Supervised STT

Name: COMPRISE Weakly Supervised Speech-to-Text

Keywords: Speech recognition, Language model, Acoustic Model

Functional Description: COMPRISE Weakly Supervised Speech-to-Text provides two main components for training Speech-to-Text (STT) models. These two components represent the two main approaches proposed in the COMPRISE project, namely (a) semi-supervised training driven by error predictions and (b) weakly supervised training based on utterance level weak labels. These two approaches can be used independently or together. The implementation builds on the Kaldi toolkit. It mainly focuses on obtaining reliable transcriptions of un-transcribed speech data which can be used for training both STT acoustic model (AM) and language model (LM). AM can be any type, although we choose the state-of-the-art TDNN Chain AM in our examples. Statistical n-gram LMs are chosen to support limited data scenarios.

News of the Year: We added a new Confusion Network based Language Model Training (CN2LM) component. In addition, we updated the setup script, and made a few additional fixes to the code and the documentation.

URL: https://gitlab.inria.fr/comprise/speech-to-text-weakly-supervised-learning

Authors: Imran Sheikh, Emmanuel Vincent, Irina Illina

Contact: Emmanuel Vincent

7.1.3 Asteroid

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

Keywords: Source Separation, Deep learning

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

News of the Year: - Added GEVD beamformer - Added recipe for Multi-Decoder DPRNN - Enable WER evaluation with GPU - Added model and support for voice activity detection

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

Contact: Antoine Deleforge

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

7.1.4 Web-based Pronunciation Learning Application

Keywords: Pronunciation training, Talking head, Second language learning

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

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

News of the Year: The flash cards application is quite complete. We have completely developed the student version. The teacher version is well developed. We have completed the administration interface to add/remove a teacher/student/class account. It is planned to have a collaboration with the DANE and the rectorat Nancy-Metz to test the platform with the students of the colleges.

Contact: Slim Ouni

Participants: Theo Biasutto–Lervat, Denis Jouvet, Slim Ouni, Thomas Girod, Leon Rohrbacher

7.1.5 HUMAN

Name: Hierarchical Universal Modular ANnotator

Keyword: Annotation tool

Scientific Description: A lot of real-world phenomena are complex and cannot be captured by single task annotations. This causes a need for subsequent annotations, with interdependent questions and answers describing the nature of the subject at hand. Even in the case a phenomenon is easily captured by a single task, the high specialization of most annotation tools can result in having to switch to another tool if the task only slightly changes. HUMAN is a novel web-based annotation tool that addresses the above problems by a) covering a variety of annotation tasks on both textual and image data, and b) the usage of an internal deterministic state machine, allowing the researcher to chain different annotation tasks in an interdependent manner. Further, the modular nature of the tool makes it easy to define new annotation tasks and integrate machine learning algorithms e.g., for active learning. HUMAN comes with an easy-to-use graphical user interface that simplifies the annotation task and management.

Functional Description: Hierarchical: Supports annotation of hierarchical data. This makes it easy to annotate instances (e.g. online comments) together with their context (e.g. the thread of comments a comment was posted in). Universal: Handles both textual data with and without context as well as PDFs and image annotation. Modular: Various question types (labeling questions, multiple-choice, yes-no, setting bounding boxes etc.) that are self-contained and can be arranged in any order needed. This also makes it easy to implement new custom question types and features. ANnotator: Comes with an easy to use GUI interface for your annotators and project manager.

News of the Year: HUMAN was used for the annotation of the MPHASIS corpus

URL: https://github.com/uds-lsv/human/

Publication: hal-02958831

Contact: Ashwin D'Sa

8 New results

8.1 Beyond black-box supervised learning

Participants: Antoine Deleforge, Denis Jouvet, Emmanuel Vincent, Vincent Colotte, Irène Illina, Romain Serizel, Imran Sheikh, Pierre Champion, Adrien Dufraux, Ajinkya Kulkarni, Sewade Olaolu Ogun, Manuel Pariente, Georgios Zervakis, Akira Campbell, Hubert Nourtel, Mehmet Ali Tuğtekin Turan.

8.1.1 Integrating domain knowledge

Integration of signal processing knowledge. State-of-the-art methods for single-channel speech enhancement or separation are based on discriminative neural networks. We finalized our work on generative modeling by variational autoencoders (VAEs), which allow generalization to mixtures of sources not seen together in training. We extended the usual VAE model that represents the variance of the magnitude spectrogram into a new VAE model that represents the covariance matrix over the entire complex-valued spectrogram. Manuel Pariente successfully defended his PhD [75], which includes a chapter on this topic.

Integration of deep learning and symbolic knowledge. Word embeddings play a fundamental role in natural language processing. Retrofitting is a simple and effective technique for refining distributional word embeddings based on word similarity relations from a semantic lexicon. Inspired by this technique, we designed two methods for incorporating similarity relations into contextualized BERT (Bidirectional Encoder Representations from Transformer) embeddings and evaluated them for medical relation extraction and sentiment analysis tasks. We showed that these methods do not substantially impact the performance, and conducted a qualitative analysis of this negative result [67].

8.1.2 Learning from little/no labeled data

Training automatic speech recognition (ASR) language models on uncertain ASR hypotheses. ASR language models are typically trained on a large amount of text data comprising the target domain. Yet, in early development stages or privacy-critical applications, only a limited amount of in-domain speech data and an even smaller amount of manual text transcriptions, if any, are available. We proposed a sampling method to train and adapt recurrent neural network (RNN) language models on uncertain ASR hypotheses embedded in ASR confusion networks and achieved up to 12% relative reduction in perplexity with respect to training on 1-best hypotheses, without any manual transcriptions [82]. We extended this work to Transformer based language models [83].

Transfer learning applied to speech synthesis. We worked on the disentanglement of speaker, emotion and content for transferring expressivity information from one speaker to another one, particularly when only neutral speech data is available for the latter. A deep metric learning framework based on multiclass n-pair loss has been used for improving the latent representation of expressivity in a multispeaker text-to-speech system setting, which results in improved expressivity transfer. Using a deep metric learning helps to reduce the intra-class variance and increase the inter-class variance. We transfer the expressivity by using the latent variables for each emotion to generate expressive speech in the voice of a different speaker for which no expressive speech is available. The approach has been applied using an end-to-end text-to-speech synthesis system based on Tacotron 2 [42].

8.1.3 Preserving privacy

Speech signals convey a lot of private information. To protect speakers, we pursued our investigation of x-vector based voice anonymization. We conducted an extensive study of the impact of four design choices (speaker distance metric, target region of x-vector space, target gender, speaker- or utterance-level target selection) on privacy and utility [84]. We have studied the modification of the fundamental frequency to improve consistency with the selected target x-vector, especially in the case of cross gender voice conversion [23], and investigated the behavior of the anonymization process with respect to the selected x-vector target identities under a white-box assessment [24]. We have also explored attack scenarios of the voice anonymization system using various techniques of embeddings alignment [25], and evaluated

the impact of the voice anonymization process on emotional speech data [49]. Finally, we showed that slicing utterances into shorter segments further improves privacy at no cost in utility [81].

In a complementary line of work, we studied the adaptation of ASR language models trained on anonymized text data to the statistics of the original text data [87].

We analyzed the results of the 1st Voice Privacy Challenge which we had organized in 2020 in an article [86] and a detailed technical report [85]. We presented a survey of our work in this area at the 1st ISCA Symposium on Security and Privacy in Speech Communication [18].

8.2 Speech production and perception

Participants: Anne Bonneau, Dominique Fohr, Denis Jouvet, Yves Laprie, Vincent Colotte, Slim Ouni, Agnes Piquard-Kipffer, Louis Abel, Théo Biasutto-Lervat, Shakeel Ahmad Sheikh, Vinicius Souza Ribeiro, Seyed Ahmad Hosseini.

8.2.1 Articulatory modeling

Construction of a rt-MRI (real-time Magnetic Resonance Imaging) database for French. Despite their interest there are very few MRI corpora for languages other than English and none for French. In collaboration with the IADI laboratory, we have created a real-time MRI corpus for 10 healthy French speakers who each pronounced 77 sentences covering all consonantal contexts including the vowels /a,i,u,y/ [11]. A real-time MRI technology with temporal resolution of 20 ms was used to acquire vocal tract images of the participants speaking. The sound was recorded simultaneously along with MRI, denoised and temporally aligned with the images. The speech was transcribed to obtain phoneme-wise segmentation of the sound signal. We also acquired static 3D MR images for a wide list of French phonemes. In addition, we included annotations of spontaneous swallowing. This database is available here.

Estimating the shape of MRI para-sagittal slices during the production of CV (consonant followed by a vowel). The estimation of the 3D dynamic shape of the vocal tract is a challenge to better understand the behavior of speech articulators during speech production. We used a database of rt-MRI covering several para-sagittal slices for a limited number of CV and 8 speakers. Unlike the previous database the challenge is to align several para-sagittal rt-MRI slices through geometrical transformations. The learned transformations are applied to the midsagittal frames of the test speaker in order to estimate the neighboring sagittal frames. Several mono speaker models are combined to produce the final frame estimation. To evaluate the results [28], image cross-correlation between the original and the estimated frames was used. Results show good agreement between the original and the estimated shapes.

Prediction of the vocal tract shape from a sequence of phonemes to be articulated. In this work, we address the prediction of speech articulators' temporal geometric position from the sequence of phonemes to be articulated, supplemented by their target duration. For this purpose we exploited a set of real-time MRI sequences uttered by a female French speaker. The contours of five articulators were tracked automatically in each of the frames in the MRI video. Then, we explored the capacity of a bidirectional gated recurrent units to correctly predict each articulator's shape and position given the sequence of phonemes and their duration. We showed that our model [52] can achieve good results with minimal data, producing very realistic vocal tract shapes.

Multimodal coarticulation modeling. We have investigated labial coarticulation to animate a virtual face from speech. We have used phonetic information as input to ensure speaker independence. We used gated recurrent units to account for the dynamics of the articulation which is an essential point of the model. The initialization of the last layers of the network has greatly eased the training and helped to handle coarticulation [90]. It relies on dimensionality reduction strategies, which have allowed us to inject knowledge of a useful latent representation of the visual data into the network. The robustness

of the model allowed us to predict lip movements for French and German, and tongue movements for English and German. The evaluation of the model was carried out by means of objective measurements of the quality of the trajectories and by evaluating the realization of the critical articulatory targets. We also conducted a subjective evaluation of the quality of the lip animation of the talking head [73].

Identifying disfluency in stuttered speech. Within the ANR project BENEPHIDIRE, the goal is to automatically identify typical kinds of stuttering disfluency using acoustic and visual cues for their automatic detection. This year, we proposed StutterNet [57], a deep learning based stuttering detection system, capable of detecting and identifying various types of disfluencies. Currently, our method relies solely on the acoustic signal. We use a time-delay neural network suitable for capturing contextual aspects of the disfluent utterances. Our method achieves promising results and outperforms the state-of-the-art residual neural network based method. We continue collecting French audiovisual data of subjects who stutter.

8.2.2 Multimodal expressive speech

Expressive audiovisual synthesis. In the thesis of Sara Dahmani (defended end of 2020) we studied the application of unsupervised learning techniques for emotional speech modeling as well as methods for restructuring emotions representation to make it continuous and more flexible. By manipulating the latent vectors, we were able to generate nuances of a given emotion and to generate new emotions that do not exist in our database, with a coherent articulation. These work has been published in a journal [8].

Emotion recognition. Speech emotion recognition is an active research topic in the affective computing community. Although deep learning based methods relying on Mel spectrogram features or on raw audio show state-of-the-art results, their performance is not yet suitable for real-world deployment. We improved it by replacing the Mel spectrogram by a constant-Q transform (CQT) input representation [59]. In another work [58], we introduced the deep scattering network for speech emotion recognition. Our study reveals that the time and frequency invariance of scattering coefficients provides a representation that is robust against irrelevant variations.

8.2.3 Categorization of sounds and prosody

Non-native speech production. We investigated voicing assimilations produced by French learners of German, -knowing that French and German voicing assimilations are respectively regressive and progressive-, inside groups of obstruents made up of a voiceless stop followed by a voiced fricative. To that purpose, we exploited the corpus recorded in 2020. Assimilations have been the object of a number of perceptual studies in L2 ; some of them tend to show that they are compensated for by advanced speakers (speakers are able to recognize non native assimilated forms). There have been fewer studies on production but they tend to show that even advanced speakers did not realize correctly typical L2 assimilations. Our results are also in favor of a poor acquisition, even by advanced speakers. The nature of assimilations, that involves universal mechanisms and language specificities, themselves dependent upon phonetic implementation (at sound level) should be taken into account to understand both learners' realizations and differences among perceptual and production studies [19, 68].

Language and reading acquisition by children. We studied the impact of lip-reading on speech perception in French-speaking children at-risk for reading failure. We followed a group of children at risk for reading failure and another group not at risk from age 5 to 7. Our hypothese was that, in the context of the COVID-19 pandemic while most teachers wear masks, it could affect learning to read, especially for children with poor phonemic discrimination skills. The results revealed a positive effect of lip-reading condition only for the at-risk group at both ages, suggesting that in the context of the COVID-19 pandemic in which teachers wear masks, this condition may interfere with learning to read for children at risk due to poor phonemic discrimination skills [15].

Computer assisted language learning. The goal of the METAL project is to provide tools to assist in foreign language pronunciation learning. We have developed a web-based learning platform that presents tutoring aspects illustrated by a talking head to show proper articulation of words and sentences; as well as using automatic tools derived from speech recognition technology, for analyzing student pronunciations. The front-end and back-end of the web application are almost finished and will be used by teachers to prepare pronunciation lessons, and by secondary school students learning German. The automatic analysis of student pronunciation is still not completed, and more development will be continued.

Prosody. The investigation of prosodic correlates of a few discourse particles has been finalized. In particular prosodic correlates of pragmatic functions have been compared accross languages (French and English) on prepared speech, and accross various speech styles. Lou Lee successfully defended her PhD thesis on this topic.

8.3 Speech in its environment

Participants: Antoine Deleforge, Dominique Fohr, Denis Jouvet, Paul Magron, Mostafa Sadeghi, Md Sahidullah, Emmanuel Vincent, Irène Illina, Romain Serizel, Félix Gontier, Tulika Bose, Can Cui, Stephane Dilungana, Sandipana Dowerah, Ashwin Geet D'sa, Raphaël Duroselle, François Effa, Nicolas Furnon, Xuechen Liu, Mauricio Michel Olvera Zambrano, Tom Sprunck, Prerak Srivastava, Nicolas Turpault, Nicolas Zampieri, Ismaël Bada, Joris Cosentino, Louis Delebecque, Francesca Ronchini.

8.3.1 Acoustic environment analysis

Ambient sound recognition. Sound event tagging is the task of finding what sound events occurred in a given time window. Since obtaining a large dataset with strongly labeled events (i.e., with onset and offset timestamps) is prohibitive, weakly labeled data (i.e., without timestamps) is often used instead. We explored the limitations induced by relying only on such weak labels [88]. Nicolas Turpault successfully defended his PhD, which includes a chapter on this topic [77]. An alternative is to generate synthetic soundscapes for which strong annotations are cheap to obtain at the cost of a domain mismatch with recorded evaluation data. We studied the impact of non-target events in such synthetic soundscapes [53]. We also proposed an efficient domain adaptation approach that relies on an auxiliary foreground-background classifier [50].

An additional problem when working with real, complex soundscapes is that they can involve multiple overlapping sound events. We proposed to adapt a standard sound separation algorithm and used it as a front-end to sound event detection [51].

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) 2021 Challenge [53] and published a detailed analysis of the submissions to the previous iteration of this task in 2020 [63, 31]. In 2021, the task still focused on the problem of learning from audio segments that are either weakly labeled or unlabeled with an additional focus on investigating the use of sound separation as a pre-processing to sound event detection, in particular in order to mitigate the problem of overlapping sound events [64].

Automatic audio captioning. We started working on automatic audio captioning focusing on incorporating knowledge from pre-trained audio tagging and natural language processing models within an audio captioning solution adapted to a specific corpus [36].

Acoustical room properties. We also pursued our work on the estimation of acoustical room properties (room shape, reverberation time, absorption coefficients) from recorded audio. While existing methods operate on single-channel recordings, we proposed a method that leverages two-channel recordings from

multiple, unknown source-receiver positions [62]. We also proposed a method for estimating the mean absorption coefficients of the walls in a room from an impulse response [9]. This is the first learning-based work in this field that studies in depth different simulation strategies for training and their impact on real-data results.

8.3.2 Speech enhancement and noise robustness

Overlapped speech detection and speaker counting. We pursued our study of overlapped speech detection and speaker counting using distant microphone arrays. We introduced a Transformer based architecture for this task, and proposed ways of exploiting multichannel input by means of early or late fusion of single-channel features with spatial features extracted from one or more microphone pairs. Extensive experiments on the AMI and CHiME-6 datasets showed that the proposed system significantly outperforms previous ones [7].

Speech enhancement. We pursued our investigation of multichannel speech separation. We analyzed the impact of speaker localization errors on speech separation for automatic speech recognition [60].

Nicolas Furnon succesfully defended his PhD thesis on multi-node deep neural network (DNN) based mask estimation integrated for speech enhancement with ad-hoc microphone arrays. The approach proposed allows for efficiently exploiting the diversity of the information provided by each node of the array during the mask estimation [10]. Extensions of the algorithm have been proposed using attention to enforce robustness to missing nodes [33] or with application to speech separation in a meeting setup [34].

Finally, we used a feature attribution-based explanation method to analyze the impact of the type of acoustic noise in the training data for speech enhancement models on the performance of the resulting models [61].

Speaker recognition and diarization. Developing a robust speaker recognition system remains a challenging task due to the variations in environmental conditions, channel effect, speech duration, and spoofing attacks. We explored a range of input features that substantially improve performance with respect to the commonly Mel frequency cepstral coefficients (MFCCs), based on replacing the linear transforms in the MFCC processing chain by learned transforms [44] or by a learnable multi-taper spectrum estimator [13], optimizing power-normalized cepstral coefficients [45] for speaker recognition, and parameterized cepstral mean normalization [46]. We also showed that utterance partitioning substantially improves text-independent speaker recognition performance with short utterances [17], and participated in the 1st Short-duration Speaker Verification (SdSV) Challenge [55].

Concerning spoofing attacks, we demonstrated that spoofing detection becomes more challenging if the speech recording is partially spoofed and proposed a solution based on frame selection [12]. We investigated the role of different factors in cross-corpora spoofing detection and found that state-of-the-art countermeasures are strongly impacted by speaker characteristics [26]. We proposed a framework to assess the similarity or complementarity of different classifiers for speaker recognition and anti-spoofing [41]. We summarized the findings and achievements of the ASVspoof 2019 challenge [14]. We co-organized the ASVspoof 2021 challenge which introduces a new subtask called *deepfake* detection. The results show that even though state-of-the-art spoofing detectors achieve good performance in known spoofing conditions, their generalization needs further investigation [65].

We participated in the third DIHARD challenge, whose goal is to perform speaker diarization of audio data collected from diverse real-world conditions including wide-band audio and telephone speech. We substantially improved the state-of-the-art baseline by integrating a domain identification method and making further processing domain-dependent [69, 80].

We investigated the robustness of speaker recognition systems with respect to environmental noises and reverberation. One approach was based on the use of a denoising autoencoder applied on the x-vectors to compensate for distortions due to noise and/or reverberation [47]. The other approach relies on an enhancement of the multichannel speech signals before giving the enhanced signal to the speaker verification system [79].

Language identification. State-of-the-art spoken language identification systems consist of three modules: a frame level feature extractor, a segment level embedding extractor (that provides x-vectors) and

a classifier. The performance of these systems degrades when facing mismatch between training and testing data. Although most domain adaptation methods focus on adaptation of the classifier, we have developed an unsupervised domain adaptation method for the segment level embedding extractor, which consists in adding a regularisation term associated to domain mismatch. Experiments were conducted with respect to transmission channel mismatch between telephone and radio channels using the RATS corpus. Another approach has been investigated, which relies on combining a classification loss with the metric learning n-pair loss for training the x-vector DNN model. Modeling and training strategies for the feature extractor (bottleneck features) have been investigated in details [30]. The various DNN based approaches for language identification have been combined with a conventional Gaussian mixture model approach, and the resulting system has been ranked first for cross channel language recognition, and for noisy data language identification at the Oriental Language Recognition challenge (OLR 2020) [29]. Raphaël Duroselle successfully defended his PhD thesis on these domain robust language identification approaches.

We have studied the cross-corpora performance for spoken language recognition with three corpora of Indian languages. The environment mismatch between corpora leads to significant performance degradation. Feature level compensation reduces the corpora mismatch, which leads to a significant improvement in the cross-corpora performance [27].

Unsupervised audio-visual speech enhancement and separation. Visual modality (lip movements of speaker) has proven to be very effective for speech enhancement and separation. While most of the existing works follow a supervised approach for audio-visual speech enhancement and separation, which require huge corpora and very deep neural networks for satisfactory generalization performance, we have developed a series of unsupervised approaches based on generative modeling of clean speech, requiring much less amount of data [16, 54, 48]. The underlying methodology is to combine traditional signal processing with the power of deep neural network, and its effectiveness to achieve good generalization performance has been experimentally verified against state-of-the-art supervised and classical approaches. We have also explored noisy visual data (non-frontal face images), and we have developed a robust face frontalization methodology to be used with our unsupervised audio-visual speech enhancement and separation framework [40]. The effectiveness of the proposed methodology has been experimentally verified, both in terms of some frontalization metrics, and some widely used speech enhancement performance metrics.

8.3.3 Linguistic and semantic processing

Detection of hate speech in social media. DNN-based classifiers have gained increased attention in hate speech classification. However, hate speech datasets consist of only a small amount of labeled data. To counter this, we explored data augmentation techniques to increase the amount of labeled samples, using a single class conditioned Generative Pre-Trained Transformer-2 (GPT-2). Adding a few hundred samples significantly improves the classifier's performance [35].

We proposed to use multiword expressions for automatic hate speech detection based on DNNs. Multiword expressions are lexical units greater than a word that have idiomatic and compositional meanings. We conducted experiments on two hate speech tweet corpora with two types of multiword expression embeddings, word2vec and BERT. Our experiments demonstrated that the proposed system significantly outperformed the baseline system in terms of macro-F1 [66]. We also conducted a comparative study of different features for efficient automatic hate speech detection [70].

State-of-the-art supervised models performance degrades when they are evaluated on abusive comments that differ from the training corpus. We have investigated if the performance of cross-corpora abuse detection can be improved by incorporating additional information from topic models. Our performance analysis revealed that topic models were able to capture abuse-related topics that could transfer across corpora, and resulted in improved generalisability [20]. We also investigated the effectiveness of several unsupervised domain adaptation approaches for the task of cross-corpora abusive language detection. Our evaluation showed that this resulted in sub-optimal performance, while the masked language model fine-tuning did better. A detailed analysis revealed the limitations of unsupervised domain adaptation [21].

Introduction of semantic information in an ASR system We aim to improve ASR performance by modeling long-term semantic relations. We proposed to perform this through DNN-based rescoring of the ASR n-best hypotheses, that combine semantic, acoustic, and linguistic information. Our DNN rescoring models are aimed at selecting hypotheses that have better semantic consistency and therefore lower word error rate. We investigated a powerful representation as part of input features to our DNN model: dynamic contextual embeddings from BERT. We performed experiments on the publicly available dataset TED-LIUM. The proposed rescoring approaches lead to significant performance improvement [38, 32].

9 Bilateral contracts and grants with industry

9.1 Bilateral grants with industry

9.1.1 Ministère des Armées

- Company: Ministère des Armées (France)
- Duration: Sep 2018 – Aug 2021
- Participants: Raphaël Duroselle, Denis Jouvet, Irène Illina
- Abstract: This contract corresponds to the PhD thesis of Raphaël Duroselle on the application of deep learning techniques for domain adaptation in speech processing.

9.1.2 Facebook

- Company: Facebook AI Research (France)
- Duration: Nov 2018 – Nov 2021
- Participants: Adrien Dufraux, Emmanuel Vincent
- Abstract: This CIFRE contract funds the PhD thesis of Adrien Dufraux. Our goal is to explore cost-effective weakly supervised learning approaches, as an alternative to fully supervised or fully unsupervised learning for automatic speech recognition.

10 Partnerships and cooperations

10.1 European initiatives

10.1.1 FP7 & H2020 projects

COMPRISE

Title: Cost-effective, Multilingual, Privacy-driven voice-enabled Services

Duration: Dec 2018 - Nov 2021

Coordinator: Emmanuel Vincent

Partners:

- Inria - also including MAGNET team (France)
- Ascora GmbH (Germany)
- Netfective Technology SA (France)
- Rooter Analysis SL (Spain)
- Tilde SIA (Latvia)
- Universität des Saarlandes (Germany)

Participants: Akira Campbell, Irène Illina, Denis Jouvet, Imran Sheikh, Brij Mohan Lal Srivastava, Mehmet Ali Tugtekin Turan, Emmanuel Vincent

Summary: COMPRISE has defined a fully private-by-design methodology and tools that reduce the cost and increase the inclusiveness of voice interaction technologies.

CPS4EU

Title: Cyber-physical systems for Europe

Duration: Jun 2019 - Jun 2022

Coordinator: Philippe Gougeon (Valeo)

Partners: 42 institutions and companies all across Europe

Participant: Francesca Ronchini, Romain Serizel

Summary: CPS4EU aims to develop key enabling technologies, pre-integration and development expertise to support the industry and research players' interests and needs for emerging interdisciplinary cyber-physical systems (CPS) and securing a supply chain ahead CPS enabling technologies and products. MULTISPEECH investigates approaches for audio event detection with applications to smart cities, tackling problems related to acoustic domain mistmatch, noisy mixtures or privacy preservation.

HumanE-AI-Net

Title: Making artificial intelligence human-centric

Duration: Sep 2020 - Aug 2023

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

Partners: 53 institutions and companies all across Europe

Participant: Slim Ouni

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

TAILOR

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

Duration: Sep 2020 - Aug 2023

Coordinator: Fredrik Heintz (Linköpings Universitet)

Partners: 53 institutions and companies all across Europe

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 2023

Coordinator: Holger Hoos (Universiteit Leiden)

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)
- PricewaterhouseCoopers Public Sector Srl (Italy)
- Thales SIX GTS France (France)
- Universiteit Leiden (Netherlands)
- University College Cork – National University of Ireland, Cork (Ireland)

Participant: Emmanuel Vincent

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

10.1.2 Other european programs/initiatives

M-PHASIS

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

Duration: Mar 2019 - Aug 2022

Program: ANR-DFG

Coordinators: Angeliki Monnier (CREM) and Christian Schemer (Johannes Gutenberg university)

Partners:

- CREM (Univ de Lorraine, France)
- LORIA (Univ de Lorraine, France)
- JGUM (Johannes Gutenberg-Universität, Germany)
- SAAR (Saarland University, Germany)

Participants: Ashwin Geet D'sa, Dominique Fohr, Irène Illina

Summary: Focusing on the social dimension of hate speech, M-PHASIS seeks to study the patterns of hate speech related to migrants, and to provide a better understanding of the prevalence and emergence of hate speech in user-generated content in France and Germany. Our contribution mainly concerns the automatic detection of hate speech in social media.

IMPRESS

Title: Improving Embeddings with Semantic Knowledge

Duration: Sep 2020 - Aug 2023

Partners:

- Inria (France)
- Deutsche Forschungszentrum für Künstliche Intelligenz GmbH (Germany)

Inria contact: Pascal Denis

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.

10.2 National initiatives

PIA2 ISITE LUE

Title: *Lorraine Université d'Excellence*

Duration: Avr 2016 - Jul 2021

Coordinator: Univ de Lorraine

Participants: Tulika Bose, Dominique Fohr, Irène Illina

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

E-FRAN METAL

Title: Modèles Et Traces au service de l'Apprentissage des Langues

Duration: Oct 2016 - Dec 2021

Coordinator: Anne Boyer (LORIA, Nancy)

Partners: LORIA, Interpsy, LISEC, ESPE de Lorraine, D@NTE (Univ. Versailles Saint Quentin), Sailendra SAS, ITOP Education, Rectorat.

Participants: Theo Biasutto-Lervat, Anne Bonneau, Vincent Colotte, Dominique Fohr, Denis Jouvet, Slim Ouni

Abstract: METAL aims at improving the learning of languages (written and oral) through development of new tools and analysis of numeric traces associated with students' learning. MULTISPEECH is concerned by oral language learning aspects.

ANR JCJC DiSCogs

Title: Distant speech communication with heterogeneous unconstrained microphone arrays

Duration: Sep 2018 – Aug 2022

Coordinator: Romain Serizel (LORIA, Nancy)

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

Collaborators: Télécom ParisTech, 7sensing

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

ANR DEEP-PRIVACY

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

Duration: Jan 2019 - Jun 2023

Coordinator: Denis Jouvet (Inria, Nancy)

Partners: MULTISPEECH (Inria Nancy), LIUM (Le Mans), MAGNET (Inria 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 – Jul 2023

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

ANR BENEPHIDIRE

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

Duration: Mar 2019 - Fev 2023

Coordinator: Praxiling (Toulouse)

Partners: Praxiling (Toulouse), LORIA (Nancy), INM (Toulouse), 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: Felix Gontier, Mauricio Michel Olvera Zambrano, Romain Serizel, Emmanuel Vincent, and Christophe Cerisara (CNRS - LORIA)

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: Irène 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 Flash Open Science HARPOCRATES

Title: Open data, tools and challenges for speaker anonymization

Duration: Oct 2019 - Sep 2021

Coordinator: Eurecom (Nice)

Partners: Eurecom (Nice), Inria (Nancy), LIA (Avignon)

Participants: Denis Jouvet, Md Sahidullah, Emmanuel Vincent

Abstract: HARPOCRATES supported the organization of the 1st VoicePrivacy Challenge, including data preparation and baseline software development.

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 (Inria, Nancy)

Participants: Antoine Deleforge, 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.

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

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

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

InriaHub ADT PEGASUS

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

Duration: Nov 2020 - Oct 2022

Coordinator: Antoine Deleforge

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

Abstract: This engineering project aims at further developing, expanding and transfering the Asteroid speech enhancement and separation toolkit recently released by the team [92].

ANR Full3DTalkingHead

Title: Synthèse articulatoire phonétique

Duration: Apr 2021 - Sep 2024

Coordinator: Yves Laprie (LORIA, Nancy)

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

Participants: Slim Ouni, Vinicius Ribeiro, Yves Laprie

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

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- Co-chair, 1st Inria-DFKI European Summer School on Artificial Intelligence, online, Jul 2021 (E. Vincent)

Member of the organizing committees

- Organizer, 2nd VoicePrivacy Challenge (E. Vincent)
- Organizer, ASVspoof 2021 Challenge and ASVspoof 2021 Workshop (M. Sahidullah, X. Liu)
- Organizer, 2nd Inria-DFKI Workshop on Artificial Intelligence, Kaiserslautern, Sep 2021 (E. Vincent)
- Organizer, Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge (R. Serizel, A. Deleforge)
- Organizer, Doctorales IA - Université de Lorraine, 23 November 2021 (Y. Laprie)

11.1.2 Scientific events: selection

Chair of conference program committees

- Area chair, 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (A. Deleforge, R. Serizel, E. Vincent)
- Area chair, 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (A. Deleforge, R. Serizel, E. Vincent)

Member of the conference program committees

- Member of program committee, 23rd International Conference on Speech and Computer (SPECOM 2021) (D. Jouvet)
- Member of program committee, 24th International Conference on Text, Speech and Dialogue (TSD 2021) (D. Jouvet)

Reviewer

- ASRU 2021 - IEEE Automatic Speech Recognition and Understanding Workshop (I. Illina, D. Jouvet, M. Sahidullah)
- ASVspoof 2021 - Automatic Speaker Verification and Spoofing. Countermeasures Challenge Workshop (M. Sahidullah)
- EUSIPCO 2021 - European Signal Processing Conference (V. Colotte, D. Jouvet, M. Sahidullah, R. Serizel)
- ICASSP 2021 - IEEE International Conference on Acoustics, Speech and Signal Processing (A. Bonneau, A. Deleforge, I. Illina, D. Jouvet, M. Sahidullah, R. Serizel, E. Vincent)
- ICML 2021 - International Conference on Machine Learning (A. Deleforge)
- INTERSPEECH 2021 (A. Bonneau, D. Jouvet, Y. Laprie, M. Sahidullah, R. Serizel, E. Vincent)
- JCP 2021 - *Journées de Phonétique Clinique* (Y. Laprie)

- NeurIPS 2021 - Conference and Workshop on Neural Information Processing Systems (A. Deleforge)
- PaPE 2021 - Phonetics and Phonology in Europe (A. Bonneau)
- SPECOM 2021 - International Conference on Speech and Computer (D. Jouvet)
- TSD 2021 - International Conference on Text, Speech and Dialogue (D. Jouvet)
- WASPAA 2021 - IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (A. Deleforge)

11.1.3 Journal

Member of the editorial boards

- Guest Editor of Computer Speech and Language, special issue on Voice Privacy (E. Vincent)
- Guest Editor of Neural Networks, special issue on Advances in Deep Learning Based Speech Processing (E. Vincent)
- Speech Communication (D. Jouvet)
- Associate Editor of IEEE/ACM Transactions on Audio, Speech and Language Processing (R. Serizel)
- Associate editor of EURASIP Journal on Audio, Speech, and Music Processing (Y. Laprie)

Reviewer - reviewing activities

- *Approche Neuropsychologique des Apprentissages* (A.Piquard-Kipffer)
- Computer Speech and Language (A.Bonneau, S. Ouni, M. Sahidullah)
- Digital Signal Processing (M. Sahidullah)
- EURASIP Journal on Audio, Speech, and Music Processing (A. Deleforge)
- IEEE Signal Processing Letters (M. Sahidullah, R. Serizel)
- IEEE/ACM Transactions on Audio, Speech, and Language Processing (V. Colotte, A. Deleforge, P. Magron, M. Sahidullah, R. Serizel)
- IEEE Transactions on Information Forensics and Security (M. Sahidullah)
- IEEE Transactions on Neural Networks and Learning Systems (P. Magron)
- Journal of Language, Speech and Hearing Research (Y. Laprie)
- Journal of the Acoustical Society of America (A. Deleforge, M. Sahidullah, Y. Laprie)
- Journal of the International Phonetic Association (A. Bonneau)
- Speech Communication (A. Deleforge, D. Jouvet, S. Ouni, M. Sahidullah)

11.1.4 Invited talks

- *Journée SdL (Sciences du Langage)*, Universiy of Lorraine, Nancy, April 2021 (S. Ouni)
- Intelligent Sensing Winter School, Queen Mary's University of London, December 2021 (A. Deleforge)

11.1.5 Leadership within the scientific community

- Member of the Steering Committee of ISCA's Special Interest Group on Security and Privacy in Speech Communication (E. Vincent).
- Member of DCASE steering group (R. Serizel)
- Member of IEEE acoustic and audio signal processing technical committee (A. Deleforge, R. Serizel)
- 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)

11.1.6 Scientific expertise

- Expertise for Bpifrance on a DeepTech startup funding request (A. Deleforge)
- Member of ANR Evaluation Committee 23 on Artificial Intelligence (E. Vincent)
- Member of ANR Evaluation Committe for ASTRID projects (D. Jouvet)
- Member of the Advisory Board of H2020 FVLLMONTI (E. Vincent)
- Member of the Scientific Committee of an Institute for deaf people, INJS-Metz (A. Piquard-Kipffer)
- Member of the hiring committee for Inria Senior Research Scientists (E. Vincent)
- Member of the hiring committee for a Professor, Avignon Université (D. Jouvet)
- Member of the hiring committee for Junior Research Scientists, Inria Lille - Nord Europe (E. Vincent)
- Member of the hiring committee for a CPPI Europe, Inria Nancy - Grand Est (E. Vincent)
- Member of the hiring committee for a permanent research engineer, Inria Nancy - Grand Est (E. Vincent)
- Member of the hiring committee for an Assistant Professor, Le Mans Université (D. Jouvet)
- Reviewer of ANR projects (D. Jouvet, Y. Laprie)
- Reviewer of CIFRE thesis proposal (D. Jouvet, S. Ouni)
- Reviewer of ERC projects (Y. Laprie)
- Reviewer of projects for Czech Science Foundation (I. Illina, D. Jouvet)
- Reviewer of projects for Austrian Academy of Sciences (I. Illina)

11.1.7 Research administration

- Head of the AM2I Scientific Pole of Université de Lorraine (Y. Laprie)
- Deputy Head of Science of Inria Nancy - Grand Est (E. Vincent)
- Scientific Director for the partnership between Inria and DFKI (E. Vincent)
- President of the *Commission Locale de Développement Durable* (CLDD) of Inria Nancy (A. Deleforge)
- Member of Management board of Université de Lorraine (Y. Laprie)
- Member of the CNU 27 (*Conseil National des Universités*) - Computer Science (S. Ouni)
- Member of Inria's Evaluation Committee (E. Vincent)

- Member of the *Comité Espace Transfert* of Inria Nancy - Grand Est (E. Vincent)
- Member of the commission for the scientific staff (COMIPERS) of the research center Inria Nancy - Grand Est (R. Serizel)
- Member of the commission for the technological developpment (CDT) of the research center Inria Nancy (R. Serizel)
- Member of *Commission paritaire* of Université de Lorraine (Y. Laprie)
- Member of the *Commission Locale de Développement Durable* (CLDD) of Inria Nancy (D. Fohr)
- Member of the *Commission des Utilisateurs des Moyens Informatiques* (CUMI) of Inria Nancy (D. Fohr)
- Member of the *Conseil de la Fédération Charles Hermite* (I. Illina)
- Member of the *Commission de Sélection ATER* (IUT Charlemagne) (I. Illina)

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- DUT: I. Illina, Java programming (100 hours), Linux programming (58 hours), and Advanced Java programming (40 hours), L1, University of Lorraine, France
- DUT: I. Illina, Supervision of student projects and internships (50 hours), L2, University of Lorraine, France
- DUT: R. Serizel, Introduction to office tools (108 hours), Multimedia and web (20 hours), Documents and databases (20 hours), L1, University of Lorraine, France
- DUT: R. Serizel, Multimedia content and indexing (14 hours), Content indexing and retrieval software (20 hours), L2, University of Lorraine, France
- DUT: S. Ouni, Programming in Java (24 hours), Web Programming (24 hours), Graphical User Interface (96 hours), L1, University of Lorraine, France
- DUT: S. Ouni, Advanced Algorihms (24 hours), L2, University of Lorraine, France
- Licence: A. Bonneau, Phonetics (17 hours), L2, *École d'audioprothèse*, University of Lorraine, France
- Licence: V. Colotte, Digital literacy and tools (hybrid courses, 50 hours), L1, University of Lorraine, France
- Licence: V. Colotte, System (35 hours), L3, University of Lorraine, France
- Licence: A. Piquard-Kipffer, Education Science (32 hours), L1, Département d'orthophonie, University of Lorraine, France
- Licence: A. Piquard-Kipffer, Learning to Read (34 hours), L2, Département d'orthophonie, University of Lorraine, France
- Licence: A. Piquard-Kipffer, Psycholinguistics (20 hours), Departement Orthophonie, University Pierre et Marie Curie, Paris, France
- Licence: A. Piquard-Kipffer, Dyslexia, Dysorthographia (12 hours), L3, Département d'orthophonie, University of Lorraine, France
- Licence: A. Piquard-Kipffer, Mathematics Didactics, 9 hours, L3, Departement Orthophonie, University of Lorraine, France

- Licence and Master: A. Deleforge, Introduction to Machine Learning, 12 hours L3, 12 hours M1, Télécom Physique Strasbourg, France
- Master: V. Colotte, Integration project: multimodal interaction with Pepper Robot (15 hours), M2, University of Lorraine, France
- Master: D. Jouvet and S. Ouni, Multimodal oral comunication (24 hours), M2, University of Lorraine, France
- Master: Y. Laprie, Speech corpora (30 hours), M1, University of Lorraine, France
- Master: S. Ouni, Multimedia in Distributed Information Systems (31 hours), M2, University of Lorraine, France
- Master: A. Piquard-Kipffer, Dyslexia, Dysorthographia diagnosis (6 hours), Deaf people & reading (21 hours), M1, Département d'orthophonie, University of Lorraine, France
- Master: A. Piquard-Kipffer, French Language Didactics (53 hours), M2, INSPE University of Lorraine, France
- Master: A. Piquard-Kipffer, Psychology (6 hours), M2, Departement of Psychology, University of Lorraine, France
- Executive Master : A. Piquard-Kipffer, Psychology, 12 hours, M2, Special Educational Needs with University of Lorraine, INSPÉ & UIR, International University of Rabat (Morocco)
- Master: R. Serizel, S. Ouni, P. Magron and V. Ribeiro, Oral speech processing (24 hours), M2, University of Lorraine
- Master: E. Vincent, A. Kulkarni and P. Magron, Neural networks (38 hours), M2, University of Lorraine
- Continuous training: A. Piquard-Kipffer, Special Educational Needs (53 hours), INSPE,University of Lorraine, France
- Continuous training: E. Vincent, Neural networks (14 hours), Data Scientist curriculum, University of Lorraine
- PhD: A. Piquard-Kipffer, Language Pathology (20 hours), EHESP, University of Sorbonne, Paris, France
- Other: V. Colotte, Co-Responsible for NUMOC (Digital literacy by hybrid courses) for the University of Lorraine, France (for 7000 students)
- Other: S. Ouni, Responsible of *Année Spéciale* DUT, University of Lorraine

11.2.2 Supervision

- PhD: Théo Biasutto-Lervat, "*Modélisation de la coarticulation multimodale : vers l'animation d'une tête parlante intelligible*", Jan 29, 2021, S. Ouni [73]
- PhD: Lou Lee, "*Fonctions pragmatiques et prosodie de marqueurs discursifs en français et en anglais*", Apr 6, 2021, Y. Keromnes (ATILF) and D. Jouvet
- PhD: Nicolas Turpault, "*Analyse des problématiques liées à la reconnaissance de sons ambiants en environnement réel*", May 31, 2021, R. Serizel and E. Vincent [77]
- PhD: Manuel Pariente, "Deep learning-based phase-aware audio signal modeling and estimation", Sep 29, 2021, A. Deleforge and E. Vincent [75]
- PhD: Raphaël Duroselle, "*Robustesse au canal des systèmes de reconnaissance de la langue*", Oct 28, 2021, D. Jouvet and I. Illina [74]

- PhD: Brij Mohan Lal Srivastava, “Speaker anonymization — Representation, evaluation and formal guarantees”, Dec 2, 2021, M. Tommasi (MAGNET project-team), E. Vincent and A. Bellet (MAGNET project-team) [76]
- PhD: Nicolas Furnon, “Deep-learning based speech enhancement with ad-hoc microphone arrays”, Dec 14, 2021, R. Serizel, I. Illina and S. Essid (Télécom ParisTech)
- PhD in progress: Ajinkya Kulkarni, “Expressive speech synthesis by deep learning”, Oct. 2018, V. Colotte and D. Jouvet
- PhD in progress: Adrien Dufraux, “Leveraging noisy, incomplete, or implicit labels for automatic speech recognition”, Nov 2018, E. Vincent, A. Brun (LORIA) and M. Douze (Facebook AI Research)
- PhD in progress: Ashwin Geet D’Sa, “Natural Language Processing: Online hate speech against migrants”, Apr 2019, I. Illina and D. Fohr
- PhD in progress: Tulika Bose, “Online hate speech and topic classification”, Sep 2019, I. Illina, D. Fohr and A. Monnier (CREM)
- PhD in progress: Mauricio Michel Olvera Zambrano, “Robust audio event detection”, Oct 2019, E. Vincent and G. Gasso (LITIS)
- PhD in progress: Pierre Champion, “Privacy preserving and personalized transformations for speech recognition”, Oct 2019, D. Jouvet and A. Larcher (LIUM)
- PhD in progress: Shakeel Ahmad Sheikh, “Identifying disfluency in speakers with stuttering, and its rehabilitation, using DNN”, Oct 2019, S. Ouni
- PhD in progress: Sandipana Dowerah, “Robust speaker verification from far-field speech”, Oct 2019, D. Jouvet and R. Serizel
- PhD in progress: Georgios Zervakis, “Integration of symbolic knowledge into deep learning”, Nov 2019, M. Couceiro (LORIA) and E. Vincent
- PhD in progress: Nicolas Zampieri, “Automatic classification using deep learning of hate speech posted on the Internet”, Nov. 2019, I. Illina and D. Fohr
- PhD in progress: Xuechen Liu, “Robust speaker recognition for smart assistant technology”, Jan 2020, M. Sahidullah
- PhD in progress: Prerak Srivastava, “Hearing the walls of a room: machine learning for audio augmented reality”, Oct 2020, A. Deleforge and E. Vincent
- PhD in progress: Stéphane Dilungana, “*L’intelligence artificielle au service du diagnostic acoustique : Apprendre à entendre les parois d’une salle*”, Oct 2020, A. Deleforge, C. Foy (UMR AE) and S. Faisan (iCube)
- PhD in progress: Vinicius Souza Ribeiro, “Tracking articulatory contours in MR images and prediction of the vocal tract shape from a sequence of phonemes to be articulated”, Oct 2020, Y. Laprie
- PhD in progress: François Effa, “Détection d’alarmes dans le bruit”, Jan 2021, R. Serizel, J.-P. Arz (INRS), N. Grimault (Centre de Recherche en Neurosciences de Lyon)
- PhD in progress: Seyed Ahmad Hosseini, “3D sign language generation”, Feb 2021, S. Ouni and M. Sadeghi
- PhD in progress: Louis Abel, “Expressive audio-visual speech synthesis in an interaction context”, Oct 2021, S. Ouni and V. Colotte

- PhD in progress: Can Cui, "*Séparation, diarisation et reconnaissance automatique de la parole conjointes et embarquées pour la génération de comptes-rendus de réunions*", 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: Tom Sprunck, "Hearing the Shape of a Room: Towards Acoustic Super-resolution", Nov 2021, A. Deleforge, C. Foy (UMR AE) and Y. Privat (Univ. Strasbourg)

- PhD in progress: Mickaëlla Grondin, "Modeling gestures and speech in interactions", Nov 2021, S. Ouni and F. Hirsch (Praxiling)

11.2.3 Juries

Participation in HDR and PhD juries

- Participation in the PhD jury of Hadrien Foroughmand (Sorbonne Université, Jan 2021), E. Vincent, member

- Participation in the PhD jury of Toni Heittola (Tampere University, Jun 2021), R. Serizel, reviewer and oponent

- Participation in the PhD jury of Corentin Guezenoc (Centrale Supélec, COMUE Université Bretagne Loire, Jun 2021), A. Deleforge, member

- Participation in the PhD jury of Valentin Gillot (Université Rennes 1, Sep 2021), E. Vincent, president

- Participation in the PhD jury of Théo Jourdan (Université de Lyon, Oct 2021), E. Vincent, reviewer

- Participation in the PhD jury of Andrea Vaglio (Institut Polytechnique de Paris, Nov 2021), E. Vincent, president

- Participation in the PhD jury of Kilian Schulze-Forster (Institut Polytechnique de Paris, Dec 2021), E. Vincent, president

- Participation in the PhD jury of Pierre-Amaury Grumiaux (Université de Grenoble, Dec 2021), R. Serizel, member

11.3 Popularization

11.3.1 Articles and contents

- Article "Enabling voice-based apps with European values", *ERCIM News*, Jul 2021 (A. Campbell, E. Vincent) [89]

- Interview for "*Dresseur d'intelligence artificielle, métier de demain ?*", *WE DEMAIN*, Nov 2021 (E. Vincent)

- Podcast "COMPRISE, the privacy-friendly, inclusive voice interface", podcast, Dec 2021 (E. Vincent)

- After its publication, the scientific paper "Impact of lip-reading on speech perception in French-speaking children at-risk for reading failure assessed from age 5 to 7." by A. Piquard-Kipffer et al. [15] has been mentionned and commented on several web sites (*Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation*, CNRS, Université de Lorraine, LORIA, *CNRS - le journal*, YouTube, SoundCloud(CNRS)), social networks (twitter) and journals (*L'Express*, *Le Figaro*, *revue Cerveau & Psycho*)

11.3.2 Interventions

- Chiche, Lycée Chopin, Nancy, Nov 2021 (R. Serizel)
- Chiche, Lycée Saint-Pierre Chanel, Thionville (2 classes), Dec 2021 (E. Vincent)
- Theater play "Drone Control" by *Les Sens des Mots*, written by Charlotte Lagrange, inspired by the research work of A. Deleforge, played 4 times in 2021 in front of general audiences (approx. 200 spectators in total)

12 Scientific production

12.1 Major publications

[1] S. Dahmani, V. Colotte, V. Girard and S. Ouni. 'Conditional Variational Auto-Encoder for Text-Driven Expressive AudioVisual Speech Synthesis'. In: *INTERSPEECH 2019 - 20th Annual Conference of the International Speech Communication Association*. Graz, Austria, Sept. 2019. URL: https://hal.inria.fr/hal-02175776.

[2] B. Elie and Y. Laprie. 'Acoustic impact of the gradual glottal abduction on the production of fricatives: A numerical study'. In: *Journal of the Acoustical Society of America* 142.3 (Sept. 2017), pp. 1303–1317. DOI: 10.1121/1.5000232. URL: https://hal.archives-ouvertes.fr/hal-01423206.

[3] A. A. Nugraha, A. Liutkus and E. Vincent. 'Multichannel audio source separation with deep neural networks'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 24.10 (June 2016), pp. 1652–1664. DOI: 10.1109/TASLP.2016.2580946. URL: https://hal.inria.fr/hal-01163369.

[4] I. A. Sheikh, D. Fohr, I. Illina and G. Linares. 'Modelling Semantic Context of OOV Words in Large Vocabulary Continuous Speech Recognition'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 25.3 (Jan. 2017), pp. 598–610. DOI: 10.1109/TASLP.2017.2651361. URL: https://hal.inria.fr/hal-01461617.

[5] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi and E. Vincent. 'Evaluating Voice Conversion-based Privacy Protection against Informed Attackers'. In: ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing. Barcelona, Spain, 4th May 2020, pp. 2802–2806. URL: https://hal.inria.fr/hal-02355115.

12.2 Publications of the year

International journals

[6] Z. Chelly Dagdia and C. Zarges. 'A detailed study of the distributed rough set based locality sensitive hashing feature selection technique'. In: *Fundamenta Informaticae* 182.2 (30th Sept. 2021), pp. 111–179. DOI: 10.3233/FI-2021-2069. URL: https://hal.inria.fr/hal-02880638.

[7] S. Cornell, M. Omologo, S. Squartini and E. Vincent. 'Overlapped speech detection and speaker counting using distant microphone arrays'. In: *Computer Speech and Language* 72 (11th Oct. 2021). DOI: 10.1016/j.csl.2021.101306. URL: https://hal.inria.fr/hal-03375681.

[8] 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.

[9] C. Foy, A. Deleforge and D. Di Carlo. 'Mean absorption estimation from room impulse responses using virtually supervised learning'. In: *Journal of the Acoustical Society of America* 150.2 (1st Jan. 2021), pp. 1286–1299. DOI: 10.1121/10.0005888. URL: https://hal.archives-ouvertes.fr/hal-03331250.

[10] 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.

[11] K. Isaieva, Y. Laprie, J. Leclère, I. K. Douros, J. Felblinger and P.-A. Vuissoz. 'Multimodal dataset of real-time 2D and static 3D MRI of healthy French speakers'. In: *Scientific Data* 8.1 (1st Oct. 2021), p. 258. DOI: 10.1038/s41597-021-01041-3. URL: https://hal.archives-ouvertes.fr/hal-03507532.

[12] K. A. Kumar, D. Paul, M. Pal, M. Sahidullah and G. Saha. 'Speech Frame Selection for Spoofing Detection with an Application to Partially Spoofed Audio-Data'. In: *International Journal of Speech Technology* (3rd Jan. 2021). DOI: 10.1007/s10772-020-09785-w. URL: https://hal.archives-ouvertes.fr/hal-03008912.

[13] X. Liu, M. Sahidullah and T. Kinnunen. 'Optimizing Multi-Taper Features for Deep Speaker Verification'. In: *IEEE Signal Processing Letters* 28 (2021), pp. 2187–2191. DOI: 10.1109/LSP.2021.3122796. URL: https://hal.archives-ouvertes.fr/hal-03394152.

[14] A. Nautsch, X. Wang, N. Evans, T. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi and K. A. Lee. 'ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech'. In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3.2 (Feb. 2021), pp. 252–265. DOI: 10.1109/TBIOM.2021.3059479. URL: https://hal.archives-ouvertes.fr/hal-03236124.

[15] A. Piquard-Kipffer, T. Cavadini, L. Sprenger-Charolles and E. Gentaz. 'Impact of lip-reading on speech perception in French-speaking children at risk for reading failure assessed from age 5 to 7'. In: *Annee Psychologique* 121 (June 2021), pp. 3–18. DOI: 10.3917/anpsy1.212.0003. URL: https://hal.inria.fr/hal-03482032.

[16] M. Sadeghi and X. Alameda-Pineda. 'Mixture of Inference Networks for VAE-based Audio-visual Speech Enhancement'. In: *IEEE Transactions on Signal Processing* 69 (9th Mar. 2021), pp. 1899–1909. DOI: 10.1109/TSP.2021.3066038. URL: https://hal.inria.fr/hal-02926172.

[17] N. Sen, M. Sahidullah, H. Patil, S. K. Das Mandal, S. K. Rao and T. K. Basu. 'Utterance partitioning for speaker recognition: an experimental review and analysis with new findings under GMM-SVM framework'. In: *International Journal of Speech Technology* 24 (Dec. 2021), pp. 1067–1088. DOI: 10.1007/s10772-021-09862-8. URL: https://hal.archives-ouvertes.fr/hal-03232723.

International peer-reviewed conferences

[18] J.-F. Bonastre, H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, P.-G. Noe, J. Patino, M. Sahidullah, B. M. L. Srivastava, M. Todisco, N. Tomashenko, E. Vincent, X. Wang and J. Yamagishi. 'Benchmarking and challenges in security and privacy for voice biometrics'. In: SPSC 2021 - 1st ISCA Symposium on Security and Privacy in Speech Communication. Magdeburg, Germany, 10th Nov. 2021. DOI: 10.21437/SPSC.2021-11. URL: https://hal.archives-ouvertes.fr/hal-03346196.

[19] A. Bonneau. 'Voicing assimilations by French Speakers of German in stop-fricative sequences'. In: INTERSPEECH 2021. Brno, Czech Republic, 30th Aug. 2021. DOI: 10.21437/Interspeech.2021-601. URL: https://hal.inria.fr/hal-03353139.

[20] T. Bose, I. Illina and D. Fohr. 'Generalisability of Topic Models in Cross-corpora Abusive Language Detection'. In: NLP4IF 2021 - Workshop Censorship, Disinformation, and Propaganda. Mexico city/Virtual, Mexico, 6th June 2021. URL: https://hal.inria.fr/hal-03212196.

[21] T. Bose, I. Illina and D. Fohr. 'Unsupervised Domain Adaptation in Cross-corpora Abusive Language Detection'. In: SocialNLP 2021 - The 9th International Workshop on Natural Language Processing for Social Media. Virtual, France, 10th June 2021. URL: https://hal.inria.fr/hal-03204605.

[22] E. Calò, L. Jacqmin, T. Rosemplatt, M. Amblard, M. Couceiro and A. Kulkarni. 'GECko+: a Grammatical and Discourse Error Correction Tool'. In: *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 3 : Démonstrations*. TALN 2021 - 28e Conférence sur le Traitement Automatique des Langues Naturelles. Lille / Virtual, France: ATALA, 2021, pp. 8–11. URL: https://hal.archives-ouvertes.fr/hal-03265914.

[23] P. Champion, D. Jouvet and A. Larcher. 'A Study of F0 Modification for X-Vector Based Speech Pseudonymization Across Gender'. In: PPAI 2021 - 2nd AAAI Workshop on Privacy-Preserving Artificial Intelligence. Virtual, China, 3rd Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-02995862.

[24] P. Champion, D. Jouvet and A. Larcher. 'Evaluating X-vector-based Speaker Anonymization under White-box Assessment'. In: SPECOM 2021 - 23rd International Conference on Speech and Computer. Saint Petersburg, Russia, 28th Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03351943.

[25] P. Champion, T. Thebaud, G. Le Lan, A. Larcher and D. Jouvet. 'On the invertibility of a voice privacy system using embedding alignement'. In: ASRU 2021 - IEEE Automatic Speech Recognition and Understanding Workshop. Cartagena, Colombia, 13th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03356021.

[26] B. Chettri, R. G. Hautamäki, M. Sahidullah and T. Kinnunen. 'Data Quality as Predictor of Voice Anti-Spoofing Generalization'. In: INTERSPEECH 2021. Brno, Czech Republic, 30th Aug. 2021. DOI: 10.21437/Interspeech.2021-1180. URL: https://hal.archives-ouvertes.fr/hal-03261131.

[27] S. Dey, G. Saha and M. Sahidullah. 'Cross-Corpora Language Recognition: A Preliminary Investigation with Indian Languages'. In: EUSIPCO 2021 - 29th European Signal Processing Conference. Dublin / Virtual, Ireland, 23rd Aug. 2021. DOI: 10.23919/EUSIPCO54536.2021.9616273. URL: https://hal.archives-ouvertes.fr/hal-03223314.

[28] I. K. Douros, A. Kulkarni, Y. Xie, C. Dourou, J. Felblinger, K. Isaieva, P.-A. Vuissoz and Y. Laprie. 'MRI Vocal Tract Sagittal Slices Estimation during Speech Production of CV'. In: EUSIPCO 2020 - 28th European Signal Processing Conference. Amsterdam / Virtual, Netherlands, 18th Jan. 2021. DOI: 10.23919/Eusipco47968.2020.9287834. URL: https://hal.inria.fr/hal-03090824.

[29] R. Duroselle, M. Sahidullah, D. Jouvet and I. Illina. 'Language recognition on unknown conditions: the LORIA-Inria-MULTISPEECH system for AP20-OLR Challenge'. In: INTERSPEECH 2021. Brno, Czech Republic, 30th Aug. 2021. DOI: 10.21437/Interspeech.2021-276. URL: https://hal.archives-ouvertes.fr/hal-03228823.

[30] R. Duroselle, M. Sahidullah, D. Jouvet and I. Illina. 'Modeling and training strategies for language recognition systems'. In: INTERSPEECH 2021. Brno, Czech Republic, 30th Aug. 2021. DOI: 10.21437/Interspeech.2021-277. URL: https://hal.archives-ouvertes.fr/hal-03264085.

[31] G. Ferroni, N. Turpault, J. Azcarreta, F. Tuveri, R. Serizel, Ç. Bilen and S. Krstulović. 'Improving Sound Event Detection Metrics: Insights from DCASE 2020'. In: ICASSP 2021 - 46th International Conference on Acoustics, Speech, and Signal Processing. Toronto/Virtual, Canada, 6th June 2021. DOI: 10.1109/ICASSP39728.2021.9414711. URL: https://hal.inria.fr/hal-02978422.

[32] D. Fohr and I. Illina. 'BERT-based Semantic Model for Rescoring N-best Speech Recognition List'. In: INTERSPEECH 2021. Proceedings of INTERSPEECH 2021. Brno, Czech Republic, 30th Aug. 2021. DOI: 10.21437/Interspeech.2021-313. URL: https://hal.archives-ouvertes.fr/hal-03248881.

[33] N. Furnon, R. Serizel, S. Essid and I. Illina. 'Attention-based distributed speech enhancement for unconstrained microphone arrays with varying number of nodes'. In: *European Signal Processing Conference (EUSIPCO)*. EUSIPCO 2021 - 29th European Signal Processing Conference. Dublin / Virtual, Ireland, 23rd Aug. 2021. DOI: 10.23919/EUSIPCO54536.2021.9616358. URL: https://hal.archives-ouvertes.fr/hal-03259801.

[34] N. Furnon, R. Serizel, I. Illina and S. Essid. 'Distributed speech separation in spatially unconstrained microphone arrays'. In: ICASSP 2021 - 46th International Conference on Acoustics, Speech, and Signal Processing. Toronto / Virtual, Canada, 6th June 2021. DOI: 10.1109/ICASSP39728.2021.9414758. URL: https://hal.archives-ouvertes.fr/hal-02985794.

[35] A. Geet D'Sa, I. Illina, D. Fohr, D. Klakow and D. Ruiter. 'Exploring Conditional Language Model Based Data Augmentation Approaches For Hate Speech Classification'. In: TSD 2021 - 24th International Conference on Text, Speech and Dialogue. Olomouc, Czech Republic, 6th Sept. 2021. URL: https://hal.inria.fr/hal-03244472.

[36] F. Gontier, R. Serizel and C. Cerisara. 'Automated audio captioning by fine-tuning bart with audioset tags'. In: DCASE 2021 - 6th Workshop on Detection and Classification of Acoustic Scenes and Events. Virtual, Spain, 15th Nov. 2021. URL: https://hal.inria.fr/hal-03522488.

[37] P.-A. Grumiaux, S. Kitić, P. Srivastava, L. Girin and A. Guérin. 'Saladnet: Self-Attentive Multisource Localization in the Ambisonics Domain'. In: *WASPAA 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. WASPAA 2021 - IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz / Virtual, United States: IEEE, 17th Oct. 2021, pp. 336–340. DOI: 10.1109/WASPAA52581.2021.9632737. URL: https://hal.archives-ouvertes.fr/hal-03537340.

[38] I. Illina and D. Fohr. 'DNN-based semantic rescoring models for speech recognition'. In: TSD 2021 - 24th International Conference on Text, Speech and Dialogue. proceedings of TSD 2021. Olomouc, Czech Republic, 6th Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03239211.

[39] O. Janin and A. Piquard-Kipffer. 'De codes gestuo-manuels à la Langue des Signes Française : usages et enjeux à la maternelle dans le cadre des gestes professionnels inclusifs et des adaptations didactiques'. In: IDEKI 2021 - 4ème colloque international Didactiques et métiers de l'humain. Pont-A-Mousson, France, 2nd Dec. 2021. URL: https://hal.inria.fr/hal-03130603.

[40] Z. Kang, R. Horaud and M. Sadeghi. 'Robust Face Frontalization For Visual Speech Recognition'. In: ICCVW 2021 - International Conference on Computer Vision Workshops. Montreal - Virtual, Canada: IEEE, 11th Oct. 2021, pp. 2485–2495. DOI: 10.1109/ICCVW54120.2021.00281. URL: https://hal.inria.fr/hal-03326002.

[41] T. Kinnunen, A. Nautsch, M. Sahidullah, N. Evans, X. Wang, M. Todisco, H. Delgado, J. Yamagishi and L. Kong Aik. 'Visualizing Classifier Adjacency Relations: A Case Study in Speaker Verification and Voice Anti-Spoofing'. In: INTERSPEECH 2021. Brno, Czech Republic, 30th Aug. 2021. DOI: 10.21437/Interspeech.2021-1522. URL: https://hal.archives-ouvertes.fr/hal-03261467.

[42] A. Kulkarni, V. Colotte and D. Jouvet. 'Improving transfer of expressivity for end-to-end multispeaker text-to-speech synthesis'. In: EUSIPCO 2021 - 29th European Signal Processing Conference. Dublin / Virtual, Ireland, 23rd Aug. 2021. DOI: 10.23919/EUSIPCO54536.2021.9616249. URL: https://hal.archives-ouvertes.fr/hal-02978485.

[43] M. Leitao, E. Venti, T. Sigiez, C. Laroche, M. Perini and A. Piquard-Kipffer. 'Projet LogilecSur : quelles stratégies enseignantes pour guider des élèves sourds vers l'autonomie en compréhension écrite ?' In: IDEKI 2021 - 4ème colloque international Didactiques et métiers de l'humain. Pont-à-Mousson, France, 2nd Dec. 2021. URL: https://hal.inria.fr/hal-03482203.

[44] X. Liu, M. Sahidullah and T. Kinnunen. 'Learnable MFCCs for Speaker Verification'. In: ISCAS 2021 - IEEE International Symposium on Circuits and Systems. Daegu, South Korea, 22nd May 2021. DOI: 10.1109/ISCAS51556.2021.9401593. URL: https://hal.archives-ouvertes.fr/hal-03139532.

[45] X. Liu, M. Sahidullah and T. Kinnunen. 'Optimized Power Normalized Cepstral Coefficients Towards Robust Deep Speaker Verification'. In: ASRU 2021 - IEEE Automatic Speech Recognition and Understanding Workshop. Cartagena, Colombia, 13th Dec. 2021. URL: https://hal.inria.fr/hal-03359173.

[46] X. Liu, M. Sahidullah and T. Kinnunen. 'Parameterized Channel Normalization for Far-field Deep Speaker Verification'. In: ASRU 2021 - IEEE Automatic Speech Recognition and Understanding Workshop. Cartagena, Colombia, 13th Dec. 2021. URL: https://hal.inria.fr/hal-03359174.

[47] M. Mohammadamini, D. Matrouf, J.-F. Bonastre, R. Serizel, S. Dowerah and D. Jouvet. 'Compensate multiple distortions for speaker recognition systems'. In: EUSIPCO 2021 - 29th European Signal Processing Conference. Dublin / Virtual, Ireland, 23rd Aug. 2021. DOI: 10.23919/EUSIPCO54536.2021.9615983. URL: https://hal.archives-ouvertes.fr/hal-03224675.

[48] V.-N. Nguyen, M. Sadeghi, E. Ricci and X. Alameda-Pineda. 'Deep Variational Generative Models for Audio-visual Speech Separation'. In: MLSP 2021 - IEEE International Workshop on Machine Learning for Signal Processing. Gold Coast, Australia, Oct. 2021. URL: https://hal.inria.fr/hal-02930662.

[49] H. Nourtel, P. Champion, D. Jouvet, A. Larcher and M. Tahon. 'Evaluation of Speaker Anonymization on Emotional Speech'. In: SPSC 2021 - 1st ISCA Symposium on Security and Privacy in Speech Communication. Virtual, Germany, 10th Nov. 2021. URL: https://hal.inria.fr/hal-03377797.

[50] M. Olvera, E. Vincent and G. Gasso. 'Improving Sound Event Detection with Auxiliary Foreground-Background Classification and Domain Adaptation'. In: DCASE 2021 - 6th Workshop on Detection and Classification of Acoustic Scenes and Events. Virtual, Spain, 15th Nov. 2021. URL: https://hal.inria.fr/hal-03387778.

[51] M. Olvera, E. Vincent, R. Serizel and G. Gasso. 'Foreground-Background Ambient Sound Scene Separation'. In: EUSIPCO 2020 - 28th European Signal Processing Conference. Amsterdam / Virtual, Netherlands, 18th Jan. 2021. DOI: 10.23919/Eusipco47968.2020.9287436. URL: https://hal.archives-ouvertes.fr/hal-02567542.

[52] V. Ribeiro, K. Isaieva, J. Leclère, P.-A. Vuissoz and Y. Laprie. 'Towards the prediction of the vocal tract shape from the sequence of phonemes to be articulated'. In: INTERSPEECH 2021. Brno, Czech Republic, 30th Aug. 2021. DOI: 10.21437/Interspeech.2021-184. URL: https://hal.inria.fr/hal-03360113.

[53] F. Ronchini, R. Serizel, N. Turpault and S. Cornell. 'The impact of non-target events in synthetic soundscapes for sound event detection'. In: DCASE 2021 - Detection and Classification of Acoustic Scenes and Events. Barcelona/Virtual, Spain, 15th Nov. 2021. URL: https://hal.inria.fr/hal-03355184.

[54] M. Sadeghi and X. Alameda-Pineda. 'Switching Variational Auto-Encoders for Noise-Agnostic Audio-visual Speech Enhancement'. In: ICASSP 2021 - 46th International Conference on Acoustics, Speech, and Signal Processing. Toronto / Virtual, Canada: IEEE, 6th June 2021, pp. 1–5. DOI: 10.1109/ICASSP39728.2021.9414097. URL: https://hal.inria.fr/hal-03155445.

[55] M. Sahidullah, A. Kumar Sarkar, V. Vestman, X. Liu, R. Serizel, T. Kinnunen, Z.-H. Tan and E. Vincent. 'UIAI System for Short-Duration Speaker Verification Challenge 2020'. In: SLT 2021 - IEEE Spoken Language Technology Workshop. Shenzhen / Virtual, China, Jan. 2021. DOI: 10.1109/SLT48900.2021.9383596. URL: https://hal.archives-ouvertes.fr/hal-02907037.

[56] U. Saqib, A. Deleforge and J. R. Jensen. 'Detecting acoustic reflectors using a robot's ego-noise'. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ICASSP 2021 - 46th International Conference on Acoustics, Speech, and Signal Processing. Toronto / Virtual, Canada, 6th June 2021. DOI: 10.1109/ICASSP39728.2021.9414061. URL: https://hal.archives-ouvertes.fr/hal-03430276.

[57] S. A. Sheikh, M. Sahidullah, F. Hirsch and S. Ouni. 'StutterNet: Stuttering Detection Using Time Delay Neural Network'. In: EUSIPCO 2021 - 29th European Signal Processing Conference. Dublin / Virtual, Ireland, 23rd Aug. 2021. DOI: 10.23919/EUSIPCO54536.2021.9616063. URL: https://hal.inria.fr/hal-03227223.

[58] P. Singh, G. Saha and M. Sahidullah. 'Deep scattering network for speech emotion recognition'. In: EUSIPCO 2021 - 29th European Signal Processing Conference. Dublin / Virtual, Ireland, 23rd Aug. 2021. DOI: 10.23919/EUSIPCO54536.2021.9615958. URL: https://hal.archives-ouvertes.fr/hal-03218278.

[59] P. Singh, G. Saha and M. Sahidullah. 'Non-linear frequency warping using constant-Q transformation for speech emotion recognition'. In: ICCCI 2021 - International Conference on Computer Communication and Informatics. Coimbatore, India, Jan. 2021. DOI: 10.1109/ICCCI50826.2021.9402569. URL: https://hal.archives-ouvertes.fr/hal-03134015.

[60] S. Sivasankaran, E. Vincent and D. Fohr. 'Analyzing the impact of speaker localization errors on speech separation for automatic speech recognition'. In: EUSIPCO 2020 - 28th European Signal Processing Conference. Amsterdam / Virtual, Netherlands, 18th Jan. 2021. DOI: 10.23919/Eusipco47968.2020.9287541. URL: https://hal.inria.fr/hal-02355669.

[61] S. Sivasankaran, E. Vincent and D. Fohr. 'Explaining deep learning models for speech enhancement'. In: INTERSPEECH 2021. Brno, Czech Republic, 28th Aug. 2021. DOI: 10.21437/Interspeech.2021-1764. URL: https://hal.inria.fr/hal-03257450.

[62] P. Srivastava, A. Deleforge and E. Vincent. 'Blind room parameter estimation using multiple multichannel speech recordings'. In: *WASPAA 2021- IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. WASPAA 2021 - IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz, NY, United States, 17th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03304656.

[63] N. Turpault, R. Serizel, S. Wisdom, H. Erdogan, J. R. Hershey, E. Fonseca, P. Seetharaman and J. Salamon. 'Sound Event Detection and Separation: a Benchmark on Desed Synthetic Soundscapes'. In: ICASSP 2021 - 46th International Conference on Acoustics, Speech, and Signal Processing. Toronto/Virtual, Canada, 6th June 2021. DOI: 10.1109/ICASSP39728.2021.9414789. URL: https://hal.inria.fr/hal-02984675.

[64] S. Wisdom, H. Erdogan, D. P. W. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman and J. R. Hershey. 'What's All the FUSS About Free Universal Sound Separation Data?' In: ICASSP 2021 - 46th International Conference on Acoustics, Speech, and Signal Processing. Toronto/Virtual, Canada, 6th June 2021. DOI: 10.1109/ICASSP39728.2021.9414774. URL: https://hal.inria.fr/hal-02984693.

[65] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans and H. Delgado. 'ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection'. In: ASVspoof 2021 Workshop - Automatic Speaker Verification and Spoofing Coutermeasures Challenge. Virtual, France, 16th Sept. 2021. URL: https://hal.inria.fr/hal-03360794.

[66] N. Zampieri, I. Illina and D. Fohr. 'Multiword Expression Features for Automatic Hate Speech Detection'. In: NLDB 2021 - 26th International Conference on Natural Language & Information Systems. Vol. 12801. Natural Language Processing and Information Systems. Saarbrücken/Virtual, Germany, 23rd June 2021. URL: https://hal.archives-ouvertes.fr/hal-03231047.

[67] G. Zervakis, E. Vincent, M. Couceiro and M. Schoenauer. 'On Refining BERT Contextualized Embeddings using Semantic Lexicons'. In: ECML PKDD 2021 - Machine Learning with Symbolic Methods and Knowledge Graphs co-located with European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. http://ceur-ws.org/Vol-2997/paper4.pdf. Online, Spain, 1st Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03318571.

Conferences without proceedings

[68] A. Bonneau. 'Assimilations de voisement et interférences français/allemand'. In: RéaL2 2021 - Colloque International du Réseau d'Acquisition des Langues Secondes. Toulouse, France, 5th July 2021. URL: https://hal.inria.fr/hal-03353153.

[69] K. A. Kumar, S. Waldekar, G. Saha and M. Sahidullah. 'Domain-Dependent Speaker Diarization for the Third DIHARD Challenge'. In: DIHARD 2021 - 3rd Speech Diarization Challenge Workshop. Virtual, France, 23rd Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-03117843.

[70] N. Zampieri, I. Illina and D. Fohr. 'A comparative study of different features for efficient automatic hate speech detection'. In: IPrA 2021 - 17th International Pragmatics Conference. Winterthur, Switzerland, 27th June 2021. URL: https://hal.archives-ouvertes.fr/hal-03115781.

[71] N. Zampieri, I. Illina and D. Fohr. 'A comparative study of different state-of-the-art NLP models for efficient automatic hate speech detection'. In: Comments, hate speech, disinformation and public communication regulation 2021. Zagreb, Croatia, 16th Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03347244.

Scientific book chapters

[72] B. Elie, C. Fauth and M. Barkat-Defradas. 'Histoire des machines parlantes'. In: *HISTOIRE DE LA DESCRIPTION DE LA PAROLE : DE L'INTROSPECTON À L'INSTRUMENTATION*. Honoré Champion, 2021. URL: https://hal.archives-ouvertes.fr/hal-03317598.

Doctoral dissertations and habilitation theses

[73] T. Biasutto-Lervat. 'Multimodal Coarticulation Modeling : Towards the animation of an intelligible talking head'. Université de Lorraine, 29th Jan. 2021. URL: https://hal.univ-lorraine.fr/tel-03203815.

[74] R. Duroselle. 'Robustness of language recognition system to transmission channel'. Université de Lorraine (ENIM, L-INP) / Université Paris Nanterre (ED 138 EA 369 CRIIA REDESC), 28th Oct. 2021. URL: https://hal.archives-ouvertes.fr/tel-03546267.

[75] M. Pariente. 'Implicit and explicit phase modeling in deep learning-based source separation'. Université de Lorraine, 29th Sept. 2021. URL: https://hal.univ-lorraine.fr/tel-03395953.

[76] B. M. L. Srivastava. 'Speaker Anonymization: Representation, Evaluation and Formal Guarantees'. Inria Lille Nord Europe - Laboratoire CRIStAL - Université de Lille, 2nd Dec. 2021. URL: https://hal.inria.fr/tel-03539738.

[77] N. Turpault. 'Analysis of scientific challenges in ambient sound recognition in real environments'. Université de Lorraine, 31st May 2021. URL: https://hal.inria.fr/tel-03304880.

Reports & preprints

[78] D. Di Carlo, P. Tandeitnik, C. Foy, N. Bertin, A. Deleforge and S. Gannot. *dEchorate: a Calibrated Room Impulse Response Dataset for Echo-aware Signal Processing*. 17th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03207860.

[79] S. Dowerah, R. Serizel, D. Jouvet, M. Mohammadamini and D. Matrouf. *MULTICHANNEL SPEECH ENHANCEMENT FOR SPEAKER VERIFICATION IN NOISY AND REVERBERANT ENVIRONMENTS*. Singapore, Singapore, 17th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03487420.

[80] K. A. Kumar, S. Waldekar, G. Saha and M. Sahidullah. *ABSP System for The Third DIHARD Challenge*. 4th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03130955.

[81] M. Maouche, B. M. L. Srivastava, N. Vauquier, A. Bellet, M. Tommasi and E. Vincent. *Enhancing Speech Privacy with Slicing*. 7th Oct. 2021. URL: https://hal.inria.fr/hal-03369137.

[82] I. A. Sheikh, E. Vincent and I. Illina. *Training RNN Language Models on Uncertain ASR Hypotheses in Limited Data Scenarios*. 27th Aug. 2021. URL: https://hal.inria.fr/hal-03327306.

[83] I. A. Sheikh, E. Vincent and I. Illina. *Transformer versus LSTM Language Models Trained on Uncertain ASR Hypotheses in Limited Data Scenarios*. 2nd Oct. 2021. URL: https://hal.inria.fr/hal-03362828.

[84] 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*. 28th Dec. 2021. URL: https://hal.inria.fr/hal-03197376.

[85] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'brien, A. Chanclu, J.-F. Bonastre, M. Todisco and M. Maouche. *Supplementary material to the paper The VoicePrivacy 2020 Challenge: Results and findings*. 20th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03335126.

[86] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'brien, A. Chanclu, J.-F. Bonastre, M. Todisco and M. Maouche. *The VoicePrivacy 2020 Challenge: Results and findings*. 19th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03332224.

[87] M. A. T. Turan, D. Klakow, E. Vincent and D. Jouvet. *Adapting Language Models When Training on Privacy-Transformed Data*. Brno, Czech Republic, 2nd Apr. 2021. URL: https://hal.inria.fr/hal-03189354.

[88] N. Turpault, R. Serizel and E. Vincent. *Analysis of weak labels for sound event tagging*. 21st Apr. 2021. URL: https://hal.inria.fr/hal-03203692.

12.3 Other

Scientific popularization

[89] A. Campbell, T. Kleinbauer, M. Tommasi and E. Vincent. 'Enabling voice-based apps with European values'. In: *ERCIM News* 126 (9th July 2021), pp. 38–39. URL: https://hal.inria.fr/hal-03476390.

Patents

[90] S. Ouni, T. Biasutto–Lervat and S. Dahmani. 'Audio-driven speech animation using recurrent neutral network'. WO2021023861 (United States). 11th Feb. 2021. URL: https://hal.inria.fr/hal-03167213.

12.4 Cited publications

[91] A. A. Nugraha, A. Liutkus and E. Vincent. 'Multichannel audio source separation with deep neural networks'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 24.10 (June 2016), pp. 1652–1664. DOI: 10.1109/TASLP.2016.2580946. URL: https://hal.inria.fr/hal-01163369.

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

[93] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker and R. Marxer. 'An analysis of environment, microphone and data simulation mismatches in robust speech recognition'. In: *Computer Speech and Language* 46 (July 2017), pp. 535–557. DOI: 10.1016/j.csl.2016.11.005. URL: https://hal.inria.fr/hal-01399180.