RESEARCH CENTRE

Nancy - Grand Est

IN PARTNERSHIP WITH:

CNRS, Université de Lorraine

2020

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 Team: 2014 July 01, updated into Project-Team: 2015 July 01

Keywords

Computer sciences and digital sciences

A3.4.6. – Neural networks

A3.4.8. – Deep learning

A3.5. – Social networks

A4.8. – Privacy-enhancing technologies

A5.1.7. – Multimodal interfaces

A5.7.1. – Sound

A5.7.3. – Speech

A5.7.4. – Analysis

A5.7.5. – Synthesis

A5.8. – Natural language 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]
- Mostafa Sadeghi [Inria, from Nov 2020, Starting Faculty Position]
- Md Sahidullah [Inria, Starting Research Position]
- Emmanuel Vincent [Inria, Senior Researcher, HDR]

Faculty Members

- Vincent Colotte [Univ de Lorraine, Associate Professor]
- Irène Illina [Univ de Lorraine, Associate Professor, HDR]
- Odile Mella [Univ de Lorraine, Associate Professor, until Feb 2020]
- Slim Ouni [Univ de Lorraine, Associate Professor, HDR]
- Agnes Piquard-Kipffer [Univ de Lorraine, Associate Professor]
- Romain Serizel [Univ de Lorraine, Associate Professor]

Post-Doctoral Fellows

- Elodie Gauthier [Univ de Lorraine, until Jan 2020]
- Manfred Pastätter [Inria, until Feb 2020]
- Imran Sheikh [Inria]

PhD Students

- Théo Biasutto-Lervat [Univ de Lorraine, until Nov 2020]
- Tulika Bose [Univ de Lorraine]
- Guillaume Carbajal [Invoxia, CIFRE, until Mar 2020]
- Pierre Champion [Inria]
- Sara Dahmani [Univ de Lorraine, until Mar 2020]
- Diego Di Carlo [Inria, until Sep 2020]
- Stephane Dilungana [Inria, from Oct 2020]
- Ioannis Douros [Univ de Lorraine, until Jul 2020]
- Sandipana Dowerah [Inria]
- Ashwin Geet Dsa [Univ de Lorraine]

- Adrien Dufraux [Facebook, CIFRE]
- Raphael Duroselle [Ministère des armées]
- Nicolas Furnon [Univ de Lorraine]
- Amal Houidhek [Ecole Nationale d'Ingénieurs de Tunis - Tunisia, until Feb 2020]
- Ajinkya Kulkarni [Univ de Lorraine]
- Lou Lee [Univ de Lorraine]
- Xuechen Liu [Inria, from Mar 2020]
- Mohamed Amine Menacer [Univ de Lorraine]
- Mauricio Michel Olvera Zambrano [Inria]
- Manuel Pariente [Univ de Lorraine]
- Shakeel Ahmad Sheikh [Univ de Lorraine]
- Sunit Sivasankaran [Inria, until Sep 2020]
- Vinicius Souza Ribeiro [Univ de Lorraine, from Oct 2020]
- Prerak Srivastava [Inria, from Oct 2020]
- Nicolas Turpault [Inria]
- Nicolas Zampieri [Inria]
- Georgios Zervakis [Inria]

Technical Staff

- Ismaël Bada [CNRS until Sep 2020, then Univ de Lorraine, Engineer]
- Akira Campbell [Inria, Engineer, from Nov 2020]
- Zaineb Chelly Dagdia [Inria, Engineer, until Aug 2020]
- Joris Cosentino [Inria, Engineer, from Nov 2020]
- Louis Delebecque [Inria, Engineer]
- Valérian Girard [Inria, Engineer, until Jun 2020]
- Seyed Ahmad Hosseini [Inria, Engineer]
- Mathieu Hu [Inria, Engineer]
- Krist Kostallari [Inria, Engineer, from Apr 2020 until Jun 2020]
- Stéphane Level [CNRS, Engineer, until Mar 2020]
- Léon Rohrbacher [Univ de Lorraine, Engineer, until Oct 2020]
- Francesca Ronchini [Inria, Engineer, from Dec 2020]
- Mehmet Ali Tugtekin Turan [Inria, Engineer]

Interns and Apprentices

- Tess Boivin [NXP, from Mar 2020 until Aug 2020]
- Clement Brifault [Univ de Lorraine, from Jun 2020 until Sep 2020]
- Joris Cosentino [Inria, from Feb 2020 until Aug 2020]
- Alexis Dieu [Inria, from Mar 2020 until Aug 2020]
- Anastasiia Karliuk [Univ de Lorraine, from Feb 2020 until Jul 2020]
- Maxence Naud [Univ de Lorraine, from May 2020 until Aug 2020]
- Flavie Oesch [Univ de Lorraine, from Sep 2020 until Oct 2020]
- Thierry Paulin [Ecole de l'aménagement durable des territoires, from Mar 2020 until Jul 2020]
- Anne Sancier [Univ de Lorraine, from Sep 2020 until Oct 2020]
- Stephanie Stoll [Univ de Lorraine, from Jun 2020 until Oct 2020]
- Ruoxiao Yang [Univ de Lorraine, from Mar 2020 until Aug 2020]

Administrative Assistants

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

Visiting Scientist

- Brij Mohan Lal Srivastava [Univ de Lille, until Aug 2020]

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 relied 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.

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

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.

4 Application domains

Approaches and models developed in the MULTISPEECH project 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-founded and 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

Asteroid, our Python toolbox for audio source separation and speech enhancement was released in May 2020 [52]. It has received more than 700 Github stars since then. Using this toolbox, Manuel Pariente and Michel Olvera won the first place in the PyTorch Summer Hackathon 2020[2] with DeMask, a method to enhance speech spoken by talkers wearing face masks.

Emmanuel Vincent co-organized the first VoicePrivacy Challenge [62].

The project Audio Cockpit Denoising for voice Command from the Man Machine Teaming initiative has been selected for presentation at the "Forum Innovation Défence" and to Florence Parly, Minister of the Armed Forces.

6.1 Awards

We participated in the Oriental Language Recognition Challenge (OLR 2020). The system we have proposed has been ranked in first position (among the systems proposed by about 20 teams) for the two tasks in which we participated: Task 1 on cross-channel language identification, and Task 3 on noisy data language identification.

7 New software and platforms

7.1 New software

7.1.1 COMPRISE Voice Transformer

Name: COMPRISE Voice Transformer

[2] https://pytorch.org/blog/announcing-the-winners-of-the-2020-global-pytorch-summer-hackathon/

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.

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.

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 Kaldi-web

Name: Kaldi-web

Keyword: Speech recognition

Functional Description: Today, developers willing to implement a voice interface must either rely on proprietary software or become experts in speech recognition. Conversely, researchers in speech recognition wishing to demonstrate their results need to be familiar with other technologies (e.g., graphical user interfaces). Kaldi-web is an open-source, cross-platform tool which bridges this gap by providing a user interface built around the online decoder of the Kaldi toolkit. Additionally, because we compile Kaldi to Web Assembly, speech recognition is performed directly in web browsers. This addresses privacy issues as no data is transmitted to the network for speech recognition.

URL: https://gitlab.inria.fr/kaldi.web/

Contact: Denis Jouvet

Participants: Mathieu Hu, Laurent Pierron, Denis Jouvet, Emmanuel Vincent

7.1.4 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.

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

7.1.5 DNNuep

Name: DNN uncertainty estimation and propagation

Keyword: Speech recognition

Functional Description: From a noisy signal and its noiseless version, the system estimates the uncertainty and propagates it in an automatic speech recognition system.

News of the Year: Development of version 1.0 of the software. Propagation of uncertainty with two methods and implementation in Kaldi. Performance evaluation on a noisy speech corpus.

Contact: Irina Illina

Participants: Irina Illina, Dominique Fohr, Ismaël Bada

7.1.6 DNNsem

Keywords: Speech recognition, Semantic

Functional Description: DNNsem is a software for rescoring the N-best list of hypotheses of an automatic speech recognition (ASR) system. It is useful in the situation when the training and testing conditions differ due to noise or other acoustic distortions. To improve performance in such conditions, DNNsem accounts for long-term semantic relations. It takes as inputs the list of N-best ASR hypotheses and the corresponding acoustic scores, and it outputs a reranked list by favoring words that better correspond to the semantic context of the sentence. To do so, it uses two continuous word representations: word2vec or FastText.

News of the Year: Development of version 1.0 of the software. Implementation of two methods, based on static word embeddings (word2vec) or dynamic word embeddings (BERT). Performance evaluation on a noisy speech corpus.

Contact: Irina Illina

Participants: Irina Illina, Dominique Fohr

7.1.7 Web-based Pronunciation Learning Application

Keywords: Pronunciation training, Talking head, Second language learning

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.

Contact: Slim Ouni

Participants: Thomas Girod, Leon Rohrbacher, Slim Ouni, Denis Jouvet

7.1.8 Grapheme-phoneme aligner

Keywords: Grapheme-to-phoneme converter, Grapheme-phoneme alignment

Functional Description: This software processes French words or sentences to determine their pronunciation, and to provide the association between letters and sounds. It calls SOJA to preprocess the text, and LORIA-PHON to determine the pronunciation of the words. It then aligns, through a set of rules, the letters of the text with the phonemes of the predicted pronunciation.

Contact: Vincent Colotte

Participants: Vincent Colotte, Louis Delebecque, Denis Jouvet

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, Manuel Pariente, Akira Campbell, Zaineb Chelly Dagdia, Mehmet Ali Tuğtekin Turan, Georgios Zervakis.

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 end-to-end neural networks including learned real-valued filterbanks. We tackled two limitations of this approach. First, to ensure that the representation properly encodes phase properties as the short time Fourier transform and other conventional time-frequency transforms, we designed complex-valued analytic learned filterbanks and defined corresponding representations and masking strategies which outperformed the popular ConvTasNet algorithm [53]. This advance formed the basis for the Asteroid toolbox [52] which provides various choices of filterbanks, network architectures and loss functions, as well as training and evaluation tools and recipes for several datasets. Asteroid performs on par with or better than the results reported in reference papers, and it has received more than 700 Github stars since its release in May 2020. Second, in order to allow generalization to mixtures of sources not seen together in training, we pursued the modeling of speech signals by variational autoencoders (VAEs), which are a variant of the probabilistic generative models classically used in source separation before the deep learning era. We extended the model developed last year for magnitude spectra to complex-valued spectra.

8.1.2 Learning from little/no labeled data

Unsupervised or semi-supervised acoustic modeling ASR systems are typically trained in a supervised fashion using manually labeled data. To reduce the cost of labeling, we investigated semi-supervised training of acoustic models in practical scenarios with a limited amount of labeled in-domain data [56]. We proposed an error detection driven semi-supervised training approach, in which an error detector controls the hypothesized transcriptions or lattices used as training targets on additional unlabeled data, and achieved word error recovery rates of 28 to 89%. We also studied the recognition of accented speech, where the accented training data is unlabeled [64]. To do so, we computed xvector-like accent embeddings and used them as auxiliary inputs to an acoustic model trained on native data only. We achieved a 15% relative word error rate reduction on accented speech w.r.t. acoustic models trained with regular spectral features only, and an additional 15% relative reduction by semi-supervised training using 1 hour of untranscribed speech per accent only.

Transfer learning applied to speech synthesis We worked on the disentanglement of speaker, emotion and content in the acoustic domain for transferring expressivity information from one speaker to another one, particularly when only neutral speech data is available for the latter. We have proposed an approach relying on multiclass N-pair based deep metric learning in a recurrent conditional variational autoencoder (RCVAE) for implementing a multispeaker expressive text-to-speech system. The proposed approach conditions the text-to-speech system on speaker embeddings, and leads to a clustering with respect to emotion in a latent space. 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 [43]. The approach has then been applied using an inverse autoregressive flow as a way to perform the variational inference [44], and more recently using an end-to-end text-to-speech synthesis system based on Tacotron 2 [92].

8.1.3 Preserving privacy

Speech signals involve a lot of private information. With a few minutes of data, the speaker identity can be modeled for malicious purposes like voice cloning, spoofing, etc. To reduce this risk, we investigated speaker anonymization strategies based on voice conversion. In contrast to prior evaluations, we argue that different types of attackers can be defined depending on the extent of their knowledge. We compared three simple conversion methods in three attack scenarios, and showed that these methods fail to protect against an attacker that has extensive knowledge of the type of conversion and how it has been applied, but may provide some protection against less knowledgeable attackers [61]. We then developed a more advanced conversion method and explored several design choices for the distance metric between the source and target speakers, the region of x-vector space where the target speaker is picked, and gender selection to find the optimal combination of design choices in terms of privacy and utility [60]. The resulting software served as a baseline for the 1st Voice Privacy Challenge [62]. We have investiagted the modification of the fundamental frequency to improve consistency with the selected target x-speaker [89, 23]. We also conducted a comparative study of speech anonymization metrics from a theoretical and experimental point of view [49].

8.2 Speech production and perception

Participants Anne Bonneau, Dominique Fohr, Denis Jouvet, Yves Laprie, Vincent Colotte, Slim Ouni, Agnes Piquard-Kipffer, Elodie Gauthier, Manfred Pastatter, Théo Biasutto-Lervat, Sara Dahmani, Ioannis Douros, Amal Houidhek, Lou Lee, Shakeel Ahmad Sheikh, Vinicius Souza Ribeiro, Louis Delebecque, Valérian Girard, Seyed Ahmad Hosseini, Mathieu Hu, Leon Rohrbacher.

8.2.1 Articulatory modeling

Exploitation of dynamic MR images Magnetic resonance imaging (MRI) has been used to study the movement of the tongue tip which is involved in the production of dental consonants. We evaluated its velocity using two independent approaches [13]. The first one consists in acquisition with a real-time technique in the mid-sagittal plane. Tracking of the tongue tip manually and with a computer vision method allows its trajectory to be found and the velocity to be calculated. The second approach - phase contrast MRI - enables velocities of the moving tissues to be measured directly. Evaluation on data from two French-speaking subjects articulating /tata/ shows that both methods are in qualitative agreement and consistent with other techniques used for evaluation of the tongue tip velocity.

Tongue contour extraction from real-time MRI is a nontrivial task due to the presence of artifacts (as blurring or ghostly contours). In this work, the automatic tongue delineation is achieved by means of a U-Net auto-encoder convolutional neural network. We particularly investigated both intra- and inter-subject validation using real-time MRI and manually annotated 1-pixel wide contours. Predicted probability maps were post-processed in order to obtain 1-pixel wide tongue contours. The results are very good and slightly outperform published results on automatic tongue segmentation [14].

We investigated the creation a 3D dynamic atlas of the vocal tract that captures the dynamics of the articulators in all three dimensions in order to create a generic speaker model. The core steps of the proposed method are temporal alignment of the real-time MRI acquired in several sagittal planes and their combination with adaptive kernel regression [31, 32]. As a preprocessing step, a reference space is created and used to remove anatomical speaker specificities, thus keeping only the variability in speech production for the construction of the atlas [34, 33]. The adaptive kernel regression addresses the choice of atlas time points independently of the time points of the frames that are used as an input for the atlas construction. The evaluation with data from two new speakers showed that the use of the atlas helps in reducing subject variability, can capture the dynamic behavior of the articulators and is able to generalize the speech production process by creating a universal-speaker reference space.

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 a Recurrent Neural Network (RNN), more specifically Gated Recurrent Units (GRU), 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. 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 RNNs 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.

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 started working on existing stuttering acoustic speech datasets. We proposed to use a Time Delay Neural Network (TDNN) model for stuttering identification which takes into consideration the temporal evolution of the acoustic signal of the stuttered speech. We have also started collecting French audiovisual data of subjects who stutter. However the current sanitary context has slowed down this procedure. We are working on alternative remote recording protocols.

8.2.2 Multimodal expressive speech

Arabic speech synthesis We have continued working on Modern Standard Arabic text-to-speech synthesis with ENIT (École Nationale d'Ingénieurs de Tunis, Tunisia), using HMM and neural network based approaches [87]. We have also investigated deep learning modeling of the sound durations for Arabic speech synthesis taking into account specificities of the Arabic language such as vowel quantity and gemination [20].

Expressive audiovisual synthesis After having acquired a high quality expressive audio-visual corpus based on fine linguistic analysis, motion capture, and naturalistic acting techniques, we have analyzed, processed, and phonetically aligned it with speech [9, 70]. We used conditional variational autoencoders (CVAE) to generate the duration, acoustic and visual aspects of speech. The emotion clusters in the latent space were clearly distinguishable although the training was carried on without using emotion labels. Perceptual experiments have confirmed the capacity of our system to generate recognizable emotions. Moreover, the generative nature of the CVAE allowed us to generate well-perceived nuances of the six emotions and to blend different emotions together. The PhD thesis related to these works has been defended [85].

8.2.3 Categorization of sounds and prosody

Non-native speech production The voicing contrast is realized differently in German and in French, either in the phonetic dimension or in the phonological one, and voicing assimilations appear in opposite direction in these two languages (regressive assimilation in French, progressive in German). We have designed a corpus devoted to the analysis of assimilations made by French people learning German, and the determination of possible links between various aspects of German voicing mastery in French/German productions. We have recorded, segmented and analysed 20 French people learning German.

A corpus of a series of German fricatives have been designed and recorded with the articulograph of the laboratory by four speakers (one German and three French speakers).

Language and reading acquisition by children having some language impairments We continued investigating the acquisition of language by hard of hearing children via cued speech. In cooperation with DevAH-EA3450 (Univ de Lorraine), we have devised a protocol to examine the use of a digital book and of a children's picture book for hard-of-hearing children in order to compare scaffolding by the speech therapist or the teacher in these two situations. We also questioned nearly two thousand kindergarten teachers regarding their use of visual language encoding gestures strengthening spoken French Language. The 493 answers received show that teachers use both gestures, French Sign Language or Signed Supported French, with children who don't have hearing loss more than with deaf children with the aim of developing a better communicational base.

We started examining the intelligibility of the talking head developed by MULTISPEECH. We undertook a scoping study comparing three different modalities: a talking head, a human speaker and a strictly auditory modality. First qualitative results from 8 children with deafnes showed that the avatar, which provides additional visual cues, allows for faster and better understanding of sentences, and was most appreciated by the children.

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 web application is almost finished and will be used by teachers to prepare pronunciation lessons, and by secondary school students learning German. The analysis of student pronunciation is still not completed, and more development will be continued.

The ALOE project dealt with children learning to read. In this project, we were involved with tutoring aspects based on a talking head, and with grapheme-to-phoneme conversion which is a critical tool for the development of the digitized version of ALOE reading learning tools (tools which were previously developed and offered only in a paper form). We have developed a text coder, which predicts the pronunciation of French sentences and returns the alignment between the letters and the sounds.

Prosody Prosodic correlates of a few discourse particles have been investigated further. In particular prosodic correlates of pragmatic functions have been compared accross languages (French and English) on prepared speech [72], and accross various speech styles [45].

8.3 Speech in its environment

Participants Emmanuel Vincent, Denis Jouvet, Antoine Deleforge, Dominique Fohr, Mostafa Sadeghi, Md Sahidullah, Irène Illina, Odile Mella, Romain Serizel, Tulika Bose, Guillaume Carbajal, Diego Di Carlo, Stephane Dilungana, Sandipana Dowerah, Ashwin Geet Dsa, Raphaël Duroselle, Nicolas Furnon, Xuechen Liu, Mohamed Amine Menacer, Mauricio Michel Olvera Zambrano, Sunit Sivasankaran, Prerak Srivastava, Nicolas Turpault, Nicolas Zampieri, Ismaël Bada, Joris Cosentino, Louis Delebecque, Mathieu Hu, Stephane Level, Krist Kostallari, Francesca Ronchini.

8.3.1 Acoustic environment analysis

Sound event detection is the task of finding what sound event occurred in a recording and when. As it is prohibitive to get a large dataset with so-called strong labeled soundcases (i.e., with onset and offset timestamps), one alternative is to rely on so-called weakly labeled soundscapes (i.e., without timestamps) that are considerably cheaper to obtain. We explored the limitations introduced by relying only on such weak labels [66]. Another alternative would be to generate synthetic soundscapes for which strong annotations are then cheap to obtain but a the cost of possible domain mismatch with recorded evaluation data. We studied the impact of training a sound event detection system using a heterogeneous dataset (including both recorded and synthetic soundscapes) and different label granularity (strong, weak) [65]. 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 on such complex soundscapes [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) 2020 Challenge [65, 67] and published a detailed analysis of the submissions to the previous iteration of this task in 2019 [55]. In 2020, the task still focused on the problem of learning from audio segments that are either weakly labeled or unlabeled, targeting domestic applications. We also proposed to investigate possible improvement obtained with a sound separation front-end used jointly with sound event detection in the case of complex soundscapes [67]. This additional aspect attracted researchers from the sound separation community, including some researchers from the team (not involved in the task organization) who proposed two different approaches to combine sound event detection and sound separation [25, 24].

We also pursued our work in estimating acoustical properties of environments from recorded audio, e.g., room shape, reverberation time or absorption coefficients. Much of the information is contained in early acoustic echoes, stemming from the sound interaction with reflective materials in the room. A new anaytical method for blind early acoustic echo retrieval based on the framework of continuous dictionary learning was proposed in [30]. A new approach for mean absorption coefficient estimation from impulse reponses using virtually-supervised learning was presented in [22].

8.3.2 Speech enhancement and noise robustness

Sound source localization and counting We studied the problem of detecting the activity and counting overlapping speakers in distant-microphone recordings [26]. We proposed to treat supervised voice activity detection, overlapped speech detection, and speaker counting as instances of a general supervised classification task. We designed a temporal convolutional network (TCN) based method to address it and showed that it significantly outperforms state-of-the-art methods on two real-world distant speech datasets.

Speech enhancement We pursued our investigation of multichannel speech separation. We proposed a deflation method which iteratively estimates the location of one speaker, derives the corresponding

time-frequency mask and removes the estimated source from the mixture before estimating the next one [59]. Sunit Sivasankaran successfully defended his PhD on this topic [88].

We also finalized our work on joint reduction of acoustic echo, reverberation and noise. Our method models the target and residual signals after linear echo cancellation and dereverberation using a multi-channel Gaussian modeling framework and jointly represents their spectra by means of a neural network. We developed an iterative block-coordinate ascent algorithm to update all the filters. The proposed approach outperforms in terms of overall distortion a cascade of the individual approaches and a joint reduction approach which does not rely on a spectral model of the target and residual signals [6]. Guillaume Carbajal successfully defended his PhD on this topic [84].

In the context of ad-hoc acoustic antennas, we proposed to extend the distributed adaptive node-specific signal estimation approach to a neural network framework. At each node, a local filtering is performed to send one signal to the other nodes where a mask is estimated by a neural network in order to compute a global multi-channel Wiener filter. In an array of two nodes, we showed that this additional signal can be efficiently taken into account to predict the masks and leads to better speech enhancement performance than when the mask estimation relies only on the local signals [39, 38, 76]. We also proposed an extension of the approach to speech separation from several concurrent speakers [91].

Robust speech recognition Achieving robust speech recognition in reverberant, noisy, multi-source conditions requires not only speech enhancement and separation but also robust acoustic modeling. In order to motivate further work by the community, we created the series of CHiME Speech Separation and Recognition Challenges in 2011. This year, we organized the 6th edition of the Challenge [68]. Compared to the 5th edition, we introduced a second track, which is the first challenge activity in the community to tackle an unsegmented multispeaker speech recognition scenario with a complete set of reproducible open source baselines providing speech enhancement, speaker diarization, and speech recognition modules.

In the framework of the MMT project, we have revised two methods that take into account the uncertainty, that is the variance of the residual distortion of speech after enhancement: the DNNU method which propagates the uncertainty through the acoustic models and the GMMD method which modifies the acoustic vectors. Evaluations were carried on the TED corpus with additive noises and on 2 corpora close to our aeronautical application. We evaluated different acoustic models: the noiseless model, the noisy model and the noisy and enhanced model. The experimental results show that on the TED corpus, the GMMD uncertainty method with the noisy and enhanced model improves recognition results compared to the other models studied.

Speaker recognition Developing a robust speaker recognition system remains a challenging task due to the variations in environmental conditions, channel effect, speakers' intrinsic characteristics, etc. To improve robustness, we have investigated a data-driven acoustic feature extraction method [18]. We have explored statistical methods to compute optimal filterbank parameters such as the frequency warping scale and filter shape from the audio dataset. The proposed scheme showed considerable improvement over the popular handcrafted feature such as mel-frequency cepstral coefficients (MFCCs) for clean and noisy conditions. Acoustic front-ends developed for improving robustness in a statistical speaker recognition framework were not investigated so far in a deep learning framework. In [47], we compared different acoustic front-ends for deep speaker embeddings. Our extensive study revealed that robust speech features involving long-term processing are more effective than commonly used MFCCs, especially in noisy conditions. Our study also demonstrates the potentiality of phase-based features for robust deep speaker embeddings. In another work [48], we explored learnable MFCCs where differentiable units replace all the linear modules of the MFCC processing chain. The results indicate that learnable MFCCs are substantially better than MFCCs computed with fixed parameters.

We have also participated in the first short-duration speaker verification (SdSV) challenge, where the key problem was to recognize speakers from short-duration utterances spoken in varying channel conditions [54]. Our study demonstrates that phonetic bottleneck features are promising for text-dependent speaker recognition. Our final submission to the challenge ranked fifth among 20 submissions in the text-dependent subtask of the challenge. We have also participated in the third DIHARD challenge, where the key problem was the speaker diarization on audio-data collected from diverse real-world conditions.

We have substantially improved the challenge baseline system by integrating domain-identification and domain-dependent processing [74].

Speaker recognition systems are highly prone to the spoofing attacks performed with voice conversion and speech synthesis technology [19, 16]. The spoofing is more prevalent due to the recent technological advancements in creating fake media contents popularly known as *deepfakes*. In a recent study [17], we have demonstrated that spoofing detection becomes a more challenging task when a natural speech signal is augmented with a small portion of synthetic speech. We have proposed a solution with frame-selection, which substantially improves the spoofing detection performance for such a scenario.

Language identification State-of-the-art spoken language identification systems consist of three modules: a frame level feature extractor, a segment level embedding extractor 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 embedding extractor. The proposed approach consists in adding a regularisation term in the loss function used for training the segment level embedding extractor. Experiments were conducted with respect to transmission channel mismatch between telephone and radio channels using the RATS corpus. The proposed method is superior to adaptation of the classifier and perform on par with published language identification approaches but without using labelled data from the target domain [71, 35]. Another approach has been investigated to control the domain missmatch, which relies on combining a classification loss with the metric learning n-pair loss for training the x-vector DNN model. Such a system achieves comparable robustness to a system trained with a domain adaptation loss function but without using the domain information [36].

These 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).

8.3.3 Linguistic and semantic processing

Transcription, translation, summarization and comparison of videos Within the AMIS project, we studied different subjects related to the processing of videos. One objective of the project was to summarize videos (for example in Arabic) into a target language (for example English). The demonstrator exploits research carried on in several areas including video summarization, speech recognition, machine translation, audio summarization [21].

Detection of hate speech in social media The spectacular expansion of the Internet led to the development of a new research problem in natural language processing, the automatic detection of hate speech, since many countries prohibit hate speech in public media. In the context of the M-PHASIS project, we explored a label propagation-based semi-supervised learning system for the task of hate speech classification. We showed that pre-trained representations are label agnostic, and when used with label propagation yield poor results. Neural network-based fine-tuning was adopted to learn task-specific representations using a small amount of labeled data [29].

We also designed binary classification and regression-based approaches aiming to determine whether a comment is toxic or not. We compared different unsupervised word representations and different DNN based classifiers. Moreover, we studied the robustness of the proposed approaches to adversarial attacks by adding one (healthy or toxic) word. Our experiments showed that using BERT fine-tuning outperforms feature-based BERT, Mikolov's and fastText representations with different DNN classifiers [40] [28] [11].

In the framework of the M-PHASIS project, a new hate speech corpus has been created. More than 8,000 comments (about 4,000 in French and 4,000 in German) have been collected on News websites and manually annotated.

Introduction of semantic information in an automatic speech recognition system Current Automatic Speech Recognition (ASR) systems mainly take into account acoustic, lexical and local syntactic information. Long term semantic relations are not used. The ASR performance significantly degrades

when the training and testing conditions differ due to noise, etc. In this case the acoustic information can be less reliable. To improve performance in such conditions, we propose to supplement the ASR system with a semantic module. This module re-evaluates the N-best list of ASR hypotheses and can be seen as a form of adaptation in the context of noise. Words in the processed sentence that could have been poorly recognized are replaced by words that better correspond to the semantic context of the sentence. To achieve this, we introduced the notions of a context part and possibility zones that measure the similarity between the semantic context of the document and the corresponding possible hypotheses. We conducted experiments on the publicly available TED conferences dataset (TED-LIUM) mixed with real noise. The proposed method achieves a significant reduction of the word error rate (WER) over the ASR system without using semantic information [46] [75] [73].

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

9.1.1 Dassault and Thalès - Man Machine Teaming Initiative

- Company: Dassault and Thalès (France)
- Duration: Apr 2019 - Sept 2020
- Participants: Irène Illina, Dominique Fohr, Ismaël Bada, Stéphane Level
- Abstract: The primary goal of the project is to develop a new approach that allows coupling speech enhancement with semantic analysis for improving speech recognition robustness.

9.2 Bilateral grants with industry

9.2.1 Invoxia

- Company: Invoxia SAS (France)
- Duration: Mar 2017 – Apr 2020
- Participants: Guillaume Carbajal, Romain Serizel, Emmanuel Vincent
- Abstract: This CIFRE contract funded the PhD thesis of Guillaume Carbajal. We designed a unified deep learning based speech enhancement system that integrates all steps in the current speech enhancement chain (acoustic echo cancellation and suppression, dereverberation, and denoising) for improved hands-free voice communication.

9.2.2 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.

9.2.3 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 International initiatives

10.1.1 Inria international partners

Informal international partners

- Samuele Cornell & Stefano Squartini, Università Politecnica delle Marche (Italy): speech/audio source separation and counting [26, 25, 24, 53, 52]
- Junichi Yamagishi, National Institute of Informatics (Japan): speaker recognition & spoofing countermeasures [19, 16], voice anonymization [60, 62]
- Scott Wisdom, Hakan Erdogan, John Hershey, Google Research (United States); Justin Salamon, Adobe Research (United States); Eduardo Fonseca, Universitat Pompeu Fabra (Spain); and Prem Seetharaman, Descript (United States): Sound event detection and separation [55, 67, 96, 95]
- Tomi Kinnunen, University of Eastern Finland (Finland): speaker recognition and anti-spoofing [19, 16, 47, 54]
- Goutam Saha, Indian Institute of Technology Kharagpur (India): Speaker recognition, anti-spoofing, and speaker diarization [18, 17, 74]
- Zheng-Hua Tan, Aalborg University (Denmark): Speaker verification [54]

10.2 European initiatives

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

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

AI4EU

Title: European Artificial Intelligence On-Demand Platform and Ecosystem

Duration: Jan 2019 - Dec 2021

Coordinator: Patrick Gatellier (THALES)

Partners: 80 partners from 22 countries

Participants: Seyed Ahmad Hosseini, Slim Ouni

Summary: The aim of AI4EU is to develop a European Artificial Intelligence ecosystem, from knowledge and algorithms to tools and resources. MULTISPEECH participates in WP6 (AI4Media) in collaboration with Interdigital. The goal is to perform an audiovisual dubbing; more precisely to adapt the animation of the face of a speaker for a video which is translated from English to French. The final result is the face of the original speaker speaking and animated such that it is synchronized with the speech (translation) in the target language. We have used our lipsync technique to perform the core of this speech animation.

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.

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.

10.2.2 Collaborations in European programs, except FP7 and H2020

M-PHASIS

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

Duration: Mar 2019 - Feb 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: Irène Illina, Dominique Fohr, Ashwin Geet D'sa

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 concern the automatic detection of hate speech in social media.

10.2.3 Collaborations with major European organizations

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 will also develop open source software and lexical resources, focusing on video activity recognition as a practical testbed.

10.3 National initiatives

ANR ArtSpeech

Title: Synthèse articulatoire phonétique

Duration: Oct 2015 - Aug 2020

Coordinator: Yves Laprie (LORIA, Nancy)

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

Participants: Ioannis Douros, Yves Laprie

Abstract: The objective was to synthesize speech via the numerical simulation of the human speech production processes, i.e. the articulatory, aerodynamic and acoustic aspects. Articulatory data comes from MRI and EPGG acquisitions.

ANR JCJC KAMoulox

Title: Kernel additive modelling for the unmixing of large audio archives

Duration: Jan 2016 - May 2020

Coordinator: Antoine Liutkus (Inria Zenith)

Participants: Mathieu Fontaine

Abstract: The objective was to develop theoretical and applied tools to embed audio denoising and separation tools in web-based audio archives. The applicative scenario was to deal with the notorious audio archive "*Archives du CNRS — Musée de l'Homme*", gathering recordings dating back to the early 1900s.

PIA2 ISITE LUE

Title: *Lorraine Université d'Excellence*

Duration: Avr 2016 - Dec 2020

Coordinator: Univ de Lorraine

Participants: Ioannis Douros, Yves Laprie, 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 challenge number 6: "Knowledge engineering" has funded the PhD thesis of Ioannis Douros on articulatory modeling. 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 - Sep 2020

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, Elodie Gauthier, Thomas Girod, Denis Jouvet, Odile Mella, Slim Ouni, Leon Rohrbacher

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 VOCADOM

Title: Robust voice command adapted to the user and to the context for ambient assisted living (http://vocadom.imag.fr/)

Duration: Jan 2017 - Dec 2020

Coordinator: CNRS - LIG (Grenoble)

Partners: CNRS - LIG (Grenoble), Inria (Nancy), Univ. Lyon 2 - GREPS, THEORIS (Paris)

Participants: Dominique Fohr, Md Sahidullah, Sunit Sivasankaran, Emmanuel Vincent

Abstract: The goal is to design a robust voice control system for smart home applications. MULTISPEECH is responsible for wake-up word detection, overlapping speech separation, and speaker recognition.

ANR JCJC DiSCogs

Title: Distant speech communication with heterogeneous unconstrained microphone arrays

Duration: Sep 2018 – Mar 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 - Dec 2022

Coordinator: Denis Jouvet (Inria, Nancy)

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

Participants: Pierre Champion, Denis Jouvet, Emmanuel Vincent

Abstract: The objective 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 – Mar 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 LEAUDS

Title: Learning to understand audio scenes

Duration: Apr 2019 - Sep 2022

Coordinator: Université de Rouen Normandie

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

Participants: 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 thousands 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 BENEPHIDIRE

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

Duration: Mar 2019 - Dec 2023

Coordinator: Praxiling (Toulouse)

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

Participants: Yves Laprie, Slim Ouni, Shakeel Ahmad Sheikh

Abstract: This 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 HAIKUS

Title: Artificial Intelligence applied to augmented acoustic Scenes

Duration: Dec 2019 - May 2023

Coordinator: Ircam (Paris)

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

Participants: Antoine Deleforge, 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 Flash Open Science HARPOCRATES

Title: Open data, tools and challenges for speaker anonymization

Duration: Oct 2019 - Mar 2021

Coordinator: Eurecom (Nice)

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

Participants: Denis Jouvet, Md Sahidullah, Emmanuel Vincent

Abstract: HARPOCRATES will form a working group that will collect and share the first open datasets and tools in the field of speech privacy, and launch the first open challenge on speech privacy, specifically on the topic of voice de-identification.

InriaHub Carnot *Technologies Vocales*

Title: InriaHub Carnot *Technologies Vocales*

Duration: Jan 2019 - Dec 2020

Coordinator: Denis Jouvet

Participants: Mathieu Hu, Denis Jouvet, Dominique Fohr, Vincent Colotte, Emmanuel Vincent, Romain Serizel

Abstract: This project aims to adjust and finalize the speech synthesis and recognition modules developed for research purposes in the team, so that they can be used in interactive mode.

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, Cédric Foy, Stéphane Dilungana

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: Antoine Deleforge, Joris Cosentino, 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 [52].

10.4 Regional initiatives

CPER LCHN

Title: CPER "*Langues, Connaissances et Humanités Numériques*"

Duration: 2015 - 2020

Coordinator: Bruno Guillaume (LORIA) & Alain Polguère (ATILF)

Participants: Dominique Fohr, Denis Jouvet, Odile Mella, Yves Laprie

Abstract: The main goal is related to experimental platforms for supporting research activities in the domain of languages, knowledge and numeric humanities engineering. MULTISPEECH contributed to automatic speech recognition, speech-text alignment and prosody aspects.

ALOE Project (*Région Grand-Est - Economie Numérique*)

Title: *Logiciel éducatif Aloé 2.0*

Duration: Mar 2019 - Aug 2020

Coordinator: Com-Medic (France)

Partners: Com-Medic (France), MULTISPEECH (Inria, Nancy), 2LPN (Univ de Lorraine, Nancy), MJC / Centre Social Nomade (Vandoeuvre-Lès-Nancy)

Participants: Denis Jouvet, Vincent Colotte, Slim Ouni, Louis Delebecque

Abstract: ALOE is a method of reading relying on a specific representation of sounds. Our involvement in the project is to develop tools to translate automatically and align text sentences into phone sequences as required by the ALOE system, and to provide audio and video tutoring examples.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- General co-chair, 1st Inria-DFKI Workshop on Artificial Intelligence, Nancy, Jan 2020 (E. Vincent)
- General co-chair, 6th CHiME Speech Separation and Recognition Challenge, May 2020 (E. Vincent)
- General co-chair, 6th International Workshop on Speech Processing in Everyday Environments, May 2020 (E. Vincent)
- General co-chair, 1st Voice Privacy Challenge, Nov 2020 (E. Vincent)
- General co-chair, Detection and Classification of Acoustic Scenes and Events Challenge, Nov 2020 (R. Serizel)
- Area chair, 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (A. Deleforge, E. Vincent)
- Area chair, 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (R. Serizel)
- General co-chair, JEP-TALN-RECITAL 2020, Nancy, Jun 2020 (S. Ouni)
- Area chair, 2021 IEEE Spoken Language Technology (SLT) Workshop (M. Sahidullah)

Member of the organizing committees

- Co-organizer of INTERSPEECH 2020 special session on Voice Privacy (E. Vincent)
- Chair of Quiet Drones 2020 special session on drone audition (A. Deleforge)
- Co-organizer of DCASE 2021 Sound Event Localization and Detection task (A. Deleforge)

11.1.2 Scientific events: selection

Member of the conference program committees

- SPECOM 2020 - 22nd International Conference on Speech and Computer (D. Jouvet)
- TSD 2020 - 23rd International Conference on Text, Speech and Dialogue (D. Jouvet)
- JEP-TALN-RECITAL 2020 (S. Ouni)
- IDEKI 2020 (A. Piquard-Kipffer)

Reviewer - reviewing activities

- AIMLAI 2020 - Workshop on Advances in Interpretable Machine Learning and Artificial IntelligencFFe (E. Vincent)
- CHiME 2020 - International Workshop on Speech Processing in Everyday Environments (E. Vincent)
- DCASE 2020 - Workshop on Detection and Classification of Acoustic Scenes and Events (R. Serizel, E. Vincent)
- EUSIPCO 2020 - European Signal Processing Conference (V. Colotte, D. Jouvet, E. Vincent)
- ICASSP 2021 - IEEE International Conference on Acoustics, Speech and Signal Processing (A. Bonneau, I. Illina, D. Jouvet, R. Serizel, E. Vincent, A. Deleforge, M. Sahidullah)

- ISSP2020 - International Seminar on Speech Production (Y. Laprie)
- INTERSPEECH 2020 (A. Bonneau, D. Jouvet, I. Illina, A. Piquard-Kipffer, E. Vincent, Md. Sahidullah)
- JEP-TALN-RECITAL 2020 (A. Bonneau, D. Jouvet, V. Colotte, S. Ouni)
- Joint Workshop of Voice Conversion Challenge and Blizzard Challenge 2020 (V. Colotte)
- SLT 2021 - IEEE Spoken Language Technology Workshop (D. Jouvet, I. Illina, M. Sahidullah)
- SP 2020 - 10th International Conference on Speech Prosody (A. Bonneau, D. Jouvet)
- SPECOM 2020 - 22nd International Conference on Speech and Computer (D. Jouvet)
- TAILOR 2020 - Workshop on Foundations of Trustworthy AI (E. Vincent)
- TSD 2020 - 23rd International Conference on Text, Speech and Dialogue (D. Jouvet)
- NeurIPS 2020 - 34th Conference on Neural Information Processing Systems (A. Deleforge)
- ICLR 2021 - 9th International Conference on Learning Representations (A. Deleforge)
- ICME 2020 - IEEE International Conference on Multimedia and Expo (A. Deleforge)
- SII 2021 - IEEE/SICE International Symposium on System Integration (A. Deleforge)
- Odyssey 2020: The Speaker and Language Recognition Workshop (M. Sahidullah)

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)
- Guest Editor of Computer Speech and Language, special issue on Advances in Automatic Speaker Verification Anti-spoofing (M. Sahidullah)
- Guest Editor for EURASIP Journal on Audio, Speech, and Music Processing, special issue on Advances in Audio Signal Processing for Robots and Drones (A. Deleforge)
- Journal on Audio, Speech, and Music Processing (Y. Laprie)
- Speech Communication (D. Jouvet)
- Springer Circuits, Systems, and Signal Processing (M. Sahidullah)
- IET Signal Processing (M. Sahidullah)

Reviewer - Reviewing Activities

- Journal of the Acoustical Society of America (Y. Laprie)
- Journal of Language, Speech and Hearing Research (Y. Laprie)
- IEEE Transactions on Affective Computing (I. Illina)
- Speech Communication (A. Bonneau, D. Jouvet, M. Sahidullah)
- Computer Speech and Language (S. Ouni, M. Sahidullah)
- Computer Animation and Virtual Worlds (S. Ouni)

- Approche Neuropsychologique des Apprentissages (A.Piquard-Kipffer)
- EURASIP Journal on Audio, Speech, and Music Processing (A. Deleforge, M. Sahidullah)
- Elsevier Signal Processing (A. Deleforge)
- IEEE Transactions on Audio, Speech and Language Processing (A. Deleforge, M. Sahidullah)
- IEEE Transactions on Multimedia (A. Deleforge)
- IEEE Transactions on Robotics (A. Deleforge)
- IEEE Transactions on Cybernetics (A. Deleforge)
- IEEE Transactions on Signal Processing (A. Deleforge)
- IEEE Journal of Selected Topics in Signal Processing (A. Deleforge)
- IEEE Robotics and Automation Letters (A. Deleforge)
- IEEE Transactions on Information Forensics and Security (M. Sahidullah)
- Neural Networks (M. Sahidullah)
- IEEE Access (M. Sahidullah)

11.1.4 Invited talks

- A brief introduction to multichannel noise reduction with deep neural networks, 12th Speech in Noise Workshop, Toulouse, Jan 2020 (R. Serizel)
- Multimodal data acquisition and processing for spoken communication, Technologies du Langage Humain et Multimodalité, TLH-AFIA, Oct 2020 (S. Ouni)
- Language pathology, *Séminaire Dépistage des troubles des apprentissages*, EHESP, University of Sorbonne, Jan 2020 (A. Piquard-Kipffer)
- Screening and in-school caring for children with special educational needs – *Dépistage et prise en charge des élèves à besoins éducatifs particuliers*, Feb. 2020. INSPE & UIR, Casablanca, Morocco (A. Piquard-Kipffer)

11.1.5 Leadership within the scientific community

- Member of the Steering Committee of ISCA's Special Interest Group on Security and Privacy in Speech Communication (E. Vincent).
- Member of the Steering Committee of the Detection and Classification of Acoustic Scenes and Events (DCASE) (R. Serizel)
- Secretary/Treasurer, executive member of AVISA (Auditory-VIsual Speech Association), an ISCA Special Interest Group (S. Ouni)

11.1.6 Scientific expertise

- Reviewer of ANR projects (D. Jouvet, Y. Laprie)
- Member of the Scientific Committee of an Institute for deaf children and teenagers, INJS-Metz (A. Piquard- Kipffer)

11.1.7 Research administration

- Member of Management board of Université de Lorraine (Y. Laprie)
- 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)
- Co-Chair of the joint Inria-Loria *Commission pour l'Action et la Responsabilité Ecologique* (CARE, CLDD) (A. Deleforge)
- Member of Inria's Evaluation Committee (E. Vincent)
- Member of the *Comité Espace Transfert* of Inria Nancy - Grand Est (E. Vincent)
- Member of the national hiring committee for Inria Junior Research Scientists (E. Vincent)
- Member of the hiring committee for Junior Research Scientists, Inria Rennes (E. Vincent)
- Member of *Commission paritaire* of Université de Lorraine (Y. Laprie)
- Member of the *Commission de développement technologique* of Inria Nancy - Grand Est (R. Serizel)
- Member of the *Commission du personnel scientifique* of Inria Nancy - Grand Est (R. Serizel)
- Member of a recruitment committee for Professor at ENIM-LCFC, Université de Lorraine (Y. Laprie)
- Member of a recruitment committee for Assistant Professor at Université Paris-Sud (D. Jouvet)
- Member of a recruitment committee for Assistant Professor at Le Mans Université (D. Jouvet)
- Member of the HCERES (*Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur*) evaluation committee for Gipsa-Lab, 2020 (S. Ouni)
- Member of the CNU-27 (*Conseil National des Universités*) - Computer Science (S. Ouni)
- Member of the *Commission Information et Edition Scientifique* (CIES) of Inria Nancy - Grand Est (A. Deleforge)

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, Speech manipulations (2 hours), L1, *Département d'orthophonie*, 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: O. Mella, Computer Networking (64 hours), L2-L3, University of Lorraine, France
- Licence: O. Mella, Introduction to Web Programming (24 hours) L1, University of Lorraine, France
- Licence: O. Mella, Digital tools (18 hours) L1, University of Lorraine, France
- Licence: A. Piquard-Kipffer, Education Science (40 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
- Master: V. Colotte, Introduction to Speech Analysis and Recognition (18 hours), M1, University of Lorraine, France
- Master: V. Colotte, Integration project: multimodal interaction with Pepper (15 hours), M2, University of Lorraine, France
- Master: D. Jouvet and S. Ouni, Multimodal oral comunication (24 hours), M2, University of Lorraine
- Master: Y. Laprie, Speech corpora (30 hours), M1, University of Lorraine, France
- Master: O. Mella, Computer Networking (10 hours), M1, University of Lorraine, France
- Master: S. Ouni, Multimedia in Distributed Information Systems (31 hours), M2, University of Lorraine
- Master: A. Piquard-Kipffer, Dyslexia, Dysorthographia diagnosis (9 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 and S. Ouni, Oral speech processing (24 hours), M2, University of Lorraine
- Master: E. Vincent and A. Kulkarni, Neural networks (38 hours), M2, University of Lorraine
- Continuous training: A. Piquard-Kipffer, Special Educational Needs (53 hours), INSPE,University of Lorraine, France
- Doctorat: 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: Amal Houidhek, "*Synthèse paramétrique de parole arabe*", Fev 12, 2020 , *cotutelle*, V. Colotte, D. Jouvet and Z. Mnasri (ENIT, Tunisia) [87].
- PhD: Guillaume Carbajal, "*Apprentissage profond bout-en-bout pour le rehaussement de la parole*", Université de Lorraine, Apr 24, 2020, R. Serizel, E. Vincent and É. Humbert (Invoxia) [84].
- PhD: Ioannis Douros, "Towards a 3 dimensional dynamic generic speaker model to study geometry simplifications of the vocal tract using magnetic resonance imaging data", Sep 2, 2020, P.-A. Vuissoz (IADI) and Y. Laprie [86].
- PhD: Sunit Sivasankaran, "Localization guided speech separation", Sep 4, 2020, D. Fohr and E. Vincent [88].
- PhD: Sara Dahmani, "*Synthèse audiovisuelle de la parole expressive : modélisation des émotions par apprentissage profond*", Nov 13, 2020, S. Ouni and V. Colotte [85].
- PhD: Amine Menacer, "*Traduction automatique de vidéos*", Nov 17, 2020, K. Smaïli (LORIA) and D. Jouvet.
- PhD: Diego Di Carlo, "Echo-aware signal processing for audio scene analysis", Dec 4, 2020, A. Deleforge and N. Bertin (Inria Rennes).
- PhD in progress: Théo Biasutto, "Multimodal coarticulation modeling: Towards the animation of an intelligible speaking head", S. Ouni.
- PhD in progress: Lou Lee, "*Fonctions pragmatiques et prosodie de marqueurs discursifs en français et en anglais*", Oct 2017, Y. Keromnes (ATILF) and D. Jouvet.
- PhD in progress: Nicolas Turpault, "Deep learning for sound scene analysis in real environments", Jan 2018, R. Serizel and E. Vincent.
- PhD in progress: Raphaël Duroselle, "*Adaptation de domaine par réseaux de neurones appliquée au traitement de la parole*", Sep 2018, D. Jouvet and I. Illina.
- PhD in progress: Nicolas Furnon, "Deep-learning based speech enhancement with ad-hoc microphone arrays", Oct 2018, 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: Manuel Pariente, "Deep learning-based phase-aware audio signal modeling and estimation", Oct 2018, A. Deleforge and E. Vincent.
- 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: Xuechen Liu, "Robust speaker recognition for smart assistant technology", Jan 2020, M. Sahidullah.
- 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: 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.

11.2.3 Juries

Participation in HDR and PhD juries

- Participation in the PhD jury of Adrien Gresse (Avignon Université, Feb 2020), E. Vincent, reviewer
- Participation in the PhD jury of Thien-Hoa Le (Lorraine university, May 2020), I. Illina, member
- Participation in the PhD jury of Salima Mdhaffar (Le Mans université, Jul 2020), I. Illina, reviewer
- Participation in the PhD jury of Hadrien Pujol (HESAM Université, Oct 2020), E. Vincent, reviewer, A. Deleforge, examiner
- Participation in the PhD jury of Meysam Shamsi (Université de Rennes, Oct 2020), S. Ouni, reviewer
- Participation in the PhD jury of Weipeng He (EPFL, Nov 2020), A. Deleforge, reviewer
- Participation in the PhD jury of Dodji Gbedahou (Université Paul-Valéry Montpellier 3, Nov 2020), S. Ouni, member
- Participation in the HDR jury of Xavier Alameda-Pineda (Université Grenoble Alpes, Dec 2020), E. Vincent, reviewer
- Participation in the PhD jury of Mirco Pezzoli (Politecnico di Milano, Dec 2020), A. Deleforge, reviewer
- Participation in the PhD jury of Félix Gontier (Ecole Centrale Nantes, Dec 2020), R. Serizel, member
- Participation in the PhD jury of Laurine Dalle (Université Paul-Valéry Montpellier 3, Dec 2020), A.Piquard-Kipffer, member

Participation in other juries

- Participation in CAFIPEMPF Jury - Master Learning Facilitator (Académie de Nancy-Metz & Lorraine University) April, May 2020, A. Piquard-Kipffer
- Participation in CRPE Jury - Master Teaching and Education Competitive Entrance (Académie de Nancy-Metz & Lorraine University) Apr & Jun 2020, A. Piquard-Kipffer
- Participation in the Competitive Entrance Examination into Speech-Language Pathology Departement (University of Lorrain) April 2020, A. Piquard-Kipffer

11.3 Popularization

11.3.1 Articles and contents

- Article "*Peut-on faire confiance aux IA ?*" in *The Conversation*, Nov 20, 2020 (E. Vincent) [97]
- Interview "*Protection de la vie privée : 2 outils de transformation de la voix et de texte*", *Radio Village Innovation*, Sep 16 & Oct 7, 2020 (E. Vincent)
- Interview for the CNIL White Paper "*À votre écoute – Exploration des enjeux éthiques, techniques et juridiques des assistants vocaux*", Sep 7, 2020 (E. Vincent)
- Interview for France 3 Lorraine TV journal "Acoust.IA: *le projet d'application destiné aux acousticiens*"[3], Dec 2020 (A. Deleforge)

11.3.2 Interventions

- Talk "*Assistants vocaux, vie privée — Enjeux scientifiques et technologiques*", Meetup CNIL, Sep 2020 (E. Vincent)
- Animation of a round-table meeting "*Langues des signes et numérique : quels défis, quels enjeux pour les apprentissages ?*", International Sign Language Day - Sep 2020, INJS-Metz (A. Piquard-Kipffer)
- Animation of a booth on "Teaching Robots to Hear Us" for Fête de la Science, Nancy, Oct 2020 (A. Deleforge)
- Participation to the Science-Theater project Binôme, compagnie Les Sens des Mots, Oct 2020 - Oct 2021 (A. Deleforge).
- Presentation of the project Audio Cockpit Denoising for voice Command at the Forum Innovation Defence, Dec 2020 (D. Fohr)
- Presentation of the project Audio Cockpit Denoising for voice Command to Florence Parly, Minister of the Armed Forces, Dec 2020 (I. Illina)
- Presentation of METAL project at JANE (*journée académique du numérique*), Feb 2020 (S. Ouni)
- Talk: "*la scolarisation des élèves dyslexiques*", Training of trainers - Académie de Nancy-Metz & INSPE de l'Académie de Nancy-Metz, Jan 2020 (A.Piquard-Kipffer)

11.3.3 Creation of media or tools for science outreach

- Video "COMPRISE Voice Transformer", https://www.youtube.com/watch?v=kh8no66BSDM
- Popular science blog post on group testing COVID-19, https://members.loria.fr/ADeleforge/les-maths-du-group-testing-melanger-des-prelevements-pour-accelerer-la-detection-du-covid-19/ (A. Deleforge)

[3] https://youtu.be/daruKG2TXtc

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] K. Nathwani, E. Vincent and I. Illina. 'DNN Uncertainty Propagation using GMM-Derived Uncertainty Features for Noise Robust ASR'. In: *IEEE Signal Processing Letters* (Jan. 2018). DOI: 10.1109/LSP.2018.2791534. URL: https://hal.inria.fr/hal-01680658.

[4] 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.

[5] 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.

12.2 Publications of the year

International journals

[6] G. Carbajal, R. Serizel, E. Vincent and E. Humbert. 'Joint NN-Supported Multichannel Reduction of Acoustic Echo, Reverberation and Noise'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* (18th July 2020). DOI: 10.1109/TASLP.2020.3008974. URL: https://hal.inria.fr/hal-02372579.

[7] Z. Chelly Dagdia and C. Zarges. 'A detailed study of the distributed rough set based locality sensitive hashing feature selection technique'. In: *Fundamenta Informaticae* (2020). DOI: 10.3233/FI-2016-0000. URL: https://hal.inria.fr/hal-02880638.

[8] Z. Chelly Dagdia, C. Zarges, G. Beck and M. Lebbah. 'A scalable and effective rough set theory-based approach for big data pre-processing'. In: *Knowledge and Information Systems (KAIS)* (2nd May 2020). DOI: 10.1007/s10115-020-01467-y. URL: https://hal.inria.fr/hal-02880626.

[9] S. Dahmani, V. Colotte and S. Ouni. 'Some consideration on expressive audiovisual speech corpus acquisition using a multimodal platform'. In: *Language Resources and Evaluation* (26th July 2020). DOI: 10.1007/s10579-020-09500-w. URL: https://hal.archives-ouvertes.fr/hal-02907046.

[10] M. Fontaine, R. Badeau and A. Liutkus. 'Separation of Alpha-Stable Random Vectors'. In: *Signal Processing* (Jan. 2020), p. 107465. DOI: 10.1016/j.sigpro.2020.107465. URL: https://hal.inria.fr/hal-02433213.

[11] A. Geet D'Sa, I. Illina and D. Fohr. 'Classification of Hate Speech Using Deep Neural Networks'. In: *Revue d'Information Scientifique & Technique*. From Data and Information Processing to Knowledge Organization : Architectures, Models and Systems 25.01 (22nd Dec. 2020). URL: https://hal.archives-ouvertes.fr/hal-03101938.

[12] I. Illina and D. Fohr. 'RNN Language Model Estimation for Out-of-Vocabulary Words'. In: *Lecture Notes in Artificial Intelligence* 12598 (2020). DOI: 10.1007/978-3-030-66527-2_15. URL: https://hal.archives-ouvertes.fr/hal-03054936.

[13] K. Isaieva, Y. Laprie, F. Odille, I. K. Douros, J. Felblinger and P.-A. P. Vuissoz. 'Measurement of Tongue Tip Velocity from Real-Time MRI and Phase-Contrast Cine-MRI in Consonant Production'. In: *Journal of Imaging* 6.5 (May 2020), p. 31. DOI: 10.3390/jimaging6050031. URL: https://hal.univ-lorraine.fr/hal-02923466.

[14] K. Isaieva, Y. Laprie, N. Turpault, A. Houssard, J. Felblinger and P.-A. Vuissoz. 'Automatic Tongue Delineation from MRI Images with a Convolutional Neural Network Approach'. In: *Applied Artificial Intelligence* (4th Oct. 2020), pp. 1–9. DOI: 10.1080/08839514.2020.1824090. URL: https://hal.archives-ouvertes.fr/hal-02962336.

[15] M. Jerbi, Z. Chelly Dagdia, S. Bechikh, M. Makhlouf and L. B. Said. 'On the Use of Artificial Malicious Patterns for Android Malware Detection'. In: *Computers and Security* 92 (2020). DOI: 10.1016/j.cose.2020.101743. URL: https://hal.inria.fr/hal-02464180.

[16] T. Kinnunen, H. Delgado, N. Evans, K.-A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi and D. A. Reynolds. 'Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing*. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (17th July 2020), pp. 2195–2210. DOI: 10.1109/TASLP.2020.3009494. URL: https://hal.archives-ouvertes.fr/hal-02900931.

[17] A. K. 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.

[18] S. Sarangi, M. Sahidullah and G. Saha. 'Optimization of data-driven filterbank for automatic speaker verification'. In: *Digital Signal Processing* 104 (Sept. 2020). DOI: 10.1016/j.dsp.2020.102795. URL: https://hal.archives-ouvertes.fr/hal-02900353.

[19] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. Le Maguer, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang and Z.-H. Ling. 'ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech'. In: *Computer Speech and Language* 64 (Nov. 2020), p. 101114. DOI: 10.1016/j.csl.2020.101114. URL: https://hal.archives-ouvertes.fr/hal-02945493.

[20] I. Zangar, Z. Mnasri, V. Colotte and D. Jouvet. 'Duration modelling and evaluation for Arabic statistical parametric speech synthesis'. In: *Multimedia Tools and Applications* (2nd Nov. 2020). DOI: 10.1007/s11042-020-09901-7. URL: https://hal.inria.fr/hal-03007287.

International peer-reviewed conferences

[21] M. Amine Menacer, D. Fohr, D. Jouvet, K. Abidi, D. Langlois and K. Smaïli. 'AMIS project : automatic summarization and translation of video'. In: *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 4 : Démonstrations et résumés d'articles internationaux*. 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 4 : Démonstrations et résumés d'articles internationaux. Nancy, France, 2020, pp. 53–56. URL: https://hal.archives-ouvertes.fr/hal-02768513.

[22] C. Bastien, A. Deleforge and C. Foy. 'Mean Absorption Coefficient Estimation From Impulse Responses: Deep Learning vs. Sabine'. In: *Forum Acusticum 2020*. Forum Acusticum 2020. Virtual, France: https://fa2020.universite-lyon.fr/fa2020/english-version/welcome-46678.kjsp?RH=FA2020, 1st Jan. 2020, p. 2. URL: https://hal.inria.fr/hal-03045556.

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

[24] S. Cornell, M. Olvera, M. Pariente, G. Pepe, E. Principi, L. Gabrielli and S. Squartini. 'Domain-Adversarial Training and Trainable Parallel Front-end for the DCASE 2020 Task 4 Sound Event Detection Challenge'. In: DCASE 2020 - 5th Workshop on Detection and Classification of Acoustic Scenes and Events. Virtual, Japan, 2nd Nov. 2020. URL: https://hal.inria.fr/hal-02962911.

[25] S. Cornell, M. Olvera, M. Pariente, G. Pepe, E. Principi, L. Gabrielli and S. Squartini. 'Task-Aware Separation for the DCASE 2020 Task 4 Sound Event Detection and Separation Challenge'. In: DCASE 2020 - 5th Workshop on Detection and Classification of Acoustic Scenes and Events. Virtual, Japan, Mar. 2020. URL: https://hal.inria.fr/hal-02962907.

[26] S. Cornell, M. Omologo, S. Squartini and E. Vincent. 'Detecting and counting overlapping speakers in distant speech scenarios'. In: INTERSPEECH 2020. Shanghai, China, 25th Oct. 2020. URL: https://hal.inria.fr/hal-02908241.

[27] P. de las Cuevas, P. Garcia-Sanchez, Z. Chelly Dagdia, M.-I. Garcia-Arenas and J. J. Merelo. 'Automatic rule extraction from access rules using Genetic Programming'. In: EvoCOP 2020 - 20th European Conference on Evolutionary Computation in Combinatorial Optimisation. Seville, Spain: http://www.evostar.org/2020/, 9th Apr. 2020. URL: https://hal.inria.fr/hal-02880764.

[28] A. G. D'Sa, I. Illina and D. Fohr. 'BERT and fastText Embeddings for Automatic Detection of Toxic Speech'. In: SIIE 2020 - Information Systems and Economic Intelligence. Tunis, Tunisia, 6th Feb. 2020. URL: https://hal.inria.fr/hal-02448197.

[29] A. G. D'Sa, I. Illina, D. Fohr, D. Klakow and D. Ruiter. 'Label Propagation-Based Semi-Supervised Learning for Hate Speech Classification'. In: Insights from Negative Results Workshop, EMNLP 2020. Punta Cana, Dominican Republic, 19th Nov. 2020. URL: https://hal.inria.fr/hal-02964065.

[30] D. Di Carlo, C. Elvira, A. Deleforge, N. Bertin and R. Gribonval. 'BLASTER: An Off-Grid Method for Blind and Regularized Acoustic Echoes Retrieval – with supplementary material'. In: ICASSP 2020 - IEEE International Conference on Acoustic Speech and Signal Processing. Barcelona, Spain, 6th Feb. 2020. URL: https://hal.archives-ouvertes.fr/hal-02469901.

[31] I. K. Douros, C. Dourou, Y. Xie, J. Felblinger, K. Isaieva, P.-A. Vuissoz and Y. Laprie. 'Synthesize MRI vocal tract data during CV production'. In: ISSP 2020 - 12th International Seminar on Speech Production. Providence / Virtual, United States: https://issp2020.yale.edu/, 14th Dec. 2020. URL: https://hal.inria.fr/hal-03090873.

[32] I. K. Douros, A. Kulkarni, C. Dourou, Y. Xie, J. Felblinger, K. Isaieva, P.-A. Vuissoz and Y. Laprie. 'Using Silence MR Image to Synthesise Dynamic MRI Vocal Tract Data of CV'. In: INTERSPEECH 2020. Shangaï / Virtual, China: http://www.interspeech2020.org/, 25th Oct. 2020. URL: https://hal.inria.fr/hal-03090808.

[33] 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: https://eusipco2020.org/, 18th Jan. 2021. URL: https://hal.inria.fr/hal-03090824.

[34] I. K. Douros, Y. Xie, C. Dourou, J. Felblinger, K. Isaieva, P.-A. Vuissoz and Y. Laprie. 'Vocal tract sagittal slices estimation from MRI midsagittal slices during speech production of CV'. In: ISSP 2020 - 12th International Seminar on Speech Production. Providence / Virtual, United States: https://issp2020.yale.edu/program.html, 14th Dec. 2020. URL: https://hal.inria.fr/hal-03090865.

[35] R. Duroselle, D. Jouvet and I. Illina. 'Metric learning loss functions to reduce domain mismatch in the x-vector space for language recognition'. In: INTERSPEECH 2020. Shangaï / Virtual, China, 26th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02920460.

[36] R. Duroselle, D. Jouvet and I. Illina. 'Unsupervised regularization of the embedding extractor for robust language identification'. In: Odyssey 2020 - The Speaker and Language Recognition Workshop. Tokyo, Japan, 2nd Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-02544156.

[37] A. Elmerich, A. Amelot, S. Maeda, Y. Laprie, J. F. Papon and L. Crevier-Buchman. 'F1 and F2 measurements for French oral vowel with a new pneumotachograph mask'. In: ISSP 2020 - 12th International Seminar on Speech Production. Providence / Virtual, United States: https://issp2020.yale.edu/, 14th Dec. 2020. URL: https://hal.inria.fr/hal-03090851.

[38] 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, Canada, 6th June 2021. URL: https://hal.archives-ouvertes.fr/hal-02985794.

[39] N. Furnon, R. Serizel, I. Illina and S. Essid. 'DNN-Based Distributed Multichannel Mask Estimation for Speech Enhancement in Microphone Arrays'. In: ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing. Barcelona, Spain, 4th May 2020. URL: https://hal.archives-ouvertes.fr/hal-02389159.

[40] A. Geet D'Sa, I. Illina and D. Fohr. 'Towards Non-Toxic Landscapes: Automatic Toxic Comment Detection Using DNN'. In: TRAC-2020, Second Workshop on Trolling, Aggression and Cyberbullying (LREC, 2020). Marseille, France, 16th May 2020. URL: https://hal.inria.fr/hal-02530879.

[41] M. Hu, L. Pierron, E. Vincent and D. Jouvet. 'Kaldi-web: An installation-free, on-device speech recognition system'. In: INTERSPEECH 2020 Show & Tell. Shanghai, China, 25th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02910876.

[42] K. Isaieva, Y. Laprie, A. Houssard, J. Felblinger and P.-A. Vuissoz. 'Tracking the tongue contours in rt-MRI films with an autoencoder DNN approach'. In: ISSP 2020 - 12th International Seminar on Speech Production. Providence / Virtual, United States: https://issp2020.yale.edu/, 14th Dec. 2020. URL: https://hal.inria.fr/hal-03090859.

[43] A. Kulkarni, V. Colotte and D. Jouvet. 'Deep variational metric learning for transfer of expressivity in multispeaker text to Speech'. In: SLSP 2020 - 8th International Conference on Statistical Language and Speech Processing. Cardiff / Virtual, United Kingdom, 14th Oct. 2020. URL: https://hal.inria.fr/hal-02573885.

[44] A. Kulkarni, V. Colotte and D. Jouvet. 'Transfer learning of the expressivity using flow metric learning in multispeaker text-to-speech synthesis'. In: INTERSPEECH 2020. Shanghai / Virtual, China, 26th Oct. 2020. URL: https://hal.inria.fr/hal-02572106.

[45] L. Lee, D. Jouvet, K. Bartkova, Y. Keromnes and M. Dargnat. 'Correlation between prosody and pragmatics: case study of discourse markers in French and English'. In: INTERSPEECH 2020. Shanghai, China, 26th Oct. 2020. URL: https://hal.inria.fr/hal-02968475.

[46] S. Level, I. Illina and D. Fohr. 'Introduction of semantic model to help speech recognition'. In: TSD 2020 - Twenty-third International Conference on Text, Speech and Dialogue. Brno, Czech Republic, 8th Sept. 2020. URL: https://hal.archives-ouvertes.fr/hal-02862245.

[47] X. Liu, M. Sahidullah and T. Kinnunen. 'A Comparative Re-Assessment of Feature Extractors for Deep Speaker Embeddings'. In: INTERSPEECH 2020. Shanghai, China, 25th Oct. 2020. URL: https://hal.inria.fr/hal-02909105.

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

[49] M. Maouche, B. M. L. Srivastava, N. Vauquier, A. Bellet, M. Tommasi and E. Vincent. 'A comparative study of speech anonymization metrics'. In: INTERSPEECH 2020. Shanghai, China, 25th Oct. 2020. URL: https://hal.inria.fr/hal-02907918.

[50] E. Marquer, A. Kulkarni and M. Couceiro. 'Embedding Formal Contexts Using Unordered Composition'. In: FCA4AI - 8th International Workshop "What can FCA do for Artificial Intelligence?" (colocated wit ECAI2020). Santiago de Compostela, Spain, 2020. URL: https://hal.archives-ouvertes.fr/hal-02912874.

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

[52] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge and E. Vincent. 'Asteroid: the PyTorch-based audio source separation toolkit for researchers'. In: Interspeech 2020. Shanghai, China: http://interspeech2020.org/, 26th Oct. 2020. URL: https://hal.inria.fr/hal-02962964.

[53] M. Pariente, S. Cornell, A. Deleforge and E. Vincent. 'Filterbank design for end-to-end speech separation'. In: ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing. Barcelona, Spain, 4th May 2020. URL: https://hal.archives-ouvertes.fr/hal-02355623.

[54] 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: IEEE Spoken Language Technology Workshop 2021. Shenzhen, China, Jan. 2022. URL: https://hal.archives-ouvertes.fr/hal-02907037.

[55] R. Serizel, N. Turpault, A. Shah and J. Salamon. 'Sound event detection in synthetic domestic environments'. In: ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing. Barcelona, Spain, 4th May 2020. URL: https://hal.inria.fr/hal-02355573.

[56] I. Sheikh, E. Vincent and I. Illina. 'On semi-supervised LF-MMI training of acoustic models with limited data'. In: INTERSPEECH 2020. Shanghai, China, 25th Oct. 2020. URL: https://hal.inria.fr/hal-02907924.

[57] P. Singh, G. Saha and M. Sahidullah. 'Non-linear frequency warping using constant-Q transformation for speech emotion recognition'. In: 2021 International Conference on Computer Communication and Informatics (ICCCI -2021). Coimbatore, India, 27th Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-03134015.

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

[59] S. Sivasankaran, E. Vincent and D. Fohr. 'SLOGD: Speaker Location Guided Deflation Approach to Speech Separation'. In: ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing. Barcelona, Spain, 4th May 2020. URL: https://hal.inria.fr/hal-02355613.

[60] B. M. L. Srivastava, N. Tomashenko, X. Wang, E. Vincent, J. Yamagishi, M. Maouche, A. Bellet and M. Tommasi. 'Design Choices for X-vector Based Speaker Anonymization'. In: INTERSPEECH 2020. Shanghai, China, 25th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02610447.

[61] 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.

[62] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé and M. Todisco. 'Introducing the VoicePrivacy initiative'. In: INTERSPEECH 2020. Shanghai, China, 25th Oct. 2020. URL: https://hal.inria.fr/hal-02562199.

[63] A. Tsukanova, I. K. Douros and Y. Laprie. 'DNN-Based Parametric Speech Synthesis Enhanced With Articulatory Information'. In: ISSP 2020 - 12th International Seminar on Speech Production. Providence / Virtual, United States: https://issp2020.yale.edu/, 14th Dec. 2020. URL: https://hal.inria.fr/hal-03090869.

[64] M. A. T. Turan, E. Vincent and D. Jouvet. 'Achieving multi-accent ASR via unsupervised acoustic model adaptation'. In: INTERSPEECH 2020. Shanghai, China, 25th Oct. 2020. URL: https://hal.inria.fr/hal-02907929.

[65] N. Turpault and R. Serizel. 'Training Sound Event Detection On A Heterogeneous Dataset'. In: DCASE Workshop. Tokyo, Japan, Mar. 2020. URL: https://hal.inria.fr/hal-02891665.

[66] N. Turpault, R. Serizel and E. Vincent. 'Limitations of weak labels for embedding and tagging'. In: ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing. Barcelona, Spain, 4th May 2020. URL: https://hal.inria.fr/hal-02467401.

[67] N. Turpault, S. Wisdom, H. Erdogan, J. R. Hershey, R. Serizel, E. Fonseca, P. Seetharaman and J. Salamon. 'Improving Sound Event Detection In Domestic Environments Using Sound Separation'. In: DCASE Workshop 2020 - Detection and Classification of Acoustic Scenes and Events. Tokyo / Virtual, Japan, Mar. 2020. URL: https://hal.inria.fr/hal-02891700.

[68] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka and N. Ryant. 'CHiME-6 Challenge: Tackling multispeaker speech recognition for unsegmented recordings'. In: CHiME 2020 - 6th International Workshop on Speech Processing in Everyday Environments. Barcelona / Virtual, Spain, 4th May 2020. URL: https://hal.inria.fr/hal-02546993.

[69] M. Wolf, D. Ruiter, A. G. D'Sa, L. Reiners, J. Alexandersson and D. Klakow. 'HUMAN: Hierarchical Universal Modular ANnotator'. In: EMNLP 2020 System Demonstration. Punta Cana (Virtual), Dominican Republic, Nov. 2020. URL: https://hal.inria.fr/hal-02958831.

National peer-reviewed Conferences

[70] S. Dahmani, V. Colotte and S. Ouni. 'Étude comparative des paramètres d'entrée pour la synthèse expressive audiovisuelle de la parole par DNNs'. In: *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*. 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole. Nancy, France, 2020, pp. 127–135. URL: https://hal.archives-ouvertes.fr/hal-02798526.

[71] R. Duroselle, D. Jouvet and I. Illina. 'Adaptation de domaine non supervisée pour la reconnaissance de la langue par régularisation d'un réseau de neurones'. In: *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*. 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole. Nancy, France, 2020, pp. 190–198. URL: https://hal.archives-ouvertes.fr/hal-02798536.

[72] L. Lee, D. Jouvet, K. Bartkova, Y. Keromnes and M. Dargnat. 'Étude comparative de corrélats prosodiques de marqueurs discursifs français et anglais selon leur fonction pragmatique'. In: *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*. 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole. Nancy, France, 2020, pp. 335–343. URL: https://hal.archives-ouvertes.fr/hal-02798556.

[73] S. Level, I. Illina and D. Fohr. 'Introduction d'informations sémantiques dans un système de reconnaissance de la parole'. In: *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*. 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole. Nancy, France, 2020, pp. 362–369. URL: https://hal.archives-ouvertes.fr/hal-02798559.

Conferences without proceedings

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

[75] S. Level, I. Illina and D. Fohr. 'Semantic Context Model for Efficient Speech Recognition'. In: ICCAS 2020 - The first International Conference on Cognitive Aircraft Systems. Toulouse, France, 18th Mar. 2020. URL: https://hal.inria.fr/hal-02500428.

[76] R. Serizel. 'A brief introduction to multichannel noise reduction with deep neural networks'. In: SpiN 2020 - 12th Speech in Noise Workshop. Toulouse, France, 9th Jan. 2020. URL: https://hal.inria.fr/hal-02506387.

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

Scientific book chapters

[78] Z. Chelly Dagdia and M. Mirchev. 'When Evolutionary Computing Meets Astro- and Geoinformatics'. In: *Knowledge Discovery in Big Data from Astronomy and Earth Observation*. https://www.sciencedirect.com/science/article/pii/B9780128191545000266, 1st May 2020, pp. 283–306. URL: https://hal.inria.fr/hal-02880731.

[79] A. Lelu and M. Cadot. 'Importance of Dataspace Embeddings when Evaluating Text Clustering Methods'. In: *Data Analysis and Rationality in a Complex World*. Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03053176.

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

[80] C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla and S. Schneider, eds. *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition. Volume 2 : Traitement Automatique des Langues Naturelles*. JEP-TALN-RECITAL 2020. Vol. 2. Volume 2 : Traitement Automatique des Langues Naturelles. Nancy, France: http://talnarchives.atala.org/TALN/TALN-2020/, 2020. URL: https://hal.archives-ouvertes.fr/hal-02784750.

[81] C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla and S. Schneider, eds. *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*. Nancy, France, 2020. URL: https://hal.archives-ouvertes.fr/hal-02798507.

[82] C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla and S. Schneider, eds. *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL*. JEP-TALN-RECITAL 2020 : 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Vol. 3. Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL. Nancy, France: http://talnarchives.atala.org/RECITAL/RECITAL-2020/, 2020. URL: https://hal.archives-ouvertes.fr/hal-02786181.

[83] C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla and S. Schneider, eds. *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 4 : Démonstrations et résumés d'articles internationaux*. JEP-TALN-RECITAL 2020. Vol. 4. Volume 4 : Démonstrations et résumés d'articles internationaux. Nancy, France: http://talnarchives.atala.org/TALN/TALN-2020/, 2020. URL: https://hal.archives-ouvertes.fr/hal-02768750.

Doctoral dissertations and habilitation theses

[84] G. Carbajal. 'End-to-end deep learning for speech enhancement'. Université de Lorraine, 24th Apr. 2020. URL: https://hal.univ-lorraine.fr/tel-02877545.

[85] S. Dahmani. 'Audiovisual synthesis of expressive speech : modeling of emotions with deep learning'. Université de Lorraine, 13th Nov. 2020. URL: https://hal.inria.fr/tel-03079349.

[86] I. K. Douros. 'Towards a 3 dimensional dynamic generic speaker model to study geometry simplifications of the vocal tract using magnetic resonance imaging data'. Université de Lorraine, 2nd Sept. 2020. URL: https://hal.univ-lorraine.fr/tel-03008224.

[87] A. Houidhek. 'Parametric synthesis of Arabic speech'. Université de Lorraine; Université de Tunis El Manar (Tunisie), 11th Feb. 2020. URL: https://hal.univ-lorraine.fr/tel-03050597.

[88] S. Sivasankaran. 'Localization guided speech separation'. Université de Lorraine, 4th Sept. 2020. URL: https://hal.univ-lorraine.fr/tel-02961882.

Reports & preprints

[89] P. Champion, D. Jouvet and A. Larcher. *Speaker information modification in the VoicePrivacy 2020 toolchain*. INRIA Nancy, équipe Multispeech; LIUM - Laboratoire d'Informatique de l'Université du Mans, 5th Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-02995855.

[90] G. Ferroni, N. Turpault, J. Azcarreta, F. Tuveri, R. Serizel, Ç. Bilen and S. Krstulović. *Improving Sound Event Detection Metrics: Insights from DCASE 2020*. 26th Oct. 2020. URL: https://hal.inria.fr/hal-02978422.

[91] N. Furnon, R. Serizel, I. Illina and S. Essid. *DNN-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays*. 2nd Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-02985867.

[92] A. Kulkarni, V. Colotte and D. Jouvet. *Improving Latent Representation For End To End Multispeaker Expressive Text To Speech System*. 26th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02978485.

[93] A. Kulkarni, I. K. Douros, V. Colotte and D. Jouvet. *Emotion recognition from phoneme-duration information*. 29th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02983229.

[94] A. K. 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.

[95] 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*. 31st Oct. 2020. URL: https://hal.inria.fr/hal-02984675.

[96] 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?* 31st Oct. 2020. URL: https://hal.inria.fr/hal-02984693.

12.3 Other

Scientific popularization

[97] E. Vincent. 'Peut-on faire confiance aux IA ?' In: *The Conversation* (21st Nov. 2020). URL: https://hal.inria.fr/hal-03017840.

Softwares

[98] [SW] S. Sivasankaran, I. Illina and E. Vincent, *voiceHome-2 corpus - automatic speech recognition baseline - scripts*, 12th Oct. 2020. HAL: ⟨hal-02963802⟩, URL: https://hal.inria.fr/hal-02963802, SWHID: ⟨swh:1:dir:e61ed9084af0d3e8542cd4ab3a990d24314a6724;origin=https://hal.archives-ouvertes.fr/hal-02963802;visit=swh:1:snp:b958e3aa64f6b1663929789c8cf28d019f55f57d;anchor=swh:1:rev:6b9bf3964385d0c16d262796d9e4a3a30a52dafd;path=/⟩.