IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2019

Project-Team MULTISPEECH

Speech Modeling for Facilitating Oral-Based Communication

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

RESEARCH CENTER
Nancy - Grand Est

THEME
Language, Speech and Audio

Table of contents

Project-Team MULTISPEECH

Creation of the Team: 2014 July 01, updated into Project-Team: 2015 July 01

Keywords:

Computer Science and Digital Science:

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
A5.9.5. - Sparsity-aware processing
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

1. Team, Visitors, External Collaborators

Research Scientists

Denis Jouvet [Team leader, Inria, Senior Researcher, on secondment from Corps des Mines, HDR]

Anne Bonneau [CNRS, Researcher]
Antoine Deleforge [Inria, Researcher]
Dominique Fohr [CNRS, Researcher]
Yves Laprie [CNRS, Senior Researcher, HDR]
Emmanuel Vincent [Inria, Senior Researcher, HDR]
Md Sahidullah [Inria, Starting Research Position]

Faculty Members

Vincent Colotte [Univ de Lorraine, Associate Professor]
Irène Illina [Univ de Lorraine, Associate Professor, HDR]
Odile Mella [Univ de Lorraine, Associate Professor]
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, granted by ANR & Région Grand-Est]
Manfred Pastatter [Inria, from Jun 2019, partially granted by Région Grand-Est]
Imran Sheikh [Inria, from Jun 2019, granted by Europe]

PhD Students

Théo Biasutto-Lervat [Univ de Lorraine, granted by ANR]
Tulika Bose [Univ de Lorraine, from Sep 2019]
Guillaume Carbajal [Invoxia]
Pierre Champion [Inria, from Oct 2019, granted by ANR & Région Grand-Est]
Sara Dahmani [Univ de Lorraine]
Diego Di Carlo [Inria]
Ioannis Douros [Univ de Lorraine]
Sandipana Dowerah [Inria, from Oct 2019, granted by ANR]
Ashwin Geet Dsa [Univ de Lorraine, from Apr 2019, granted by ANR]
Adrien Dufraux [Facebook AI Research]
Raphaël Duroselle [Ministère des Armées]
Mathieu Fontaine [Inria, until Jun 2019, granted by ANR & Région Lorraine]
Nicolas Furnon [Univ de Lorraine, granted by ANR]
Amal Houidhek [Univ. de Lorraine & École Nationale d'Ingénieurs de Tunis, Tunisia]
Ajinkya Kulkarni [Univ de Lorraine]
Lou Lee [Univ de Lorraine]
Mohamed Amine Menacer [Univ de Lorraine]
Mauricio Michel Olvera Zambrano [Inria, from Oct 2019, granted by ANR]
Manuel Pariente [Univ de Lorraine, granted by École Normale Supérieure]
Lauréline Perotin [Orange Labs, until Nov 2019]
Shakeel Ahmad Sheikh [Univ de Lorraine, from Oct 2019, granted by ANR & Région Grand-Est]
Sunit Sivasankaran [Inria, granted by ANR]
Anastasiia Tsukanova [Univ de Lorraine, until Jul 2019]
Nicolas Turpault [Inria, partially granted by Région Grand-Est]
Nicolas Zampieri [Inria, from Nov 2019, partially granted by Région Grand-Est]
Georgios Zervakis [Inria, from Nov 2019]

Technical staff

Ismaël Bada [CNRS, from Sep 2019, granted by ANR]
Yassine Boudi [Inria, until Mar 2019]
Zaineb Chelly Dagdia [Inria, from Mar 2019, granted by Europe]
Louis Delebecque [Inria, from Sep 2019, granted by Région Grand-Est]
Valérian Girard [Inria]

Thomas Girod [Univ de Lorraine, until Oct 2019, granted by ANR]
Seyed Ahmad Hosseini [Inria, from Oct 2019, granted by Europe]
Mathieu Hu [Inria]
Stephane Level [CNRS, from Nov 2019, granted by ANR]
Leon Rohrbacher [Univ de Lorraine, from Oct 2019, granted by ANR]
Mehmet Ali Tugtekin Turan [Inria, from Jul 2019, granted by Europe]

Interns and Apprentices

Louis Abel [Univ de Lorraine, from Apr 2019 until May 2019]
Corto Bastien [Sorbonne Université, Paris, from May 2019 until Aug 2019]
Luc Cheng [Univ de Lorraine, from Apr 2019 until Jun 2019]
Stephanie Deckert [Univ de Lorraine, Jun 2019]
Sandipana Dowerah [Univ de Lorraine, from Mar 2019 until Aug 2019]
Pierre Goncalves [Univ de Lorraine, from May 2019 until Jul 2019]
Alexis Houssard [Ecole Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise, Evry, from Jul 2019 until Aug 2019]
Léna Joyeux [Université de Technologie de Compiègne, Compiègne, until Feb 2019]
Annya Kar [Univ de Lorraine, from Feb 2019 until Jul 2019]
Mathieu Leclaire [Univ de Lorraine, from May 2019 until Jul 2019]
Stephane Level [Université de Lorraine, from Apr 2019 until Sep 2019]
Romain Marlier [Univ de Lorraine, from Jun 2019 until Jul 2019]
Quentin Noirot [Univ de Lorraine, from Apr 2019 until Jun 2019]
Mauricio Michel Olvera Zambrano [National Autonomous University of Mexico, Mexico, from Apr 2019 until Aug 2019]

Administrative Assistants

Hélène Cavallini [Inria]
Delphine Hubert [Univ de Lorraine]
Martine Kuhlmann [CNRS, until Apr 2019]
Anne-Marie Messaoudi [CNRS, since May 2019]

Visiting Scientist

Brij Mohan Lal Srivastava [Université de Lille, Lille, since Sep 2019]

External Collaborator

Imene Zangar [Ecole Nationale d'Ingénieurs de Tunis, Tunisia]

2. Overall Objectives

2.1. 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 pronunciations 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 project proposal finalized 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 intend to 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 will 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 will 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 will also keep studying **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 will 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 will rely on further 2D and 3D modeling of the vocal tract as well as of the **dynamics of the vocal tract** from real-time MRI data. The prediction of glottis opening will also be 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 will help characterize links between the vocal tract and the face, and illustrate inner mouth articulation to learners. The suspension of articulatory movements in stuttering speech will also be 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, will continue 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 will be considered, and development will continue towards a multilingual talking head. We will investigate further the modeling of **expressivity** both for audio-only and for audiovisual speech synthesis. We will 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 will be set on the impact of L2 prosody on segmental realizations. Prosody will also be considered for its implication on the structuration of speech communication, including on discourse particles. Moreover, we will 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 will 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 will keep working on source localization in the presence of nearby acoustic reflectors. We will 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 will 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 will 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 will 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 will 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

4.1. Introduction

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. Foreseen application domains are related to multimodal computer interaction, annotation and processing of spoken documents, health and autonomy (more precisely aided communication and monitoring), and computer assisted learning.

4.2. Multimodal Computer Interactions

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 intermediate 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 interactions.

4.3. Annotation and Processing of Spoken Documents and Audio Archives

A first type of annotation consists in transcribing a spoken document in order to get the corresponding sequences of words, with possibly some complementary information, such as the structure (punctuation) or the modality (affirmation/question) of the utterances to make the reading and understanding easier. Typical applications of the automatic transcription of radio or TV shows, or of any other spoken document, include making possible their access by deaf people, as well as by text-based indexing tools.

A second type of annotation is related to speech-text alignment, which aims at determining the starting and ending times of the words, and possibly of the sounds (phonemes). This is of interest in several cases such as for annotating speech corpora for linguistic studies, and for synchronizing lip movements with speech sounds (for example, for avatar-based communications). Although good results are currently achieved on clean data, automatic speech-text alignment needs to be improved for properly processing noisy spontaneous speech data and needs to be extended to handle overlapping speech.

Finally, there is also a need for speech signal processing techniques in the field of multimedia content creation and rendering. Relevant techniques include speech and music separation, speech equalization, speech enhancement, prosody modification, and speaker conversion.

4.4. Aided Communication and Monitoring

Source separation techniques should help for locating and monitoring people through the detection of sound events inside apartments, and speech enhancement is mandatory for hands-free vocal interactions. A foreseen application aims at improving the autonomy of elderly or disabled people, and also fits with smartroom applications. In a longer perspective, adapting speech recognition technologies to the voice of elderly people should also be useful for such applications, but this requires the recording of adequate databases. Sound monitoring in other application fields (security, environmental monitoring) can also be envisaged.

4.5. Computer Assisted Learning

Although speaking seems quite natural, learning foreign languages, or learning the 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 precision and quality of the segmentation of the speech utterance into phones, and of the computed prosodic parameters.

5. Highlights of the Year

5.1. Highlights of the Year

We developed the first deep learning-based multichannel speech enhancement algorithm that jointly reduces acoustic echo, reverberation, and background noise [57].

E. Vincent gave a keynote at the Voice Tech Paris 2019 trade fair [18].

A. Deleforge organized the IEEE Signal Processing Cup 2019 on "Search & Rescue with Drone-Embedded Sound Source Localization", to which 20 teams of undergraduate students from 18 universities in 11 countries participated, for a total of 132 participants [5]. The final took place on May the 13th at the international conference ICASSP in Brighton. The associated DREGON dataset, which was made publicly available afterwards, has received over 1,000 file downloads as of December 2019.

5.1.1. Awards

L. Perotin obtained the Best Poster Award of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) .

BEST PAPER AWARD:

[43]

L. PEROTIN, A. DÉFOSSEZ, E. VINCENT, R. SERIZEL, A. GUÉRIN. *Regression versus classification for neural network based audio source localization*, in "WASPAA 2019 - IEEE Workshop on Applications of Signal Processing to Audio and Acoustics", New Paltz, United States, IEEE, October 2019, https://hal.inria.fr/hal-02125985

6. New Software and Platforms

6.1. dnnsep

Multichannel audio source separation with deep neural networks

KEYWORDS: Audio - Source Separation - Deep learning

SCIENTIFIC DESCRIPTION: dnnsep is the only source separation software relying on multichannel Wiener filtering based on deep learning. Deep neural networks are used to initialize and reestimate the power spectrum of the sources at every iteration of an expectation-maximization (EM) algorithm.

FUNCTIONAL DESCRIPTION: Combines deep neural networks and multichannel signal processing for speech enhancement and separation of musical recordings.

NEWS OF THE YEAR: Version 1.1 was slightly modified in order to issue a test license to the French Ministry of Interior.

- Participants: Aditya Nugraha, Emmanuel Vincent and Antoine Liutkus
- Contact: Emmanuel Vincent

6.2. KATS

Kaldi-based Automatic Transcription System

KEYWORD: Speech recognition

FUNCTIONAL DESCRIPTION: KATS is a multipass system for transcribing audio data, and in particular radio or TV shows in French, English or Arabic. It is based on the Kaldi speech recognition tools. It relies on Deep Neural Network (DNN) modeling for speech detection and acoustic modeling of the phones (speech sounds). Higher order statistical language models and recurrent neural network language models can be used for improving performance through rescoring of multiple hypotheses.

NEWS OF THE YEAR: New models have been trained for German, as well as two bilingual models: one combining French and German phones, and one combining French and English phones. Also, a web server has been set up for on line real-time speech recognition.

- Participants: Dominique Fohr, Odile Mella, Mathieu Hu, Denis Jouvet and Irina Illina
- Contact: Dominique Fohr

6.3. SOJA

Speech Synthesis platform in JAva

KEYWORDS: Speech Synthesis - Audio

SCIENTIFIC DESCRIPTION: SOJA relies on a non-uniform unit selection algorithm. Phonetic and linguistic features are extracted and computed from the text to drive the selection of speech units in a recorded corpus. The selected units are concatenated to obtain the speech signal corresponding to the input text.

FUNCTIONAL DESCRIPTION: SOJA is a software for Text-To-Speech synthesis (TTS). It performs all steps from text input to speech signal output. A set of associated tools is available for elaborating a corpus for a TTS system (transcription, alignment, etc.). Currently, the corpus contains about 3 hours of speech recorded by a female speaker. Most of the modules are in Java, some are in C. The SOJA software runs under Windows and Linux. It can be launched with a graphical user interface or directly integrated in a Java code or by following the client-server paradigm.

RELEASE FUNCTIONAL DESCRIPTION: Version 3.0 integrates a phonetization based on a deep learning algorithm. In addition, the phonetization step is managed by API REST (client/server mode). The NLP part provides an output of descriptors in the format that can be used by HTS and Merlin systems.

NEWS OF THE YEAR: The latest version can use the LORIA-PHON deep learning based grapheme-to-phoneme converter through a web API.

- Participants: Alexandre Lafosse and Vincent Colotte
- Contact: Vincent Colotte

6.4. LORIA-PHON

LORIA grapheme-to-phoneme converter

KEYWORDS: Grapheme-to-phoneme converter - Neural networks

FUNCTIONAL DESCRIPTION: LORIA-PHON is a deep-learning based software for grapheme-to-phoneme conversion. It currently works for French. A web API is available for using it in a client/server mode. It properly interfaces with the SOJA software used for speech synthesis.

NEWS OF THE YEAR: new software

- Participants: Mathieu Hu, Denis Jouvet, Vincent Colotte and Louis Delebecque
- Contact: Vincent Colotte

6.5. Dynalips-Player

High realistic lip synchronization for 3d animated characters

KEYWORDS: 3D animation - Graphics - Speech Synthesis

FUNCTIONAL DESCRIPTION: Dynalips provides a solution to synchronize precisely and automatically the movements of the lips of a 3D character with speech (we address 3D animation movies and video games). We have developed a demonstrator that illustrates the whole process: from audio + text to the generation of the animation trajectory, and controlling the animation of a 3D model (e.g. an avatar). The demonstrator is composed mainly by the player developed in Unity 3D (but can be used with any other system) and plays the animation synchronously with speech in realtime. It is possible to generate an animation for Autodesk Maya 3D.

NEWS OF THE YEAR: The player has been extended to be multilingual thanks to two developments within two projects. In fact, within the METAL project, a lipsync for German has been developed. In addition, within the ATT Dynalips, we have built a lipsync for English.

- Partners: Université de Lorraine - Sayens (SATT Grand Est)
- Contact: Slim Ouni
- URL: http://www.dynalips.com

6.6. VisArtico

Visualization of multimodal speech data

KEYWORDS: Data visualization - 3D movement - Speech processing - Videos

SCIENTIFIC DESCRIPTION: VisArtico is a visualization software of multimodal data. It is possible to visualize the positions of real or virtual sensors and to animate them simultaneously with acoustics. This software can be useful for researchers in speech production, audiovisual speech synthesis or articulatory speech analysis.

FUNCTIONAL DESCRIPTION: VisArtico is a user-friendly software which allows visualizing multimodal data acquired by several systems: an articulograph, motion capture system, depth camera. This visualization software has been designed so that it can directly use the data provided by the different systems to display the spatial and temporal positions of the sensors (real and virtual). Moreover, VisArtico allows viewing the sensors augmented with visual information by indicating graphically the data for the tongue, lips and jaw.

RELEASE FUNCTIONAL DESCRIPTION: The current version allows the user to manage different modalities (articulatory, gestural, acoustic and video). It is possible to do automatic alignment, or even speech recognition. Several spatial data processing tools have been added (referential change, head movement suppression, merging data from multiple sources, ...).

NEWS OF THE YEAR: The software has undergone several improvements. Mainly, several branches have been merged in order to have as many features as possible available within the master branch.

- Participants: Ilef Ben Farhat, Loïc Mangeonjean, Slim Ouni and Louis Abel
- Partners: CNRS - Université de Lorraine
- Contact: Slim Ouni
- Publication: VisArtico: a visualization tool for articulatory data
- URL: http://visartico.loria.fr

6.7. Xarticulators

KEYWORDS: Medical imaging - Natural language processing

FUNCTIONAL DESCRIPTION: The Xarticulators software is intended to delineate contours of speech articulators in X-ray and MR images, construct articulatory models and synthesize speech from X-ray films. This software provides tools to track contours automatically, semi-automatically or by hand, to make the visibility of contours easier, to add anatomical landmarks to speech articulators and to synchronize images with the sound. In addition we also added the possibility of processing digitized manual delineation results made on sheets of papers when no software is available. Xarticulators also enables the construction of adaptable linear articulatory models from the X-ray or MR images and incorporates acoustic simulation tools to synthesize speech signals from the vocal tract shape. Recent work was on the possibility of synthesizing speech from 2D-MRI films, and on the construction of better articulatory models for the velum, lips and epiglottis.

RELEASE FUNCTIONAL DESCRIPTION: The new version allows MRI films to be processed and, above all, it offers a better transition from the shape of the vocal tract to the area function, which corresponds to an approximation of the vocal tract using a series of elementary tubes from the glottis to the lips.

NEWS OF THE YEAR: This year we completed the software to evaluate the articulatory model built from static images on dynamic images and we added a module to monitor the contour of the language using deep learning.

- Contact: Yves Laprie
- Publication: Articulatory model of the epiglottis

6.8. DCASE 2019 baseline

Baseline system for the task 4 of DCASE 2019 Challenge

KEYWORDS: Audio signal processing - Audio source classification - Machine learning - Smart home

FUNCTIONAL DESCRIPTION: This is the baseline system for the task 4 of the challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) 2019. The algorithm performs sound events detection and classification. From an audio recording, the target of the system is to provide not only the event class but also the event time boundaries given that multiple events can be present in an audio recording. The baseline approach relies on convolutionnal and recurrent neural networks (CRNN) and a mean-teacher model to exploit a large amount of unbalanced and unlabeled training data together with a small weakly annotated (without timestamps) training set and a strongly annotated (with timestamps) synthetic set to improve system performance.

RELEASE FUNCTIONAL DESCRIPTION: This version includes a mean teacher model to exploit the various training sets with that have different levels of annotations, as provided in the task 4 of the DCASE 2019 challenge.

- Participants: Nicolas Turpault and Romain Serizel
- Contact: Nicolas Turpault
- Publication: Sound event detection in domestic environments with weakly labeled data and soundscape synthesis
- URL: https://github.com/turpaultn/DCASE2019_task4/tree/public/baseline

7. New Results

7.1. Beyond black-box supervised learning

Participants: Emmanuel Vincent, Denis Jouvet, Antoine Deleforge, Vincent Colotte, Irène Illina, Romain Serizel, Imran Sheikh, Pierre Champion, Adrien Dufraux, Ajinkya Kulkarni, Manuel Pariente, Georgios Zervakis, Zaineb Chelly Dagdia, Mehmet Ali Tugtekin Turan, Brij Mohan Lal Srivastava.

This year marked a significant increase in our research activities on domain-agnostic challenges relating to deep learning, such as the integration of domain knowledge, data efficiency, or privacy preservation. Our vision was illustrated by a keynote [18] and several talks [19], [17] on the key challenges and solutions.

7.1.1. Integrating domain knowledge

7.1.1.1. 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 [59]. Second, in order to allow generalization to mixtures of sources not seen together in training, we explored the modeling of speech spectra by variational autoencoders (VAEs), which are a variant of the probabilistic generative models classically used in source separation before the deep learning era. The VAEs are trained separately for each source and used to infer the source signals underlying a given mixture. Compared with existing iterative inference algorithms involving Gibbs sampling or gradient descent, we proposed a computationally efficient variational inference method based on an analytical derivation in which the encoder of the pre-learned VAE can be used to estimate the variational approximation of the true posterior [42], [55].

7.1.2. Learning from little/no labeled data

7.1.2.1. Learning from noisy labels

ASR systems are typically trained in a supervised fashion using manually labeled data. This labeling process incurs a high cost. Classical semi-supervised learning and transfer learning approaches to reduce the transcription cost achieve limited performance because the amount of knowledge that can be inferred from unlabeled data is intrinsically lower. We explored the middle ground where the training data are neither accurately labeled nor unlabeled but a not-so-expensive “noisy” transcription is available instead. We proposed a method to learn an end-to-end ASR model given a noise model and a single noisy transcription per utterance by adapting the auto segmentation criterion (ASG) loss to account for several possible transcriptions. Because the computation of this loss is intractable, we used a differentiable beam search algorithm that samples only the best alignments of the best transcriptions [32].

7.1.2.2. Transfer learning

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. In [36], we proposed to transfer the expressive characteristics through layer adaptation during the learning step. The obtained results highlighted that there is a difficult trade-off between speaker's identity to remove and the expressivity to transfer. We are now working on an approach relying on multiclass N-pair based deep metric learning in recurrent conditional variational autoencoder (RCVAE) for implementing a multispeaker expressive text-to-speech (TTS) 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. The deep metric learning helps to reduce the intra-class variance and increase the inter-class variance. We transfer the expressivity by using the latent variables for each emotion to generate expressive speech in the voice of a different speaker for which no expressive speech is available. The performance measured shows the model's capability to transfer the expressivity while preserving the speaker's voice in synthesized speech.

7.1.3. 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 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 [64]. As an alternative, we proposed an adversarial approach to learn representations that perform well for ASR while hiding speaker identity. Our results demonstrate that adversarial training dramatically reduces the closed-set speaker classification accuracy, but this does not translate into increased open-set speaker verification error [45]. We are currently organizing the 1st Voice Privacy Challenge in which these and other approaches will be further assessed and compared.

7.2. Speech production and perception

7.2.1. Articulatory modeling

Participants: Denis Jouvet, Anne Bonneau, Dominique Fohr, 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, Anastasiia Tsukanova, Louis Delebecque, Valérian Girard, Thomas Girod, Seyed Ahmad Hosseini, Mathieu Hu, Leon Rohrbacher, Imene Zangar.

7.2.1.1. Articulatory speech synthesis

A number of simplifying assumptions have to be made in articulatory synthesis to enable the speech signal to be generated in a reasonable time. They mainly consist of approximating the propagation of the sound in the vocal tract as a plane wave and approximating the 3D vocal tract shape from the mid-sagittal shape [30], and also simplifying the vocal tract topology by removing small cavities [29]. The posture of the subject in the MRI machine was also investigated [31]. Vocal tract resonances were evaluated from the 3D acoustic simulation computed with the K-wave Matlab package from the complete 3D vocal tract shape recovered from MRI and compared to those of real speech [27].

We also developed an approach for using articulatory features for speech synthesis. The approach is based on a deep feed-forward neural network-based speech synthesizer trained with the standard recipe of Merlin on the audio recorded during real-time MRI (RT-MRI) acquisitions: denoised (and yet containing a residual noise of the MRI machine) speech in French and force-aligned state labels encoding phonetic and linguistic information [26]. The synthesizer was augmented with eight parameters representing articulatory information (lips opening and protrusion, distances between the tongue and the velum, between the velum and the pharyngeal wall, and between the tongue and the pharyngeal wall) that were automatically extracted from the captures and aligned with the audio signal and the linguistic specification.

7.2.1.2. Dynamics of vocal tract and glottal opening

The problem of creating a 3D dynamic atlas of the vocal tract that captures the dynamics of the articulators in all three dimensions has been addressed [28]. The core steps of the method are using 2D real time MRI in several sagittal planes and, after temporal alignment, combine them using adaptive kernel regression. As a preprocessing step, a reference space was created to be used in order to remove anatomical information of the speakers and keep only the variability in speech production for the construction of the atlas. Using adaptive kernel regression makes the choice of atlas time points independent of the time points of the frames that are used as an input for the atlas construction.

We started the development of a database of realistic glottal gestures which will be used to design the glottal opening dynamics in articulatory synthesis paradigms. Experimental measurements of glottal opening dynamics in VCV and VCCV sequences uttered by real subjects have been achieved thanks to a specifically designed external photoglottographic device (ePGG) [33]. The existence of different patterns of glottal opening is evidenced according to the class of the consonant articulated.

7.2.1.3. Multimodal coarticulation modeling

We have investigated labial coarticulation to animate a virtual face from speech. We experimented a sequential deep learning model, bidirectional gated recurrent networks, that have been used successfully in addressing the articulatory inversion problem. We have used phonetic information as input to ensure speaker independence. 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, allowing injecting knowledge of useful latent representation of the visual data into the network. We have trained and evaluated the model with a corpus consisting of 4 hours of French speech, and we got a good average RMSE (Root Mean Square Error) close to 1.3 mm [21].

7.2.1.4. 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 analyzing existing stuttering acoustic speech datasets to characterize the kind of data.

7.2.2. Multimodal expressive speech

7.2.2.1. Arabic speech synthesis

We have continued working on Modern Standard Arabic speech synthesis with ENIT (École Nationale d'Ingénieurs de Tunis, Tunisia), using HMM and NN based approaches. This year we investigated the modeling of the fundamental frequency for Arabic speech synthesis with feedforward and recurrent DNN, and using specific linguistic features for Arabic like vowel quantity and gemination [50].

7.2.2.2. Expressive audiovisual synthesis

After acquiring 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. We used conditional variational autoencoders (CVAE) to generate the duration, acoustic and visual aspects of speech 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 [23].

7.2.2.3. Lipsync - synchronization of lips movements with speech

In the ATT Dynalips-2, we have developed an English version of the system which allows us having a full multilingual lipsync system. During this ATT, we also worked on the business aspects (business plan, funding, investment, search for clients, etc.) with the goal of creating a startup, spinoff of the laboratory, during 2020.

7.2.3. Categorization of sounds and prosody

7.2.3.1. Non-native speech production

We analysed voicing in sequences of obstruents with French as L1 and German as L2, that is languages characterized by strong differences in the voicing dimension, including assimilation direction. To that purpose, we studied the realizations of two sequences of obstruents, where the first consonant, in final position, was fortis, and the second consonant, in initial position, was either a lenis stop or a lenis fricative. These sequences lead to a possible anticipation of voicing in French, a direction not allowed in German given German phonetics and phonology. Highly variable realizations were observed: progressive and regressive assimilation, and absence of assimilation, often accompanied by an unexpected pause [22].

We also started investigating non-native phoneme productions of French learners of German in comparison to phoneme productions by native German speakers. A set of research questions has been developed for which a customized French/German corpus was designed, and recorded by one reference native speaker of German so far. Based on these initial recordings and according to the targeted research questions, analysis strategies and algorithms have been elaborated and implemented, and are ready to be employed onto a larger data set. By means of these methods we expect to access phonetic and phonological grounds of recurrently occurring mis-pronunciation.

7.2.3.2. Language and reading acquisition by children having some language impairments

We continued examining the schooling experience of 170 children, teenagers and young adults with specific language impairment (dysphasia, dyslexia, dysorthographia) facing severe difficulties in learning to read. The phonemic discrimination, phonological and phonemic analysis difficulties faced in their childhoods had raised reading difficulties, which the pupils did not overcome. With 120 of these young people, we explored the presence of other neuro-developmental disorders. We also studied their reading habits to achieve better understanding of their difficulties.

We continued investigating the acquisition of language by hard-of-hearing children via cued speech (i.e. augmenting the audiovisual speech signal by visualizing the syllables uttered via a code of hand positions). We have used a digital book and a children's picture book with 3 hard-of-hearing children in order to compare scaffolding by the speech therapist or the teacher in these two situations.

We started to examine language difficulties and related problems with children with autism and to work with their parents with a view to creating an environment conducive to their progress [39].

7.2.3.3. Computer assisted language learning

In the METAL project, experiments are planned to investigate the use of speech technologies for foreign language learning and to experiment with middle and high school students learning German. This includes tutoring aspects based on 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 under development, and experiments have continued for analyzing the performance of an automatic detection of mispronunciations made by language learners.

The ALOE project deals with children learning to read. In this project, we are also 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).

7.2.3.4. Prosody

The keynote [15] summarizes recent research on speech processing and prosody, and presents the extraction of prosodic features, as well as their usage in various tasks. Prosodic correlates of discourse particles have been investigated further. It was found that occurrences of different discourse particles with the same pragmatic value have a great tendency to share the same prosodic pattern; hence, the question of their commutability have been studied [37].

7.3. Speech in its environment

Participants: Denis Jouvet, Antoine Deleforge, Dominique Fohr, Emmanuel Vincent, Md Sahidullah, Irène Illina, Odile Mella, Romain Serizel, Tulika Bose, Guillaume Carbajal, Diego Di Carlo, Sandipana Dowerah, Ashwin Geet Dsa, Adrien Dufraux, Raphaël Duroselle, Mathieu Fontaine, Nicolas Furnon, Mohamed Amine Menacer, Mauricio Michel Olvera Zambrano, Lauréline Perotin, Sunit Sivasankaran, Nicolas Turpault, Nicolas Zampieri, Ismaël Bada, Yassine Boudi, Mathieu Hu, Stephane Level.

7.3.1. Acoustic environment analysis

We are constantly surrounded by ambient sounds and rely heavily on them to obtain important information about our environment. Deep neural networks are useful to learn relevant representations of these sounds. Recent studies have demonstrated the potential of unsupervised representation learning using various flavors of the so-called triplet loss (a triplet is composed of the current sample, a so-called positive sample from the same class, and a negative sample from a different class), and compared it to supervised learning. To address real situations involving both a small labeled dataset and a large unlabeled one, we combined unsupervised and supervised triplet loss based learning into a semi-supervised representation learning approach and compared it with supervised and unsupervised representation learning depending on the ratio between the amount of labeled and unlabeled data [49].

Pursuing our involvement in the community on ambient sound recognition, we co-organized a task on large-scale sound event detection as part of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 Challenge [48]. It focused on the problem of learning from audio segments that are either weakly labeled or not labeled, targeting domestic applications. We also published a summary of the outcomes of the DCASE 2017 Challenge, in which we had organized the first version of that task [7] and a detailed analysis of the submissions to that task in 2018 [16] and 2019 [61].

7.3.2. Speech enhancement and noise robustness

7.3.2.1. Sound source localization and counting

In multichannel scenarios, source localization, counting and separation are tightly related tasks. Concerning deep learning based speaker localization, we introduced the real and imaginary parts of the acoustic intensity vector in each time-frequency bin as suitable input features. We analyzed the inner working of the neural network using layerwise relevance propagation [9]. We also defined alternative regression-based approaches for localization and compared them to the usual classification-based approach on a discrete grid [43]. Lauréline Perotin successfully defended her PhD on this topic [2]. In [24], we proposed the first deep-learning based method for blindly estimating early acoustic echoes. We showed how estimates of these echoes enable 2D sound source localization with only two microphones near a reflective surface, a task normally impossible with traditional methods. Finally, we published our former work on motion planning for robot audition [8].

We organized the IEEE Signal Processing Cup 2019, an international competition aimed at teams of undergraduate students [5]. The tasks we proposed were on sound source localization using an array embedded in a flying drone for search and rescue application. Submissions to the first phase of the competition were opened from November 2018 to March 2019, and the final took place on May the 13th at the international conference ICASSP in Brighton. 20 teams of undergraduate students from 18 universities in 11 countries participated, for a total of 132 participants. The drone-embedded sound source localization dataset we recorded for the challenge was made publically available after the competition and has received over 1,000 file downloads as of December 2019.

7.3.2.2. Speech enhancement

We investigated the effect of speaker localization accuracy on deep learning based speech enhancement quality. To do so, we generated a multichannel, multispeaker, reverberated, noisy dataset inspired from the well studied WSJ0-2mix and evaluated enhancement performance in terms of the word error rate. We showed that the signal-to-interference ratio between the speakers has a higher impact on the ASR performance than the angular distance [62]. In addition, we proposed a deflation method which estimates the sources iteratively. At each iteration, we estimate the location of the speaker, derive the corresponding time-frequency mask and remove the estimated source from the mixture before estimating the next one [63].

In parallel, we introduced a method for joint reduction of acoustic echo, reverberation and noise. This method models the target and residual signals after linear echo cancellation and dereverberation using a multichannel 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 [53], [57].

In the context of ad-hoc acoustic antennas, we proposed to extend the distributed adaptive node-specific signal estimation approach to a neural networks 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 multichannel 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 performances than when the mask estimation relies only on the local signals [58].

We have been pursuing our work on non-Gaussian heavy-tail models for signal processing, and notably investigated whether such models could be of use to devise new cost functions for the training of deep generative models for source separation [34]. In the case of speech enhancement, it turned out that the related log-likelihood functions could advantageously replace the more constraining squared-error and lead to significant performance gains.

We have also been pursuing our theoretical work on multichannel alpha-stable models, devising two new multichannel filtering methods that are adequate for processing multivariate heavy-tailed vectors. The related work is presented in Mathieu Fontaine's PhD manuscript [1].

7.3.2.3. 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. We are now organizing the 6th edition of the Challenge, and released the French dataset for ambient assisted living applications previously collected as part of the FUI VOICEHOME project [4].

7.3.2.4. Speaker recognition

Automatic speaker recognition systems give reasonably good recognition accuracy when adequate amount of speech data from clean conditions are used for enrollment and test. However, performance degrades substantially in real-world noisy conditions as well as due to the lack of adequate speech data. Apart from these two practical limitations, speaker recognition performance also degrades in presence of spoofing attacks [51] where playback voice or synthetic speech generated with voice conversion or speech synthesis methods are used by attackers to access a system protected with voice biometrics.

We have explored a new speech quality measure for quality-based fusion of speaker recognition systems. The quality metric is formulated with the zero-order statistics estimated during i-vector extraction. The proposed quality metric is shown to capture the speech duration information, and it has outperformed absolute-duration based quality measures when combining multiple speaker recognition systems. Noticeable improvement over existing methods have been observed specifically for the short-duration conditions [10].

We have also participated in speaker recognition evaluation campaigns NIST SREs and VoxSRC. For the NIST SREs [54], the key problem was to recognize speakers from low-quality telephone conversations. In addition, the language mismatch between system development and data under test made the problem more challenging. In VoxSRC, on the other hand, the main problem was to recognize speakers speaking short sentences of about 10 sec where the speech files are extracted from Youtube video clips. We have explored acoustic feature extraction, domain adaptation, parameter optimization and system fusion for these challenges. For VoxSRC, our system has shown substantial improvement over baseline results.

We also introduced a statistical uncertainty-aware method for robust i-vector based speaker verification in noisy conditions, that is the first one to improve over simple chaining of speech enhancement and speaker verification on the challenging NIST-SRE corpus mixed with real domestic noise and reverberation [44].

Robust speaker recognition is an essential component of speaker diarization systems. We have participated in the second DIHARD challenge where the key problem was the diarization of speech signals collected from diverse real-world conditions. We have explored speech activity detection, domain grouping, acoustic features, and speech enhancement for improved speaker recognition. Our proposed system has shown considerable improvement over the Kaldi-based baseline system provided by the challenge organizer [60].

We have co-organized the ASVspoof 2019 challenge, as an effort to develop next-generation countermeasures for automatic detection of spoofed/fake audio [46]. This involved creating the audio dataset, designing experiments, evaluating and analyzing the results. 154 teams or individuals participated in the challenge. The database is available for research and further exploration from Edinburgh DataShare, and has been downloaded/viewed more than a thousand times so far.

We have also analyzed whether target speaker selection can help in attacking speaker recognition systems with voice impersonation [35]. Our study reveals that impersonators were not successful in attacking the systems, however, the speaker similarity scores transfer well from the attacker's system to the attacked system [12]. Though there were modest changes in F0 and formants, we found that the impersonators were able to considerably change their speaking rates when mimicking targets.

7.3.2.5. Language identification

State-of-the-art spoken language identification systems are constituted 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 of the embedding extractor. The proposed approach consists in a modification of the loss of the segment level embedding extractor by adding a regularisation term. 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 obtain the same performance as published language identification results but without using labelled data from the target domain.

7.3.3. Linguistic and semantic processing

7.3.3.1. Transcription, translation, summarization and comparison of videos

Within the AMIS project, we studied different subjects related to the processing of videos. The first one concerns the machine translation of Arabic-English code-switched documents [41]. Code-switching is defined as the use of more than one language by a speaker within an utterance. The second one deals with the summarization of videos into a target language [11]. This exploits research carried on in several areas including video summarization, speech recognition, machine translation, audio summarization and speech segmentation. One of the big challenges of this work was to conceive a way to evaluate objectively a system composed of several components given that each of them has its limits and that errors propagate through the components. A third aspect was a method for extracting text-based summarization of Arabic videos [40]. The automatic speech recognition system developed to transcribe the videos has been adapted to the Algerian dialect, and additional modules were developed for segmenting the flow of recognized word into sentences, and for summarization. Finally the last aspect concerns the comparison of the opinions of two videos in two different languages [20]. Evaluations have been carried on comparable videos extracted from a corpus of 1503 Arabic and 1874 English videos.

7.3.3.2. 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 proposed a new approach for the classification of tweets, aiming to predict whether a tweet is abusive, hate or neither. We compare different unsupervised word representations and DNN classifiers, and study the robustness of the proposed approaches to adversarial attacks when adding one (healthy or toxic) word. We are evaluating the proposed methodology on the English Wikipedia Detox corpus and on a Twitter corpus.

7.3.3.3. Introduction of semantic information in an automatic speech recognition system

In current state-of-the-art automatic speech recognition systems, N-gram based models are used to take into account language information. They have a local view and are mainly based on syntax. The introduction of semantic information and longer term information in a recognition system should make it possible to remove some ambiguities and reduce the error rate of the system. Within the MMT project, we are proposing and evaluating methods for integrating semantic information into our speech recognition system through the use of various word embeddings.

7.3.3.4. Music language modeling

Similarly to speech, language models play a key role in music modeling. We represented the hierarchical structure of a temporal scenario (for instance, a chord progression) via a phrase structure grammar and proposed a method to automatically induce this grammar from a corpus and to exploit it in the context of machine improvisation [6].

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. Studio Maia

Company: Studio Maia SARL (France)

Other partners: Imaging Factory

Duration: Jul 2017 – March 2019

Participants: Yassine Boudi, Vincent Colotte, Mathieu Hu, Emmanuel Vincent

Abstract: We developed a software suite for voice processing in the multimedia creation chain. The software was designed for sound engineers, and relied on the team's expertise in speech enhancement, robust speech and speaker recognition, and speech synthesis.

8.1.2. Honda Research Institute Japan

Company: Honda Research Institute Japan (Japan)

Duration: Aug 2018 – Mar 2019

Participants: Nancy Bertin (CNRS - IRISA), Antoine Deleforge, Diego Di Carlo

Abstract: This was a follow-up contract targeting collaborative research on multichannel speech and audio processing and eventual software licensing in order to enable voice-based communication in challenging noisy and reverberant conditions in which current hands-free voice-based interfaces perform poorly.

8.1.3. Dassault and Thalès - Man Machine Teaming Initiative

Company: Dassault and Thalès (France)

Duration: Apr 2019 - Sept 2020

Participants: Irène Illina, Dominique Fohr, Ismael Bada, Stephane 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.

8.2. Bilateral Grants with Industry

8.2.1. Orange

Company: Orange SA (France)

Duration: Nov 2016 – Oct 2019

Participants: Lauréline Perotin, Romain Serizel, Emmanuel Vincent

Abstract: This CIFRE contract funded the PhD thesis of Lauréline Perotin. Our goal was to develop deep learning based speaker localization and speech enhancement algorithms for robust hands-free voice command. We were especially targeting difficult scenarios involving several simultaneous speakers.

8.2.2. Invoxia

Company: Invoxia SAS (France)

Duration: Mar 2017 – Apr 2020

Participants: Guillaume Carbajal, Romain Serizel, Emmanuel Vincent

Abstract: This CIFRE contract funds the PhD thesis of Guillaume Carbajal. Our goal is to design a unified end-to-end 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.

8.2.3. 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.

8.2.4. 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.

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. CPER LCHN

Project acronym: CPER LCHN

Project 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 contributes to automatic speech recognition, speech-text alignment and prosody aspects.

9.1.2. CPER IT2MP

Project acronym: CPER IT2MP

Project title: CPER “Innovation Technologique Modélisation et Médecine Personalisée”

Duration: 2015-2020

Coordinator: Faiez Zannad (Inserm-CHU-UL)

Participants: Romain Serizel, Emmanuel Vincent

Abstract: The goal is to develop innovative technologies for health, and tools and strategies for personalized medicine. MULTISPEECH will collect data for distant-microphone voice commands.

9.1.3. Com-Medic ALOE

Company: Com-Medic (France)

Duration: Mar 2019 – August 2020

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.

9.2. National Initiatives

9.2.1. ANR ArtSpeech

Project acronym: ArtSpeech

Project title: Synthèse articulatoire phonétique

Duration: October 2015 - August 2020

Coordinator: Yves Laprie

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

Participants: Ioannis Douros, Yves Laprie, Anastasiia Tsukanova

Abstract: The objective is 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.

9.2.2. ANR JCJC KAMoulox

Project acronym: KAMoulox

Project title: Kernel additive modelling for the unmixing of large audio archives

Duration: January 2016 - September 2019

Coordinator: Antoine Liutkus (Inria Zenith)

Participants: Mathieu Fontaine

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

9.2.3. PIA2 ISITE LUE

Project acronym: ISITE LUE

Project title: Lorraine Université d'Excellence

Duration: 2016 - 2020

Coordinator: Univ. Lorraine

Participants: Ioannis Douros, Yves Laprie

Abstract: LUE (Lorraine Université d'Excellence) was designed as an "engine" for the development of excellence, by stimulating an original dialogue between knowledge fields. Within challenge number 6: "Knowledge engineering" this project funds the PhD thesis of Ioannis Douros on articulatory modeling.

9.2.4. OLKI LUE

Project acronym: OLKI LUE

Project title: Open Language and Knowledge for Citizens, Lorraine Université d'Excellence

Coordinator: Christophe Cerisara (LORIA)

Participants: Tulika Bose, Dominique Fohr, Irène Illina

Abstract: The initiative aims at developing new algorithms that improve the automatic understanding of natural language documents, and a federated language resource distribution platform to enable and facilitate the sharing of open resources. This project funds the PhD thesis of Tulika Bose on the detection and classification of hate speech.

9.2.5. E-FRAN METAL

Project acronym: E-FRAN METAL

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

Duration: October 2016 - September 2020

Coordinator: Anne Boyer (LORIA)

Other partners: 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.

9.2.6. ANR VOCADOM

Project acronym: VOCADOM (http://vocadom.imag.fr/)

Project title: Robust voice command adapted to the user and to the context for ambient assisted living

Duration: January 2017 - December 2020

Coordinator: CNRS - LIG (Grenoble)

Other partners: 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.

9.2.7. ANR JCJC DiSCogs

Project acronym: DiSCogs

Project title: Distant speech communication with heterogeneous unconstrained microphone arrays

Duration: September 2018 – March 2022

Coordinator: Romain Serizel

Participants: 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.

9.2.8. ANR DEEP-PRIVACY

Project acronym: DEEP-PRIVACY

Project title: Distributed, Personalized, Privacy-Preserving Learning for Speech Processing

Duration: January 2019 - December 2022

Coordinator: Denis Jouvet

Other partners: 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.

9.2.9. ANR ROBOVOX

Project acronym: ROBOVOX

Project title: Robust Vocal Identification for Mobile Security Robots

Duration: Mar 2019 – Mar 2023

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

Other partners: Inria (Nancy), A.I. Mergence

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.

9.2.10. ANR LEAUDS

Project acronym: LEAUDS

Project title: Learning to understand audio scenes

Duration: Apr 2019 - Sep 2022

Coordinator: Université de Rouen Normandie

Other partners: 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.

9.2.11. Inria Project Lab HyAIAI

Project acronym: HyAIAI

Project title: Hybrid Approaches for Interpretable AI

Duration: Sep 2019 - Aug 2023

Coordinator: Inria LACODAM (Rennes)

Other 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.

9.2.12. ANR BENEPHIDIRE

Project acronym: BENEPHIDIRE

Project title: Stuttering: Neurology, Phonetics, Computer Science for Diagnosis and Rehabilitation

Duration: March 2019 - December 2023

Coordinator: Praxiling (Toulouse)

Other partners: 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.

9.2.13. ANR HAIKUS

Project acronym: HAIKUS

Project title: Artificial Intelligence applied to augmented acoustic Scenes

Duration: Dec 2019 - May 2023

Coordinator: Ircam (Paris)

Other partners: 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.

9.2.14. ANR Flash Open Science HARPOCRATES

Project acronym: HARPOCRATES

Project title: Open data, tools and challenges for speaker anonymization

Duration: Oct 2019 - Mar 2021

Coordinator: Eurecom (Nice)

Other partners: 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.

9.2.15. ATT Dynalips & ATT Dynalips-2

Project acronym: DYNALIPS

Project title: Automatic Lip synchronization with speech

Duration: Jul 2018 - Dec 2019

Coordinator: Slim Ouni

Participants: Valerian Girard, Slim Ouni

Abstract: This is a technology transfer project of our research solution that aims to synchronize precisely and automatically the movement of the mouth of a 3D character with speech. We address 3D animation and video game industries.

9.2.16. InriaHub Carnot Technologies Vocales

Project 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.

9.3. European Initiatives

9.3.1. FP7 & H2020 Projects

9.3.1.1. COMPRISE

Program: H2020 ICT-29-2018 (RIA)

Project acronym: COMPRISE

Project title: Cost-effective, Multilingual, Privacy-driven voice-enabled Services

Duration: Dec 2018- Nov 2021

Coordinator: Emmanuel Vincent

Other partners: Inria Magnet, Ascora GmbH, Netfective Technology SA, Rooter Analysis SL, Saarland University, Tilde SIA

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

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

9.3.1.2. AI4EU

Program: ICT-26-2018-2020

Project acronym: AI4EU

Project title: European Artificial Intelligence On-Demand Platform and Ecosystem

Duration: 2019–2021

Coordinator: THALES

Other partners: 80 partners from 22 countries

Participants: Seyed Ahmad Hosseini, Slim Ouni

Abstract: The aim of AI4EU is to develop a European Artificial Intelligence ecosystem, from knowledge and algorithms to tools and resources.

9.3.1.3. CPS4EU

Program: PSPC-ECSEL

Project acronym: CPS4EU

Project title: Cyber-physical systems for Europe

Duration: June 2019 – June 2022

Coordinator: CEA

Other partners: 42 partners from 6 countries

Participants: Antoine Deleforge, Romain Serizel

Abstract: 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 around CPS enabling technologies and products.

9.3.2. Collaborations in European Programs, Except FP7 & H2020

9.3.2.1. AMIS

Program: CHIST-ERA

Project acronym: AMIS

Project title: Access Multilingual Information opinionS

Duration: Dec 2015- Nov 2019

Coordinator: Kamel Smaïli (LORIA)

Other partners: University of Avignon, University of Science and Technology Krakow, University of DEUSTO (Bilbao)

Participants: Dominique Fohr, Denis Jouvet, Odile Mella, Mohamed Amine Menacer

Abstract: The idea is to develop a multilingual system to help people understand broadcast news in a foreign language and compare them to corresponding news available in the user's mother tongue. MULTISPEECH contributions concern mainly the speech recognition in French, English and Arabic videos.

9.3.2.2. M-PHASIS

Program: ANR-DFG

Project acronym: M-PHASIS

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

Duration: March 2019 - Feb 2022

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

Partners: CREM (UL), LORIA (UL), JGUM (Johannes Gutenberg-Universität), SAAR (Saarland University)

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

Abstract: 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. MULTISPEECH contributions concern mainly the automatic detection of hate speech in social media.

9.4. International Initiatives

9.4.1. Inria International Partners

9.4.1.1. Informal International Partners

- Alessio Brutti & Maurizio Omologo, Fondazione Bruno Kessler (Italy)
  speech enhancement and speaker recognition [60]
- Samuele Cornell & Stefano Squartini, Università Politecnica delle Marche (Italy)
  speech enhancement and speaker recognition [59], [60]
- Tomi Kinnunen, University of Eastern Finland (Finland)
  speaker recognition & spoofing countermeasures [35], [12], [51], [54], [46].
- Justin Salamon, Adobe Research (USA)
  Sound event detection [48], [61]

- Junichi Yamagishi, National Institute of Informatics (Japan)
  speaker recognition & spoofing countermeasures [51], [46].

9.5. International Research Visitors

9.5.1. Visits to International Teams

9.5.1.1. Research Stays Abroad

- 2019 Sixth Frederick Jelinek Memorial Summer Workshop (Jun.–Aug. 2019, M. Pariente, S. Sivasankaran)

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events: Organisation

10.1.1.1. General Chair, Scientific Chair

Elected chair, Steering Committee of the Latent Variable Analysis and Signal Separation (LVA/ICA) conference series (E. Vincent, until Nov. 2019)

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, Barcelona, Spain, May 2020 (E. Vincent)

General co-chair, 1st Voice Privacy Challenge, Sept. 2020 (E. Vincent)

General co-chair, Detection and Classification of Acoustic Scenes and Events Challenge (R. Serizel, since Nov. 2018)

Area chair, 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (E. Vincent)

Area chair, 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (E. Vincent, A. Deleforge)

10.1.1.2. Member of the Organizing Committees

Steering Committee of the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge series (E. Vincent until Nov. 2019, R. Serizel from Nov. 2019)

Organizing Committee of the ASVspoof 2019 Challenge: Automatic Speaker Verification Spoofing And Countermeasures Challenge (M. Sahidullah)

Organizing Committee of Special Sessions on ASVspoof 2019 at INTERSPEECH 2019 and at IEEE ASRU 2019 (M. Sahidullah).

Organizing Committee of AVSP 2019 (S. Ouni)

Main organizer of IEEE Signal Processing Cup 2019 on Search & Rescue with Drone-Embedded Sound Source Localization (A. Deleforge).

10.1.2. Scientific Events Selection

10.1.2.1. Chair of Conference Program Committees

Review chair, IEEE Technical Committee on Audio and Acoustic Signal Processing, responsible for organizing the review of the 443 papers submitted to the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) in the general AASP domain (E. Vincent)

10.1.2.2. Member of the Conference Program Committees

ICALP 2019 - International Conference on Arabic Language Processing (D. Jouvet)

SIIE 2019 - Edition of the International Conference on Information Systems and Economic Intelligence (D. Fohr, I. Illina)

SPECOM 2019 - International Conference on Speech and Computer (D. Jouvet)

TSD 2019 - International Conference Text, Speech and Dialogue (D. Jouvet)

10.1.2.3. Reviewer

ASRU 2019 - IEEE Automatic Speech Recognition and Understanding Workshop (D. Jouvet, I. Illina, S. Ouni, M. Sahidullah, E. Vincent)

AVSP 2019 - International Conference on Auditory-Visual Speech Processing (S. Ouni)

DCASE 2019 - Workshop on Detection and Classification of Acoustic Scenes and Events (R. Serizel, E. Vincent)

ECML PKDD Joint International Workshop on Advances in Interpretable Machine Learning and Artificial Intelligence & eXplainable Knowledge Discovery in Data Mining (E. Vincent)

EUSIPCO 2019 - European Signal Processing Conference (M. Sahidullah)

ICALP 2019 - International Conference on Arabic Language Processing (D. Jouvet)

ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing (A. Bonneau, A. Deleforge, D. Jouvet, M. Sahidullah, R. Serizel, E. Vincent)

ICPhS 2019 - International Congress of Phonetic Sciences (A. Bonneau, Y. Laprie)

INTERSPEECH 2019 (A. Bonneau, I. Illina, D. Jouvet, S. Ouni, M. Sahidullah, E. Vincent)

IROS 2019 - International Conference on Intelligent Robots and Systems (A. Deleforge)

IVA 2019 - Intelligent Virtual Agents Conference (S. Ouni)

NeurIPS 2019 - Conference on Neural Information Processing Systems (A. Deleforge)

PaPE 2019 - Phonetics and Phonology in Europe conference (A. Bonneau)

SLATE 2019 - Workshop on Speech and Language Technology in Education (A. Bonneau, D. Jouvet)

SPECOM 2019 - International Conference on Speech and Computer (D. Jouvet)

TSD 2019 - International Conference Text, Speech and Dialogue (D. Jouvet)

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

Speech Communication (D. Jouvet)

Journal on Audio, Speech, and Music Processing (Y. Laprie)

Associate Editor of Circuits, Systems, and Signal Processing (M. Sahidullah)

Associate Editor of IET Signal Processing (M. Sahidullah)

Guest Editor of Computer Speech and Language, Special issue on Advances in Automatic Speaker Verification Anti-spoofing (M. Sahidullah)

Guest Editor of Journal on Audio, Speech and Music Processing, special issue on Advances in Audio Signal Processing for Robots and Drones (A. Deleforge)

10.1.3.2. Reviewer - Reviewing Activities

Computer Speech and Language (A. Deleforge, D. Jouvet, M. Sahidullah)

IEEE Access (A. Deleforge)

IEEE Signal Processing Letters (A. Deleforge, M. Sahidullah)

IEEE Transactions on Audio, Speech and Language Processing (M. Sahidullah, A. Deleforge)

IEEE Transactions on Biometrics, Behavior, and Identity Science (M. Sahidullah)

IEEE Transactions on Cognitive and Developmental Systems (A. Deleforge)

IEEE Transactions on Information Forensics and Security (M. Sahidullah)

IEEE Transactions on Signal Processing (A. Deleforge)

IET Biometrics (M. Sahidullah)

JASA Express Letter (Y. Laprie, S. Ouni)

Journal of the Acoustical Society of America (Y. Laprie)

Journal of Language, Speech and Hearing Research (Y. Laprie)

Journal on Audio, Speech, and Music Processing (A. Deleforge)

Sensors (A. Deleforge)

Speech Communication (D. Jouvet)

10.1.4. Invited Talks

Keynote "Grands défis scientifiques et technologiques en traitement de la parole: quelles initiatives chez Inria et au niveau européen?", Voice Tech Paris 2019 (E. Vincent) [18]

Keynote "Speech Processing and Prosody", 22nd International Conference of Text, Speech and Dialogue (TSD 2019) (D. Jouvet) [15]

"COMPRISE", META-FORUM 2019 (E. Vincent) [17]

Lecture "Taking the Best of Physics and Machine Learning in Robot Audition", IEEE/EURASIP/ISIF 2019 Summer School of Signal Processing, Arenzano, Italy (A. Deleforge)

Language pathology, Séminaire "Dépistage des troubles des apprentissages" in EHESP, Rennes, Jan. 2019 (A. Piquard-Kipffer)

"Dyslexie, dysorthographie : quels parcours scolaires ? Quelles rémédiations et adapatations ?" in Journée nationale des Dys, Bavilliers, Oct. 2019 (A. Piquard-Kipffer)

10.1.5. Leadership within the Scientific Community

Secretary/Treasurer, executive member of AVISA (Auditory-VIsual Speech Association), an ISCA Special Interest Group (S. Ouni)

Member of the board of AFCP - Association Francophone de la Communication Parlée (S. Ouni)

Elected members of the IEEE Technical committee on Audio and Acoustic Signal Processing (A. Deleforge, R. Serizel).

10.1.6. Scientific Expertise

Member of the Scientific Committee of an Institute for deaf people (INJS-Metz), A. Piquard-Kipffer

Member of an expertise Committee for specific language disabilities (MDPH 54), A. Piquard-Kipffer

10.1.7. Research Administration

Head of the AM2I Scientific Pole of Université de Lorraine (Y. Laprie)

Deputy Head of Science of Inria Nancy - Grand Est (E. Vincent)

Member of Management board of Université de Lorraine (Y. Laprie)

Member of the Comité Espace Transfert of Inria Nancy - Grand Est (E. Vincent)

Member of the national recruitment jury for Inria Junior Research Scientists (E. Vincent)

Member of the recruitment jury for Junior Research Scientists, Inria Paris (E. Vincent)

Member of a recruitment committee for Assistant Professor at Université Paris-Sud, May 2019 (D. Jouvet)

Member of the recruitment jury for an Associate Professor position, University of Lorraine (R. Serizel)

Member of the HCERES committee for Gipsa-Lab (S. Ouni)

Member of the National Council of Universities - CNU 27 (S. Ouni)

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 the Bureau de la Commission de Mention Informatique (I. Illina, S. Ouni)

Animator of the Commission Locale Développement Durable of Inria Nancy - Grand Est (A. Deleforge)

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

DUT: I. Illina, Java programming (56 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 (45 hours), L3, University of Lorraine, France

Licence: O. Mella, Introduction to Web Programming (22 hours), Digital tools (20 hours), L1, University of Lorraine, France

Licence: O. Mella, Computer Networking (72 hours), L2-L3, University of Lorraine, France

Licence: A. Piquard-Kipffer, Education Science (32 hours), L1, Département d'orthophonie, University of Lorraine, France

Licence: A. Piquard-Kipffer, Learning to Read (34 hours), L2, Département d'orthophonie, University of Lorraine, France

Licence: A. Piquard-Kipffer, Dyslexia, Dysorthographia (12 hours), L3, Département d'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 (10 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 (67 hours), Introduction to Speech Analysis and Recognition (12 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 , Psycholinguistics (20 hours), Departement Orthophonie, University Pierre et Marie Curie, Paris, France

Master: A. Piquard-Kipffer, French Language Didactics (53 hours), ESPE, INSPE University of Lorraine, France

Master: A. Piquard-Kipffer, Psychology (6 hours), M2, Departement of Psychology, University of Lorraine, France

Master: R. Serizel, Introduction to machine listening (3 hours), M2, University of Lorraine

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: O. Mella, DIU "Teaching computer science in high school" (7 hours), Computer science courses for secondary school teachers (ISN courses, 21 hours), ESPE, University of Lorraine, France

Continuous training: A. Piquard-Kipffer, Special Educational Needs (53 hours), ESPE, INSPE, University of Lorraine, France

Other: V. Colotte, Co-Responsible for NUMOC (Digital literacy by hybrid courses) for the University of Lorraine, France (for 7000 students)

Doctorat: A. Piquard-Kipffer , Language Pathology (20 hours), EHESP, University of Sorbonne, Paris, France

10.2.2. Supervision

PhD: Mathieu Fontaine, "Processus alpha-stable pour le traitement du signal", University of Lorraine, Jun. 12, 2019, A. Liutkus and R. Badeau (Télécom ParisTech) [1].

PhD: Lauréline Perotin, "Localisation et rehaussement de sources de parole au format Ambisonique", University of Lorraine, Oct. 31, 2019, R. Serizel, E. Vincent, and A. Guérin (Orange) [2].

PhD: Anastasiia Tsukanova, "Coarticulation modeling in articulatory synthesis", University of Lorraine, Dec. 13, 2019, Y. Laprie .

PhD in progress: Amal Houidhek, "Synthèse paramétrique de parole arabe", Dec. 2015, cotutelle, V. Colotte, D. Jouvet and Z. Mnasri (ENIT, Tunisia).

PhD in progress: Amine Menacer, "Traduction automatique de vidéos", May 2016, K. Smaïli (LORIA) and D. Jouvet.

PhD in progress: Nathan Libermann, "Deep learning for musical structure analysis and generation", Oct. 2016, F. Bimbot (IRISA) and E. Vincent.

PhD in progress: Théo Biasutto, "Multimodal coarticulation modeling: Towards the animation of an intelligible speaking head", Dec. 2016, S. Ouni.

PhD in progress: Sara Dahmani, "Modeling facial expressions to animate a realistic 3D virtual talking head", Jan. 2017, S. Ouni and V. Colotte.

PhD in progress: Guillaume Carbajal, “Apprentissage profond bout-en-bout pour le rehaussement de la parole”, Mar. 2017, R. Serizel, E. Vincent and É. Humbert (Invoxia).

PhD in progress: Sunit Sivasankaran, “Exploiting contextual information in the speech processing chain”, Jul. 2017, D. Fohr and E. Vincent.

PhD in progress: Ioannis Douros, “Combining cineMRI and static MRI to analyze speech production”, Jul. 2017, P.-A. Vuissoz (IADI) and Y. Laprie.

PhD in progress: Diego Di Carlo, “Estimating the Geometry of Audio Scenes Using Virtually-Supervised Learning”, Oct. 2017, A. Deleforge and N. Bertin (Inria Rennes).

PhD in progress: Lou Lee, “Du lexique au discours: les particules discursives en français”, 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”, Sept. 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, "Synthèse de parole expressive par apprentissage profond", 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: Shakeel Ahmad Sheikh, "Identifying disfluency in speakers with stuttering, and its rehabilitation, using DNN", Oct. 2019, S. Ouni.

PhD in progress: Sandipana Dowerah, “Robust speaker verification from far-field speech”, Oct. 2019, D. Jouvet and R. Serizel.

PhD in progress: Georgios Zervakis, “Integration of symbolic knowledge into deep learning”, Nov. 2019, M. Couceiro (LORIA) and E. Vincent.

PhD in progress: Nicolas Zampieri, "Automatic classification using deep learning of hate speech posted on the Internet", Nov. 2019, I. Illina and D. Fohr.

10.2.3. Juries

10.2.3.1. Participation in HDR and PhD juries

Participation in the HDR jury of Fabrice Hirsch (Sorbonne nouvelle University, Nov. 2019), Y. Laprie.

Participation in the HDR jury of Éric Bavu (CNAM, Dec. 2019), E. Vincent, president.

Participation in the PhD jury of Corentin Louboutin (Bretagne Loire University, Mar. 2019), E. Vincent, president.

Participation in the PhD jury of Neil Zeghidour (PSL Research University, Mar. 2019), E. Vincent, reviewer and president.

Participation in the PhD jury of Zied Elloumi (Communauté Université Grenoble Alpes, Mar. 2019), D. Jouvet, reviewer.

Participation in the PhD jury of Céline Jacques (Sorbonne University, Apr. 2019), E. Vincent.

Participation in the PhD jury of João Felipe Santos (University of Québec, Jul. 2019), E. Vincent, reviewer.

Participation in the PhD jury of Alice Cohen-Hadria (Sorbonne University, Oct. 2019), E. Vincent, reviewer.

Participation in the PhD jury of Lode Vuegen (KU Leuven, Oct. 2019), R. Serizel.

Participation in the PhD jury of Anne Bouvet (Grenoble Alpes University, Nov. 2019), Y. Laprie.

Participation in the Phd jury of Kevin Vythelingum (Le Mans Université, Dec. 2019), D. Jouvet, reviewer.

Participation in the PhD jury of Karima Abidi (Univertsité de Lorraine, Dec. 2019), D. Jouvet, president.

10.2.3.2. Participation in other juries

Participation in CAFIPEMPF Jury - Master Learning Facilitator, Académie de Nancy-Metz & Université de Lorraine, Apr. & May 2019, A. Piquard-Kipffer

Participation in the Competitive Entrance Examination into Speech-Language Pathology Department, University of Lorraine, Jun. 2019, A. Piquard-Kipffer.

10.3. Popularization

10.3.1. Internal or external Inria responsibilities

Member of the Commission Information et Edition Scientifique of Inria Nancy (A. Deleforge).

10.3.2. Articles and contents

Interview for “Un assistant vocal qui protège les données”, *France 3 Grand Est*, Jun. 7, 2019 (E. Vincent)

Interview for “La voix crescendo”, *L’Usine Nouvelle*, Jun. 27, 2019 (E. Vincent)

Interview for “On transmet beaucoup plus d’informations par la voix qu’un message”, *Rue89 Strasbourg*, Nov. 20, 2019 (E. Vincent)

Participation to the white paper “AI in the media and creative industries”, New European Media (NEM) consortium (E. Vincent) [52]

10.3.3. Education

5 interventions (3 hours) : "Dyslexic pupils in mainstream and special education". Training of trainers. ESPE de l’Académie de Nancy-Metz. Feb., Mar. & May 2019 (A. Piquard-Kipffer)

10.3.4. Interventions

6 interventions (1 hour) on robot audition and artificial intelligence research in classes from 5th (*CM2*) to 12th (*terminale*) grade around Nancy, Jan. 2019 (A. Deleforge)

Talk “Parole & deep learning: succès et grands défis”, Journée IA, Langage et Citoyens, LORIA, Mar. 2019; and also at Meetup IA Nancy, Jun. 2019 (E. Vincent)

Panel discussion on “Se positionner à l’Europe, une opportunité à saisir”, General assembly of Pôle Materalia, Nancy, Apr. 2019 (E. Vincent)

Demos and talk "Aider des enfants ayant des troubles du langage, quels métiers ? Quels outils ?" Forum des métiers, Collège Péguy, Le Chesnay, Apr. 2019 (A. Piquard-Kipffer)

4 interventions (1 hour) on robot audition and artificial intelligence research in high-school classes in Serbia, Sep. 2019 (A. Deleforge)

Demo “Apprendre aux robots à nous entendre” at the “Nuit des Chercheurs” of Belgrade, Serbia in Sept. 2019 and of Nancy for the 80th anniversary of CNRS in Oct. 2019 (A. Deleforge)

Demos “Apprendre aux robots à nous entendre” and “Assistants vocaux et vie privée”, Fête de la Science, Université de Lorraine, Oct. 2019 (A. Deleforge, I. Illina, M. Sahidullah, B. M. L. Srivastava, E. Vincent)

Panel discussion “Tous connectés et après : les enjeux des applications d’interactions vocales”, Shadok, Strasbourg, Nov. 2019 (E. Vincent)

10.3.5. Internal action

“H2020 COMPRISE”, Internal “Café’In” event, Inria Nancy, Jun. 2019 (E. Vincent & Z. Chelly-Dagdia)

10.3.6. Creation of media or tools for science outreach

Video “Exposed by your own voice”, https://www.youtube.com/watch?v=gm2cC8za8Us.

Video ‘When voice assistants don’t understand”, https://www.youtube.com/watch?v=-HvADcfEOuE.

Video “Why is voice assistant integration so expensive”, https://www.youtube.com/watch?v=5LQb9X3RtUs

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] M. Fontaine. *Alpha-stable processes for signal processing*, Université de Lorraine, June 2019, https://tel.archives-ouvertes.fr/tel-02188304

[2] L. Perotin. *Localization and enhancement of speech from the Ambisonics format : analyse de scènes sonores pour faciliter la commande vocale*, Université de Lorraine, October 2019, https://hal.univ-lorraine.fr/tel-02393258

[3] A. Tsukanova. *Articulatory speech synthesis*, Univeristé de lorraine, December 2019, https://hal.archives-ouvertes.fr/tel-02433528

Articles in International Peer-Reviewed Journals

[4] N. Bertin, E. Camberlein, R. Lebarbenchon, E. Vincent, S. Sivasankaran, I. Illina, F. Bimbot. *VoiceHome-2, an extended corpus for multichannel speech processing in real homes*, in "Speech Communication", January 2019, vol. 106, pp. 68-78 [*DOI :* 10.1016/j.specom.2018.11.002], https://hal.inria.fr/hal-01923108

[5] A. Deleforge, D. Di Carlo, M. Strauss, R. Serizel, L. Marcenaro. *Audio-Based Search and Rescue with a Drone: Highlights from the IEEE Signal Processing Cup 2019 Student Competition*, in "IEEE Signal Processing Magazine", September 2019, vol. 36, n^o^ 5, pp. 138-144, https://arxiv.org/abs/1907.04655 [*DOI :* 10.1109/MSP.2019.2924687], https://hal.archives-ouvertes.fr/hal-02161897

[6] K. Déguernel, E. Vincent, J. Nika, G. Assayag, K. Smaïli. *Learning of Hierarchical Temporal Structures for Guided Improvisation*, in "Computer Music Journal", 2019, vol. 43, n^o^ 2, https://hal.inria.fr/hal-02378273

[7] A. MESAROS, A. DIMENT, B. ELIZALDE, T. HEITTOLA, E. VINCENT, B. RAJ, T. VIRTANEN. *Sound event detection in the DCASE 2017 Challenge*, in "IEEE/ACM Transactions on Audio, Speech and Language Processing", June 2019, vol. 27, n^{o} 6, pp. 992-1006 [*DOI :* 10.1109/TASLP.2019.2907016], https://hal.inria.fr/hal-02067935

[8] Q. V. NGUYEN, F. COLAS, E. VINCENT, F. CHARPILLET. *Motion planning for robot audition*, in "Autonomous Robots", December 2019, vol. 43, n^{o} 8, pp. 2293-2317 [*DOI :* 10.1007/S10514-019-09880-1], https://hal.inria.fr/hal-02188342

[9] L. PEROTIN, R. SERIZEL, E. VINCENT, A. GUÉRIN. *CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings*, in "IEEE Journal of Selected Topics in Signal Processing", February 2019, vol. 13, n^{o} 1, pp. 22-33 [*DOI :* 10.1109/JSTSP.2019.2900164], https://hal.inria.fr/hal-01839883

[10] A. PODDAR, M. SAHIDULLAH, G. SAHA. *Quality Measures for Speaker Verification with Short Utterances*, in "Digital Signal Processing", January 2019, vol. 88, pp. 66-79 [*DOI :* 10.1016/J.DSP.2019.01.023], https://hal.inria.fr/hal-01998376

[11] K. SMAÏLI, D. FOHR, C.-E. GONZÁLEZ-GALLARDO, M. L. GREGA, L. JANOWSKI, D. JOUVET, A. KOŹBIAŁ, D. LANGLOIS, M. LESZCZUK, O. MELLA, M.-A. MENACER, A. MENDEZ, E. L. L. PONTES, E. SANJUAN, J.-M. TORRES-MORENO, B. GARCIA-ZAPIRAIN. *Summarizing videos into a target language: Methodology, architectures and evaluation*, in "Journal of Intelligent and Fuzzy Systems", July 2019, vol. 1, pp. 1-12 [*DOI :* 10.3233/JIFS-179350], https://hal.archives-ouvertes.fr/hal-02271287

[12] V. VESTMAN, T. KINNUNEN, R. G. HAUTAMÄKI, M. SAHIDULLAH. *Voice Mimicry Attacks Assisted by Automatic Speaker Verification*, in "Computer Speech and Language", June 2019, vol. 59, pp. 36-54 [*DOI :* 10.1016/J.CSL.2019.05.005], https://hal.archives-ouvertes.fr/hal-02161773

Invited Conferences

[13] C. DODANE, D. BOUTET, F. HIRSCH, S. OUNI, A. MORGENSTERN. *MODALISA une plateforme intégrative pour capturer l'orchestration des gestes et de la parole*, in "Défi Instrumentation aux Limites, Colloque de restitution", Paris, France, CNRS, September 2019, https://hal.archives-ouvertes.fr/hal-02375011

[14] F. FORBES, A. DELEFORGE, R. HORAUD, E. PERTHAME. *Robust non-linear regression approach for generalized inverse problems in a high dimensional setting*, in "AIP 2019 - Applied Inverse Problem conference", Grenoble, France, July 2019, https://hal.archives-ouvertes.fr/hal-02415115

[15] D. JOUVET. *Speech Processing and Prosody*, in "TSD 2019 - 22nd International Conference of Text, Speech and Dialogue", Ljubljana, Slovenia, September 2019, https://hal.inria.fr/hal-02177210

[16] R. SERIZEL, N. TURPAULT. *Sound Event Detection from Partially Annotated Data: Trends and Challenges*, in "IcETRAN conference", Srebrno Jezero, Serbia, June 2019, https://hal.inria.fr/hal-02114652

[17] E. VINCENT. *COMPRISE*, in "META-FORUM", Bruxelles, Belgium, October 2019, https://hal.inria.fr/hal-02377051

[18] E. VINCENT. *Grands défis scientifiques et technologiques en traitement de la parole: quelles initiatives chez Inria et au niveau européen?*, in "Voice Tech Paris 2019", Paris, France, November 2019, https://hal.inria.fr/hal-02377036

[19] E. VINCENT. *Parole & deep learning : succès et grands défis*, in "Journée IA, Langage et Citoyens", Nancy, France, March 2019, https://hal.inria.fr/hal-02090623

International Conferences with Proceedings

[20] K. ABIDI, D. FOHR, D. JOUVET, D. LANGLOIS, O. MELLA, K. SMAÏLI. *A Fine-grained Multilingual Analysis Based on the Appraisal Theory: Application to Arabic and English Videos*, in "ICALP: International Conference on Arabic Language Processing", Nancy, France, Springer, August 2019, vol. Communications in Computer and Information Science book series (CCIS, volume 1108), pp. 49-61 [*DOI :* 10.1007/978-3-030-32959-4_4], https://hal.archives-ouvertes.fr/hal-02314244

[21] T. BIASUTTO–LERVAT, S. DAHMANI, S. OUNI. *Modeling Labial Coarticulation with Bidirectional Gated Recurrent Networks and Transfer Learning*, in "INTERSPEECH 2019 - 20th Annual Conference of the International Speech Communication Association", Graz, Austria, September 2019, https://hal.inria.fr/hal-02175780

[22] A. BONNEAU. *German obstruent sequences by French L2 learners*, in "ICPhS 2019 - International Congress of Phonetic Sciences", Melbourne, Australia, August 2019, https://hal.inria.fr/hal-02143360

[23] S. DAHMANI, V. COLOTTE, V. GIRARD, 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, September 2019, https://hal.inria.fr/hal-02175776

[24] D. DI CARLO, A. DELEFORGE, N. BERTIN. *Mirage: 2D Source Localization Using Microphone Pair Augmentation with Echoes*, in "ICASSP 2019 - IEEE International Conference on Acoustic, Speech Signal Processing", Brighton, United Kingdom, IEEE, May 2019, pp. 775-779, https://arxiv.org/abs/1906.08968 [*DOI :* 10.1109/ICASSP.2019.8683534], https://hal.archives-ouvertes.fr/hal-02160940

[25] C. DODANE, D. BOUTET, I. DIDIRKOVA, F. HIRSCH, S. OUNI, A. MORGENSTERN. *An integrative platform to capture the orchestration of gesture and speech*, in "GeSpIn 2019 - Gesture and Speech in Interaction", Paderborn, Germany, September 2019, https://hal.inria.fr/hal-02278345

[26] I. K. DOUROS, J. FELBLINGER, J. FRAHM, K. ISAIEVA, A. JOSEPH, Y. LAPRIE, F. ODILLE, A. TSUKANOVA, D. VOIT, P.-A. VUISSOZ. *A Multimodal Real-Time MRI Articulatory Corpus of French for Speech Research*, in "INTERSPEECH 2019 - 20th Annual Conference of the International Speech Communication Association", Graz, Austria, September 2019, https://hal.inria.fr/hal-02167756

[27] I. K. DOUROS, Y. LAPRIE, P.-A. VUISSOZ, B. ELIE. *Acoustic Evaluation of Simplifying Hypotheses Used in Articulatory Synthesis*, in "ICA 2019 - 23rd International Congress on Acoustics", Aachen, Germany, September 2019, https://hal.inria.fr/hal-02180617

[28] I. K. DOUROS, A. TSUKANOVA, K. ISAIEVA, P.-A. VUISSOZ, Y. LAPRIE. *Towards a method of dynamic vocal tract shapes generation by combining static 3D and dynamic 2D MRI speech data*, in "INTERSPEECH 2019 - 20th Annual Conference of the International Speech Communication Association", Graz, Austria, September 2019, https://hal.inria.fr/hal-02181333

[29] I. K. DOUROS, P.-A. VUISSOZ, Y. LAPRIE. *Acoustic impacts of geometric approximation at the level of velum and epiglottis on French vowels*, in "ICPhS 2019 - International Congress of Phonetic Sciences", Melbourne, Australia, August 2019, https://hal.inria.fr/hal-02180566

[30] I. K. DOUROS, P.-A. VUISSOZ, Y. LAPRIE. *Comparison between 2D and 3D models for speech production: a study of French vowels*, in "ICPhS 2019 - International Congress of Phonetic Sciences", Melbourne, Australia, August 2019, https://hal.inria.fr/hal-02180606

[31] I. K. DOUROS, P.-A. VUISSOZ, Y. LAPRIE. *Effect of head posture on phonation of French vowels*, in "ICPhS 2019 - Proceedings of International Congress of Phonetic Sciences", Melbourne, Australia, August 2019, https://hal.inria.fr/hal-02180486

[32] A. DUFRAUX, E. VINCENT, A. HANNUN, A. BRUN, M. DOUZE. *Lead2Gold: Towards exploiting the full potential of noisy transcriptions for speech recognition*, in "ASRU 2019 - IEEE Automatic Speech Recognition and Understanding Workshop", Singapour, Singapore, December 2019, https://hal.inria.fr/hal-02316572

[33] B. ELIE, A. AMELOT, Y. LAPRIE, S. MAEDA. *Glottal Opening Measurements in VCV and VCCV Sequences*, in "ICA 2019 - 23rd International Congress on Acoustics", Aachen, Germany, September 2019, https://hal.inria.fr/hal-02180626

[34] M. FONTAINE, A. A. NUGRAHA, R. BADEAU, K. YOSHII, A. LIUTKUS. *Cauchy Multichannel Speech Enhancement with a Deep Speech Prior*, in "EUSIPCO 2019 - 27th European Signal Processing Conference", Coruña, Spain, September 2019, https://hal.telecom-paristech.fr/hal-02288063

[35] T. KINNUNEN, R. G. HAUTAMÄKI, V. VESTMAN, M. SAHIDULLAH. *Can We Use Speaker Recognition Technology to Attack Itself? Enhancing Mimicry Attacks Using Automatic Target Speaker Selection*, in "ICASSP 2019 – 44th International Conference on Acoustics, Speech, and Signal Processing", Brighton, United Kingdom, May 2019, https://hal.inria.fr/hal-02051701

[36] A. KULKARNI, V. COLOTTE, D. JOUVET. *Layer adaptation for transfer of expressivity in speech synthesis*, in "LTC'19 - 9th Language & Technology Conference", Poznan, Poland, May 2019, https://hal.inria.fr/hal-02177945

[37] L. LEE, K. BARTKOVA, D. JOUVET, M. DARGNAT, Y. KEROMNES. *Can prosody meet pragmatics? Case of discourse particles in French*, in "ICPhS 2019 - International Congress of Phonetic Sciences", Melbourne, Australia, August 2019, https://hal.inria.fr/hal-02177202

[38] K. A. LEE, V. HAUTAMÄKI, T. KINNUNEN, H. YAMAMOTO, K. OKABE, V. VESTMAN, J. HUANG, G. DING, H. SUN, A. LARCHER, R. K. DAS, H. LI, M. ROUVIER, P.-M. B. BOUSQUET, W. RAO, Q. WANG, C. ZHANG, F. BAHMANINEZHAD, H. DELGADO, J. PATINO, Q. WANG, L. GUO, T. KOSHINAKA, J. ZHANG, K. SHINODA, T. NGO TRONG, M. SAHIDULLAH, F. LU, Y. TANG, M. TU, K. KUAN TEH, H. DAT TRAN, K. K. GEORGE, I. KUKANOV, F. DESNOUS, J. YANG, E. YILMAZ, L. XU, J.-F. BONASTRE, C. XU, Z. H. LIM, S. CHNG, S. RANJAN, J. H. L. HANSEN, M. TODISCO, N. EVANS. *I4U Submission to NIST SRE 2018: Leveraging from a Decade of Shared Experiences*, in "INTERSPEECH 2019 - 20th Annual Conference of the International Speech Communication Association", Graz, Austria, September 2019, https://hal.archives-ouvertes.fr/hal-02280151

[39] T. LÉONOVA, G. COFFE, A. TARASCONI, A. PIQUARD-KIPFFER, D. SARDIN, A. GOSSE, J. BORÉ. *L'impact du trouble du spectre de l'autisme sur le bien-être psychologique des parents*, in "XVIIIème

Congrès de l'Association Internationale de Formation et de Recherche en Éducation Familiale", Schoelcher, Martinique, France, May 2019, https://hal.inria.fr/hal-02179616

[40] M. A. Menacer, C. E. González-Gallardo, K. Abidi, D. Fohr, D. Jouvet, D. Langlois, O. Mella, F. Sadat, J. M. Torres-Moreno, K. Smaïli. *Extractive Text-Based Summarization of Arabic videos: Issues, Approaches and Evaluations*, in "ICALP: International Conference on Arabic Language Processing", Nancy, France, Springer, August 2019, vol. Communications in Computer and Information Science book series (CCIS, volume 1108), pp. 65-78 [*DOI :* 10.1007/978-3-030-32959-4_5], https://hal.archives-ouvertes.fr/hal-02314238

[41] M. Menacer, D. Langlois, D. Jouvet, D. Fohr, O. Mella, K. Smaïli. *Machine Translation on a parallel Code-Switched Corpus*, in "Canadian AI 2019 - 32nd Conference on Canadian Artificial Intelligence", Ontario, Canada, Lecture Notes in Artificial Intelligence, May 2019, https://hal.archives-ouvertes.fr/hal-02106010

[42] M. Pariente, A. Deleforge, E. Vincent. *A Statistically Principled and Computationally Efficient Approach to Speech Enhancement using Variational Autoencoders*, in "INTERSPEECH", Graz, Austria, September 2019, https://arxiv.org/abs/1905.01209 , https://hal.inria.fr/hal-02116165

[43] *Best Paper*
L. Perotin, A. Défossez, E. Vincent, R. Serizel, A. Guérin. *Regression versus classification for neural network based audio source localization*, in "WASPAA 2019 - IEEE Workshop on Applications of Signal Processing to Audio and Acoustics", New Paltz, United States, IEEE, October 2019, https://hal.inria.fr/hal-02125985.

[44] D. Ribas, E. Vincent. *An improved uncertainty propagation method for robust i-vector based speaker recognition*, in "ICASSP 2019 - 44th International Conference on Acoustics, Speech, and Signal Processing", Brighton, United Kingdom, May 2019, https://arxiv.org/abs/1902.05761 , https://hal.inria.fr/hal-02010199

[45] B. M. L. Srivastava, A. Bellet, M. Tommasi, E. Vincent. *Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion?*, in "INTERSPEECH 2019 - 20th Annual Conference of the International Speech Communication Association", Graz, Austria, September 2019, https://hal.inria.fr/hal-02166434

[46] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, K. A. Lee. *ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection*, in "INTERSPEECH 2019 - 20th Annual Conference of the International Speech Communication Association", Graz, Austria, September 2019, https://hal.archives-ouvertes.fr/hal-02172099

[47] A. Tsukanova, I. K. Douros, A. Shimorina, Y. Laprie. *Can static vocal tract positions represent articulatory targets in continuous speech? Matching static MRI captures against real-time MRI for the French language*, in "ICPhS 2019 - International Congress of Phonetic Sciences", Melbourne, Australia, August 2019, https://hal.inria.fr/hal-02181314

[48] N. Turpault, R. Serizel, A. Parag Shah, J. Salamon. *Sound event detection in domestic environments with weakly labeled data and soundscape synthesis*, in "Workshop on Detection and Classification of Acoustic Scenes and Events", New York City, United States, October 2019, https://hal.inria.fr/hal-02160855

[49] N. Turpault, R. Serizel, E. Vincent. *Semi-supervised triplet loss based learning of ambient audio embeddings*, in "ICASSP", Brighton, United Kingdom, 2019, https://hal.archives-ouvertes.fr/hal-02025824

[50] I. Zangar, Z. Mnasri, V. Colotte, D. Jouvet. *F0 modeling using DNN for Arabic parametric speech synthesis*, in "INNSBDDL 2019 - INNS Big Data and Deep Learning", Sestri Levante, Italy, April 2019, https://hal.inria.fr/hal-02177496

Scientific Books (or Scientific Book chapters)

[51] M. Sahidullah, H. Delgado, M. Todisco, T. Kinnunen, N. Evans, J. Yamagishi, K. A. Lee. *Introduction to Voice Presentation Attack Detection and Recent Advances*, in "Handbook of Biometric Anti-Spoofing: Presentation Attack Detection", S. Marcel, M. S. Nixon, J. Fierrez, N. Evans (editors), Advances in Computer Vision and Pattern Recognition, Springer, 2019, pp. 321-361 [*DOI :* 10.1007/978-3-319-92627-8_15], https://hal.inria.fr/hal-01974528

Research Reports

[52] B. Caramiaux, F. Lotte, J. Geurts, G. Amato, M. Behrmann, F. Bimbot, F. Falchi, A. Garcia, J. Gibert, G. Gravier, H. Holken, H. Koenitz, S. Lefebvre, A. Liutkus, A. Perkis, R. Redondo, E. Turrin, T. Viéville, E. Vincent. *AI in the media and creative industries*, New European Media (NEM), April 2019, pp. 1-35, https://arxiv.org/abs/1905.04175 , https://hal.inria.fr/hal-02125504

[53] G. Carbajal, R. Serizel, E. Vincent, E. Humbert. *Joint DNN-Based Multichannel Reduction of Acoustic Echo, Reverberation and Noise: Supporting Document*, Inria Nancy, équipe Multispeech ; Invoxia SAS, November 2019, n^{o} RR-9303, https://hal.inria.fr/hal-02372431

[54] K. A. Lee, V. Hautamäki, T. Kinnunen, H. Yamamoto, K. Okabe, V. Vestman, J. Huang, G. Ding, H. Sun, A. Larcher, R. K. Das, H. Li, M. Rouvier, P.-M. B. Bousquet, W. Rao, Q. Wang, C. Zhang, F. Bahmaninezhad, H. Delgado, J. Patino, Q. Wang, L. Guo, T. Koshinaka, J. Zhang, K. Shinoda, T. Ngo Trong, M. Sahidullah, F. Lu, Y. Tang, M. Tu, K. Kuan Teh, H. Dat Tran, K. K. George, I. Kukanov, F. Desnous, J. Yang, E. Yilmaz, L. Xu, J.-F. Bonastre, C. Xu, Z. H. Lim, S. Chng, S. Ranjan, J. H. L. Hansen, M. Todisco, N. Evans. *I4U Submission to NIST SRE 2018: Leveraging from a Decade of Shared Experiences*, I4U Consortium, April 2019, https://hal.archives-ouvertes.fr/hal-02174317

[55] M. Pariente, A. Deleforge, E. Vincent. *A Statistically Principled and Computationally Efficient Approach to Speech Enhancement using Variational Autoencoders : Supporting Document*, Inria, April 2019, n^{o} RR-9268, pp. 1-8, https://hal.inria.fr/hal-02089062

Software

[56] M. Kowalski, E. Vincent, R. Gribonval. *Underdetermined Reverberant Source Separation*, October 2019
[SWH-ID : swh:1:dir:ec4ae097465d9ea51589537ea94b2ea50e8d134d], Software, https://hal.archives-ouvertes.fr/hal-02309043

Other Publications

[57] G. Carbajal, R. Serizel, E. Vincent, E. Humbert. *Joint DNN-Based Multichannel Reduction of Acoustic Echo, Reverberation and Noise*, December 2019, working paper or preprint, https://hal.inria.fr/hal-02372579

[58] N. Furnon, R. Serizel, I. Illina, S. Essid. *DNN-Based Distributed Multichannel Mask Estimation for Speech Enhancement in Microphone Arrays*, October 2019, Submitted to ICASSP2020, https://hal.archives-ouvertes.fr/hal-02389159

[59] M. Pariente, S. Cornell, A. Deleforge, E. Vincent. *Filterbank design for end-to-end speech separation*, October 2019, Submitted to ICASSP2020, https://hal.archives-ouvertes.fr/hal-02355623

[60] M. Sahidullah, J. Patino, S. Cornell, R. Yin, S. Sivasankaran, H. Bredin, P. Korshunov, A. Brutti, R. Serizel, E. Vincent, N. Evans, S. Marcel, S. Squartini, C. Barras. *The Speed Submission to DIHARD II: Contributions & Lessons Learned*, November 2019, working paper or preprint, https://hal.inria.fr/hal-02352840

[61] R. Serizel, N. Turpault, A. Shah, J. Salamon. *Sound event detection in synthetic domestic environments*, November 2019, working paper or preprint, https://hal.inria.fr/hal-02355573

[62] S. Sivasankaran, E. Vincent, D. Fohr. *Analyzing the impact of speaker localization errors on speech separation for automatic speech recognition*, November 2019, Submitted to ICASSP 2020, https://hal.inria.fr/hal-02355669

[63] S. Sivasankaran, E. Vincent, D. Fohr. *SLOGD: Speaker Location Guided Deflation Approach to Speech Separation*, November 2019, Submitted to ICASSP 2020, https://hal.inria.fr/hal-02355613

[64] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, E. Vincent. *Evaluating Voice Conversion-based Privacy Protection against Informed Attackers*, November 2019, working paper or preprint, https://hal.inria.fr/hal-02355115