IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2017

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:

A3.1.4. - Uncertain data
A3.4.6. - Neural networks
A3.4.8. - Deep learning
A5.1.7. - Multimodal interfaces
A5.7. - Audio modeling and processing
A5.7.1. - Sound
A5.7.2. - Music
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

Other Research Topics and Application Domains:

B4.3.3. - Wind energy
B8.1.2. - Sensor networks for smart buildings
B8.4. - Security and personal assistance
B9.1.1. - E-learning, MOOC
B9.2.1. - Music, sound
B9.2.2. - Cinema, Television
B9.4.1. - Computer science
B9.4.2. - Mathematics
B9.4.5. - Data science
B9.5.8. - Linguistics
B9.5.10. - Digital humanities

1. Personnel

Research Scientists

Anne Bonneau [CNRS, Researcher]
Dominique Fohr [CNRS, Researcher]
Denis Jouvet [Team leader, Inria, Senior Researcher, HDR]
Yves Laprie [CNRS, Senior Researcher, HDR]
Antoine Liutkus [Inria, Researcher, until Sep 2017]
Emmanuel Vincent [Inria, Senior Researcher, HDR]

Faculty Members

Vincent Colotte [Univ de Lorraine, Associate Professor]
Irène Illina [Univ de Lorraine, Associate Professor, HDR]
Odile Mella [Univ de Lorraine, Associate Professor]
Slim Ouni [Univ de Lorraine, Associate Professor, HDR]
Agnès Piquard-Kipffer [Univ de Lorraine, École Supérieure du Professorat et de l'Éducation, Associate Professor]
Romain Serizel [Univ de Lorraine, Associate Professor]

Post-Doctoral Fellow

Benjamin Elie [CNRS, until Aug 2017]

PhD Students

Theo Biasutto-Lervat [Univ de Lorraine]
Guillaume Carbajal [Invoxia, from Mar 2017]
Sara Dahmani [Univ de Lorraine]
Ken Deguernel [Inria]
Ioannis Douros [Univ de Lorraine, from Jun 2017]
Mathieu Fontaine [Inria]
Amal Houidhek [École Nationale d'Ingénieurs de Tunis, Tunisia]
Nathan Libermann [Inria Rennes, Team PANAMA, until Aug 2017]
Yang Liu [Univ de Lorraine, until Feb 2017]
Van Quan Nguyen [Inria, Team LARSEN, until Oct 2017]
Aditya Nugraha [Inria]
Laureline Perotin [Orange Labs, from Nov 2016]
Imran Sheikh [Univ de Lorraine, until Mar 2017]
Sunit Sivasankaran [Inria, from Jul 2017]
Anastasiia Tsukanova [Univ de Lorraine]

Technical staff

Ismaël Bada [CNRS, until Sep 2017]
Yassine Boudi [Inria, from Jul 2017]
Valérian Girard [Univ de Lorraine, from Feb 2017]
Mathieu Hu [Inria, from Jul 2017]
Karan Nathwani [Inria, until Aug 2017]

Interns

Aman Zaid Berhe [Univ de Lorraine, from Feb 2017 until Jul 2017]
Stephanie Caillavet [Univ de Lorraine, from Jun 2017 until Jul 2017]
Kévin Champeroux [Univ de Lorraine, from Mar 2017 until Jun 2017]
Jérôme Cousinou [Univ de Lorraine, from Mar 2017 until Jun 2017]
Remi Decelle [Inria, from Apr 2017 until Sep 2017]
Diego Di Carlo [Inria, until Apr 2017]
Ismail El Mastafi [Inria, from Jun 2017 until Sep 2017]
Floris Fournier [Inria, from May 2017 until Jul 2017]
Amélie Greiner [Univ de Lorraine, from Apr 2017 until Sep 2017]
Juan Karsten [CNRS, from Mar 2017 until Jul 2017]
Gautier Loveiko [CNRS, from Apr 2017 until Aug 2017]

Romain Marlier [Inria, from Apr 2017 until Jun 2017]
Yohan Robert [CNRS, from May 2017 until Jun 2017]

Administrative Assistants

Hélène Cavallini [Inria]
Antoinette Courrier [CNRS, until Aug 2017]
Delphine Hubert [Univ de Lorraine, from Sep 2017]
Martine Kuhlmann [CNRS, from Sep 2017]
Sylvie Musilli [Univ de Lorraine, until Aug 2017]

Visiting Scientists

Md Sahidullah [University of Eastern Finland, from Aug 2017 until Oct 2017]
Venkata Vishnu Vardan Varanasi [Indian Institute of Sciences, Kanpur, from Feb 2017 until Aug 2017]
Ziteng Wang [Institute of Acoustics, Chinese Academy of Sciences, until Sep 2017]

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.

The project is organized along the three following scientific challenges:

- **The explicit modeling of speech.** Speech signals result from the movements of articulators. A good knowledge of their position with respect to sounds is essential to improve, on the one hand, articulatory speech synthesis, and on the other hand, the relevance of the diagnosis and of the associated feedback in computer assisted language learning. Production and perception processes are interrelated, so a better understanding of how humans perceive speech will lead to more relevant diagnoses in language learning as well as pointing out critical parameters for expressive speech synthesis. Also, as the expressivity translates into both visual and acoustic effects that must be considered simultaneously, the multimodal components of expressivity, which are both on the voice and on the face, will be addressed to produce expressive multimodal speech.
- **The statistical modeling of speech.** Statistical approaches are common for processing speech and they achieve performance that makes possible their use in actual applications. However, speech recognition systems still have limited capabilities (for example, even if large, the vocabulary is limited) and their performance drops significantly when dealing with degraded speech, such as noisy signals, distant microphone recording and spontaneous speech. Source separation based approaches are investigated as a way of making speech recognition systems more robust to noise. Handling new proper names is an example of critical aspect that is tackled, along with the use of statistical models for speech-text automatic alignment and for speech production.

- **The estimation and the exploitation of uncertainty in speech processing.** Speech signals are highly variable and often disturbed with noise or other spurious signals (such as music or undesired extra speech). In addition, the output of speech enhancement and of source separation techniques is not exactly the accurate "clean" original signal, and estimation errors have to be taken into account in further processing. This is the goal of computing and handling the uncertainty of the reconstructed signal provided by source separation approaches. Finally, MULTISPEECH also aims at estimating the reliability of phonetic segment boundaries and prosodic parameters for which no such information is yet available.

Although being interdependent, each of these three scientific challenges constitutes a founding research direction for the MULTISPEECH project. Consequently, the research program is organized along three research directions, each one matching a scientific challenge. A large part of the research is conducted on French speech data; English and German languages are also considered in speech recognition experiments and language learning. Adaptation to other languages of the machine learning based approaches is possible, depending on the availability of corresponding speech corpora.

3. Research Program

3.1. Explicit Modeling of Speech Production and Perception

Speech signals are the consequence of the deformation of the vocal tract under the effect of the movements of the articulators (jaw, lips, tongue, ...) to modulate the excitation signal produced by the vocal cords or air turbulence. These deformations are visible on the face (lips, cheeks, jaw) through the coordination of different orofacial muscles and skin deformation induced by the latter. These deformations may also express different emotions. We should note that human speech expresses more than just phonetic content, to be able to communicate effectively. In this project, we address the different aspects related to speech production from the modeling of the vocal tract up to the production of expressive audiovisual speech. Phonetic contrasts used by the phonological system of any language result from constraints imposed by the nature of the human speech production apparatus. For a given language these contrasts are organized so as to guarantee that human listeners can identify (categorize) sounds robustly. The study of the categorization of sounds and prosody thus provides a complementary view on speech signals by focusing on the discrimination of sounds by humans, particularly in the context of language learning.

3.1.1. Articulatory modeling

Modeling speech production is a major issue in speech sciences. Acoustic simulation makes the link between articulatory and acoustic domains. Unfortunately this link cannot be fully exploited because there is almost always an acoustic mismatch between natural and synthetic speech generated with an articulatory model approximating the vocal tract. However, the respective effects of the geometric approximation, of the fact of neglecting some cavities in the simulation, of the imprecision of some physical constants and of the dimensionality of the acoustic simulation are still unknown. Hence, the first objective is to investigate the origin of the acoustic mismatch by designing more precise articulatory models, developing new methods to acquire tridimensional Magnetic Resonance Imaging (MRI) data of the entire vocal tract together with denoised speech signals, and evaluating several approaches of acoustic simulation. The articulatory data acquisition relies on a head-neck antenna at Nancy Hospital to acquire MRI of the vocal tract, and on the articulograph Carstens AG501 available in the laboratory.

Up to now, acoustic-to-articulatory inversion has been addressed as an instantaneous problem, articulatory gestures being recovered by concatenating local solutions. The second objective is thus to investigate how more elaborated strategies (a syllabus of primitive gestures, articulatory targets...) can be incorporated in the acoustic-to-articulatory inversion algorithms to take into account dynamic aspects.

3.1.2. Expressive acoustic-visual synthesis

Speech is considered as a bimodal communication means; the first modality is audio, provided by acoustic speech signals and the second one is visual, provided by the face of the speaker. In our approach, the Acoustic-Visual Text-To-Speech synthesis (AV-TTS) is performed simultaneously with respect to its acoustic and visible components, by considering a bimodal signal comprising both acoustic and visual channels. A first AV-TTS system has been developed resulting in a talking head; the system relied on 3D-visual data and on an extension of our acoustic-unit concatenation text-to-speech synthesis system (SOJA). An important goal is to provide an audiovisual synthesis that is intelligible, both acoustically and visually. Thus, we continue working on adding visible components of the head through a tongue model and a lip model. We will also improve the TTS engine to increase the accuracy of the unit selection simultaneously into the acoustic and visual domains. To acquire the facial data, we consider using a marker-less motion capture system using a kinect-like system with a face tracking software, which constitutes a relatively low-cost alternative to the Vicon system.

Another challenging research goal is to add expressivity in the AV-TTS. The expressivity comes through the acoustic signal (prosody aspects) and also through head and eyebrow movements. One objective is to add a prosodic component in the TTS engine in order to take into account some prosodic entities such as emphasis (to highlight some important key words). One intended approach will be to explore an expressivity measure at sound, syllable and/or sentence levels that describes the degree of perception or realization of an expression/emotion (audio and 3D domain). Such measures will be used as criteria in the selection process of the synthesis system. To tackle the expressivity issue we will also investigate Hidden Markov Model (HMM) based synthesis which allows for easy adaptation of the system to available data and to various conditions.

3.1.3. Categorization of sounds and prosody for native and non-native speech

Discriminating speech sounds and prosodic patterns is the keystone of language learning whether in the mother tongue or in a second language. This issue is associated with the emergence of phonetic categories, i.e., classes of sounds related to phonemes and prosodic patterns. The study of categorization is concerned not only with acoustic modeling but also with speech perception and phonology. Foreign language learning raises the issue of categorizing phonemes of the second language given the phonetic categories of the mother tongue. Thus, studies on the emergence of new categories, whether in the mother tongue (for people with language deficiencies) or in a second language, must rely upon studies on native and non-native acoustic realizations of speech sounds and prosody, and on perceptual experiments. Concerning prosody, studies are focused on native and non-native realizations of modalities (e.g., question, affirmation, command, ...), as well as non-native realizations of lexical accents and focus (emphasis).

For language learning, the analysis of the prosody and of the acoustic realization of the sounds aims at providing automatic feedback to language learners with respect to acquisition of prosody as well as acquisition of a correct pronunciation of the sounds of the foreign language. Concerning the mother tongue we are interested in the monitoring of the process of sound categorization in the long term (mainly at primary school) and its relation with the learning of reading and writing skills [8], especially for children with language deficiencies.

3.2. Statistical Modeling of Speech

Whereas the first research direction deals with the physical aspects of speech and its explicit modeling, this second research direction investigates statistical models for speech data. Acoustic models are used to represent the pronunciation of the sounds or other acoustic events such as noise. Whether they are used for source separation, for speech recognition, for speech transcription, or for speech synthesis, the achieved performance strongly depends on the accuracy of these models. At the linguistic level, MULTISPEECH investigates models for handling the context (beyond the few preceding words currently handled by the n-gram models) and evolutive lexicons necessary when dealing with diachronic audio documents. Statistical approaches are also useful for generating speech signals. Along this direction, MULTISPEECH considers voice transformation techniques, with their application to pathological voices, and statistical speech synthesis applied to expressive multimodal speech synthesis.

3.2.1. Source separation

Acoustic modeling is a key issue for automatic speech recognition. Despite the progress made for many years, current speech recognition applications rely on strong constraints (close-talk microphone, limited vocabulary, or restricted syntax) to achieve acceptable performance. The quality of the input speech signals is particularly important and performance degrades quickly with noisy signals. Accurate signal enhancement techniques are therefore essential to increase the robustness of both automatic speech recognition and speech-text alignment systems to noise and non-speech events.

In MULTISPEECH, focus is set on source separation techniques using multiple microphones and/or models of non-speech events. Some of the challenges include getting the most of the new modeling frameworks based on alpha-stable distributions and deep neural networks, combining them with established spatial filtering approaches, modeling more complex properties of speech and audio sources (phase, inter-frame and inter-frequency properties), and exploiting large data sets of speech, noise, and acoustic impulse responses to automatically discover new models. Beyond the definition of such models, the difficulty will be to design scalable estimation algorithms robust to overfitting, integrate them into the recently developed FASST [6] and KAM software frameworks if relevant, and develop new software frameworks otherwise.

3.2.2. Linguistic modeling

MULTISPEECH investigates lexical and language models in speech recognition with a focus on improving the processing of proper names and of spontaneous speech. Proper names are relevant keys in information indexing, but are a real problem in transcribing many diachronic spoken documents which refer to data, especially proper names, that evolve over time. This leads to the challenge of dynamically adjusting lexicons and language models through the use of the context of the documents or of some relevant external information. We also investigate language models defined on a continuous space (through neural network based approaches) in order to achieve a better generalization on unseen data, and to model long-term dependencies. We also want to introduce into these models additional relevant information such as linguistic features, semantic relation, topic or user-dependent information.

Other topics are spontaneous speech and pronunciation lexicons. Spontaneous speech utterances are often ill-formed and frequently contain disfluencies (hesitations, repetitions, ...) that degrade speech recognition performance. Hence the objective of improving the modeling of disfluencies and of spontaneous speech pronunciation variants. Attention will also be set on pronunciation lexicons with respect to non-native speech and foreign names. Non-native pronunciation variants have to take into account frequent mis-pronunciations due to differences between mother tongue and target language phoneme inventories. Proper name pronunciation variants are a similar problem where difficulties are mainly observed for names of foreign origin that can be pronounced either in a French way or kept close to foreign origin native pronunciation.

3.2.3. Speech generation by statistical methods

Over the last few years statistical speech synthesis has emerged as an alternative to corpus-based speech synthesis. The announced advantages of the statistical speech synthesis are the possibility to deal with small amounts of speech resources and the flexibility for adapting models (for new emotions or new speakers), however, the quality is not as good as that of the concatenation-based speech synthesis. MULTISPEECH will focus on a hybrid approach, combining corpus-based synthesis, for its high-quality speech signal output, and HMM-based speech synthesis for its flexibility to drive selection, and the main challenge will be on its application to producing expressive audio-visual speech.

Moreover, in the context of acoustic feedback in foreign language learning, voice modification approaches are investigated to modify the learner's (or teacher's) voice in order to emphasize the difference between the learner's acoustic realization and the expected realization.

3.3. Uncertainty Estimation and Exploitation in Speech Processing

This axis focuses on the uncertainty associated with some processing steps. Uncertainty stems from the high variability of speech signals and from imperfect models. For example, enhanced speech signals resulting

from source separation are not exactly the clean original speech signals. Words or phonemes resulting from automatic speech recognition contain errors, and the phone boundaries resulting from an automatic speech-text alignment are not always correct, especially in acoustically degraded conditions. Hence it is important to know the reliability of the results and/or to estimate the uncertainty of the results.

3.3.1. Uncertainty and acoustic modeling

Because small distortions in the separated source signals can translate into large distortions in the cepstral features used for speech recognition, this limits the recognition performance on noisy data. One way to address this issue is to estimate the uncertainty of the separated sources in the form of their posterior distribution and to propagate this distribution, instead of a point estimate, through the subsequent feature extraction and speech decoding stages. Although major improvements have been demonstrated in proof-of-concept experiments using knowledge of the true uncertainty, accurate uncertainty estimation and propagation remains an open issue.

MULTISPEECH seeks to provide more accurate estimates of the posterior distribution of the separated source signals accounting for, e.g., posterior correlations over time and frequency which have not been considered so far. The framework of variational Bayesian (VB) inference appears to be a promising direction. Mappings learned on training data and fusion of multiple uncertainty estimators are also explored. The estimated uncertainties are then exploited for acoustic modeling in speech recognition and, in the future, also for speech-text alignment. This approach may later be extended to the estimation of the resulting uncertainty of the acoustic model parameters and of the acoustic scores themselves.

3.3.2. Uncertainty and phonetic segmentation

The accuracy of the phonetic segmentation is important in several cases, as for example for the computation of prosodic features, for avoiding incorrect feedback to the learner in computer assisted foreign language learning, or for the post-synchronization of speech with face/lip images. Currently the phonetic boundaries obtained are quite correct on good quality speech, but the precision degrades significantly on noisy and non-native speech. Phonetic segmentation aspects will be investigated, both in speech recognition (i.e., spoken text unknown) and in forced alignment (i.e., when the spoken text is known).

In the same way that combining several speech recognition outputs leads to improved speech recognition performance, MULTISPEECH will investigate the combination of several speech-text alignments as a way of improving the quality of speech-text alignment and of determining which phonetic boundaries are reliable and which ones are not, and also for estimating the uncertainty of the boundaries. Knowing the reliability of the boundaries will also be useful when segmenting speech corpora; this will help deciding which parts of the corpora need to be manually checked and corrected without an exhaustive checking of the whole corpus.

3.3.3. Uncertainty and prosody

Prosody information is also investigated as a means for structuring speech data (determining sentence boundaries, punctuation...) possibly in addition to syntactic dependencies. Structuring automatic transcription output is important for further exploitation of the transcription results such as easier reading after the addition of punctuation, or exploitation of full sentences in automatic translation. Prosody information is also necessary for determining the modality of the utterance (question or not), as well as determining accented words.

Prosody information comes from the fundamental frequency, the duration of the sounds and their energy. Any error in estimating these parameters may lead to a wrong decision. MULTISPEECH will investigate estimating the uncertainty of the duration of the phones (see uncertainty of phonetic boundaries above) and on the fundamental frequency, as well as how this uncertainty shall be propagated in the detection of prosodic phenomena such as accented words, utterance modality, or determination of the structure of the utterance.

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 the 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 computer assisted learning, health and autonomy (more precisely aided communication and monitoring), annotation and processing of spoken documents, and multimodal computer interaction.

4.2. 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 a comparison of the learner's production to a reference, automatic diagnoses of the learner's production can be considered, as well as perceptual feedback relying on an automatic transformation of the learner's voice. The diagnosis step strongly relies on the studies on categorization of sounds and prosody in the mother tongue and in the second language. Furthermore, reliable diagnosis on each individual utterance is still a challenge, and elaboration of advanced automatic feedback requires a temporally accurate segmentation of speech utterances into phones and this explains why accurate segmentation of native and non-native speech is an important topic in the field of acoustic speech modeling.

4.3. Aided Communication and Monitoring

A foreseen application aims at improving the autonomy of elderly or disabled people, and fit with smartroom applications. In a first step, source separation techniques could be tuned and should help for locating and monitoring people through the detection of sound events inside apartments. 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) could also be envisaged.

4.4. 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 as for example, 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.

Large audio archives are important for some communities of users, e.g., linguists, ethnologists or researchers in digital humanities in general. In France, a notorious example is the "Archives du CNRS — Musée de l'homme", gathering about 50,000 recordings dating back to the early 1900s. When dealing with very old recordings, the practitioner is often faced with the problem of noise. This stems from the fact that a lot of interesting material from a scientific point of view is very old or has been recorded in very adverse noisy conditions, so that the resulting audio is poor. The work on source separation can lead to the design of semi-automatic denoising and enhancement features, that would allow these researchers to significantly enhance their investigation capabilities, even without expert knowledge in sound engineering.

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, prosody modification, and speaker conversion.

4.5. 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 audiobook, to facilitate the access to literature (for instance for blind people or illiterate people).

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Awards

Best student paper award at LTC'2017 (8th Language & Technology Conference) [39]

Third best paper award at ICNLSP'2017 (International Conference On Natural Language, Signal and Speech Processing) [42]

BEST PAPERS AWARDS:

[39]
A. HOUIDHEK, V. COLOTTE, Z. MNASRI, D. JOUVET, I. ZANGAR. *Statistical modelling of speech units in HMM-based speech synthesis for Arabic*, in "LTC 2017 - 8th Language & Technology Conference", Poznań, Poland, November 2017, pp. 1-5, https://hal.inria.fr/hal-01649034

[42]
D. JOUVET, D. LANGLOIS, M. A. MENACER, D. FOHR, O. MELLA, K. SMAÏLI. *About vocabulary adaptation for automatic speech recognition of video data*, in "ICNLSSP'2017 - International Conference on Natural Language, Signal and Speech Processing", Casablanca, Morocco, December 2017, pp. 1-5, https://hal.inria.fr/hal-01649057

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. This results in state-of-the-art separation quality for both speech and music.
FUNCTIONAL DESCRIPTION: Combines deep neural networks and multichannel signal processing for speech enhancement and separation of musical recordings.

NEWS OF THE YEAR: In 2017, we changed the type of multichannel filter used and modified the software so that it runs online in real time.

- Participants: Aditya Nugraha, Laurent Pierron, Emmanuel Vincent, Antoine Liutkus, Romain Serizel and Floris Fournier
- 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: Better acoustic models have been developed for French, English and Arabic languages. An NN-based speech detection module has been included, as well as rescoring with RNN language models.

- 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 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. . .). 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.
NEWS OF THE YEAR: SOJA now supports the unit selection with emotion tags.

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

6.4. Xarticulators

KEYWORD: Medical imaging
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.

NEWS OF THE YEAR: New models of the lips, velum and epiglottis have been added. Xarticulators generates area functions from an MRI film annotated in terms of articulators.

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

6.5. Platforms

6.5.1. Platform MultiMod: Multimodal Acquisition Data Platform

We have set up an acquisition hardware platform to acquire multimodal data in speech communication context. The system was previously composed of the articulograph Carstens AG501 (which was acquired as part of the EQUIPEX ORTOLANG), 4 Vicon cameras (a motion capture system), and an Intel RealSense camera which contains four components: a video camera, an infrared laser projector, an infrared camera, and a microphone array. With such heterogeneous hardware the synchronization is essential; this is achieved through a trigger device. All the data processing is performed with the PLAVIS software.

This year, we have replaced the 4 Vicon cameras by 8 optitrack cameras. The new motion capture system allows acquiring higher spatial and temporal resolution data, and allows faster acquisition and processing.

We are currently using the system to acquire expressive audiovisual data to build an expressive audiovisual speech synthesis in addition to a lipsync system.

- Participants: Slim Ouni, Vincent Colotte, Valerian Girard, Sara Dahmani
- Contact: Slim Ouni

7. New Results

7.1. Explicit Modeling of Speech Production and Perception

Participants: Anne Bonneau, Vincent Colotte, Yves Laprie, Slim Ouni, Agnès Piquard-Kipffer, Benjamin Elie, Theo Biasutto-Lervat, Sara Dahmani, Ioannis Douros, Valérian Girard, Yang Liu, Anastasiia Tsukanova.

7.1.1. Articulatory modeling

7.1.1.1. Articulatory models and synthesis

The geometry of the vocal tract is essential to guarantee the success of articulatory synthesis. This year we worked on the construction of an articulatory model of the epiglottis from MRI images and X-ray films. The new model takes into account the influences of the mandible, tongue and larynx via a multi-linear regression applied to the contours of the epiglottis [44]. Once these influences are removed from the contours, principle component analysis is applied to the control points of the B-spline representing the centerline of the epiglottis. The main advantage of using the centerline is to reduce the effect of delineation errors. Following the same idea, we also developed an articulatory model of the velum.

Geometry of the vocal tract is an input of articulatory synthesis and an algorithm for controlling the positions of speech articulators (jaw, tongue, lips, velum, larynx and epiglottis) is required to produce given speech sounds, syllables and phrases. This control has to take into account coarticulation and be flexible enough to be able to vary strategies for speech production [65]. The data for the algorithm are 97 static MRI images capturing the articulation of French vowels and blocked consonant-vowel syllables. The results of this synthesis are evaluated visually, acoustically and perceptually, and the problems encountered are broken down by their origin: the dataset, its modeling, the algorithm for managing the vocal tract shapes, their translation to the area functions, and the acoustic simulation.

7.1.1.2. Acoustic simulations

The acquisition of EPGG data (ElectroPhotoGlottoGraphy) data in collaboration with LPP in Paris has allowed the exploration of the production of voiced and unvoiced fricatives with realistic glottis opening profiles. These data show that the glottal opening is gradual and starts well before the fricative itself. Production of fricatives were studied by using acoustic simulations based on classic lumped circuit element methods to compute the propagation of the acoustic wave along the vocal tract. The glottis model incorporating a glottal chink developed last year is connected to the wave solver to simulate a partial abduction of the vocal folds during their self-oscillating cycles. Area functions of fricatives at the three places of articulation of French (palato-alveolar, alveolar, and labiodental) have been extracted from static MRI acquisitions. Simulations highlight the existence of three distinct regimes, named A, B, and C, depending on the degree of abduction of the glottis. They are characterized by the frication noise level: A exhibits a voiced signal with a low frication noise level, B is a mixed noise/voiced signal, and C contains only frication noise [33], [12].

Following the same approach of coupling articulatory data and acoustic simulation we investigated the acoustic simulation of alveolar trills, and the articulatory and phonatory configurations that are required to produce them. Using a realistic geometry of the vocal tract, derived from cineMRI data of a real speaker, the mechanical behavior of a lumped two-mass model of the tongue tip was studied [13]. The incomplete occlusion of the vocal tract during linguopalatal contacts was modeled by adding a lateral acoustic waveguide. Finally, the simulation framework is used to study the impact of a set of parameters on the characteristic features of the produced alveolar trills. It shows that the production of trills is favored when the tongue tip position is slightly away of the alveolar zone, and when the glottis is fully adducted.

7.1.1.3. Acquisition of articulatory data

The effort of acquiring new articulatory data was quite strong this year: **(i)** acquisition of MRI films (136 x 136 pixel images at a sampling rate of 55Hz) of continuous speech in Max Planck Institute Göttingen with Prof. Jens Frahm. We collected 2 hours of speech for 2 male speakers covering sentences and spontaneous speech. The sentences were designed so as to contain all the consonants and consonant clusters (excepted the very rare ones) in four vocalic contexts (the three cardinal vowels and /y/) and some intermediate vowels to check how they can be derived from those extreme vowels. The acoustic speech signal was recorded and denoised. Orthographic annotations of speech are available and the phonetic alignments were computed from the denoised speech signal. **(ii)** acquisition of EPGG (ElectroPhotoGlottoGraphy) data in LPP (Laboratoire de Phonologie et de Phonétique in Paris). The principle is to measure the flow of light (infrared light) which crosses the glottis. The emitting source is placed above the glottis and a light sensor below. The flow of light crossing the obstacle is roughly proportional to the surface of the glottis. Data acquired cover VCVs for fricatives and stops and some consonant clusters. These data were used to study the coordination between glottis opening and the realization of constrictions in the vocal tract. **(iii)** acquisition of fibroscopy data in HEGP (Georges Pompidou European Hospital). The principle is to introduce a smooth endoscope through the nostrils up to the top of the pharynx so as to image the glottis opening. This technique only allows a frequency close to 50 Hz which is not sufficient to observe the smooth glottis opening profiles accompanying the production of fricatives. Data have been collected for one female speaker and two male speakers.

7.1.2. Expressive acoustic and visual synthesis

We have improved our audiovisual acquisition techniques by acquiring a very advanced 8-camera motion capture system that allows capturing 3D data with higher temporal resolution and accuracy. We have acquired a small corpus for testing and evaluation purpose.

Within the framework expressive audiovisual speech synthesis, a perceptive case study on the quality of the expressiveness of a set of emotions acted by a semi-professional actor has been conducted. We have analyzed the production of this actor pronouncing a set of sentences with acted emotions, during a human emotion-recognition task. We have observed different modalities: audio, real video, 3D-extracted data, as unimodal presentations and bimodal presentations (with audio). The results of this study show the necessity of such perceptive evaluation prior to further exploitation of the data for the synthesis system. The comparison of the

modalities shows clearly what the emotions are, that need to be improved during production and how audio and visual components have a strong mutual influence on emotional perception [57].

7.1.3. Categorization of sounds and prosody for native and non-native speech

7.1.3.1. Categorization of sounds for native speech

Concerning the mother tongue, we conducted empirical research. We followed 170 young people, aged from 6 to 20 years old, with langage deficiencies - dyslexia and Specific Language Impairment (SLI) - including categorization of sounds. We examined the links between those difficulties and their schooling experience and observed how they constituted a point of major obstacle at the time of learning to read and to write, which the pupils do not overcome. All of them were in a handicap situation [18].

We conducted two descriptive studies which aims were to give an overview of educational systems for students with special educational needs, including pupils with learning and sound categorization disabilities (LD). Around the world, schooling is different from one country to another, according to the languages, even every country follows the international movement of school for all. For these students, the question of the best mode of inclusion remains topical [16]. In France, different types of scooling are observed. We focused our study on a particular system of teaching - a local unit for inclusive education - for children aged from 6 to 12 with specific language disorders - dyslexia and SLI - and learning disabilities, in a specialised school. We described a few exemples of pedagogical multimodal accomodations [15].

7.1.3.2. Digital books for language impaired children

In the framework of Handicom ADT project [7], we used one of the digital books prototypes set up with the use of a 3D avatar as narrator and multimodal speech, combining oral, written language and visual clues (i.e. LPC, french cued speech), specially targeting children between 3 and 6. After the study conducted with digital album users, speech-therapists or re-eductors with hearing impaired impaired children, SLI and children with autism [81], we conducted another study, following children at school to investigate how technological innovations could help kindergarten children's (with and without langage difficulties) to improve their speech and language habilities.

7.1.3.3. Analysis of non-native pronunciations

Deviations in L2 intonation affect a number of prosodic characteristics including pitch range, declination line, or the rises of non-final intonation phrases, and might lead to misunderstandings or contribute to the perception of foreign-accent. This study investigates the characteristics of non-native speech at the boundary between prosodic constituents [67]. We analyzed a French declarative sentence, extracted from the IFCASL corpus (http://www.ifcasl.org), made up of four constituents and pronounced with a neutral intonation. Each constituent has three syllables and the sentence is realized typically by French speakers with four accentual –prosodic- groups, corresponding to the four constituents. Fourty German learners of French (beginners, and advanced speakers) and fifty four French speakers read the sentence once. We used the software ProsodyPro from Yi Xu for the prosodic analysis. We determined the presence of pauses and evaluated for each prosodic group: the (normalized) F0 maximum on the last syllable; the F0 excursion (max-min) of the final contour, and its maximum of velocity. In order to analyze the temporal course of F0 on the final contour, we also compared the values of the F0 excursion on the vowel and before it. On the basis of acoustic cues, non-native speakers, especially beginners, appear to realize more important prosodic boundaries (in particular higher F0 maxima, especially at the very end of the prosodic group, and more pauses) than French speakers, whereas native speakers appear to show more anticipation.

7.2. Statistical Modeling of Speech

Participants: Vincent Colotte, Dominique Fohr, Irène Illina, Denis Jouvet, Antoine Liutkus, Odile Mella, Romain Serizel, Emmanuel Vincent, Md Sahidullah, Guillaume Carbajal, Ken Deguernel, Mathieu Fontaine, Amal Houidhek, Aditya Nugraha, Laureline Perotin, Imran Sheikh, Sunit Sivasankaran, Ziteng Wang, Ismaël Bada.

7.2.1. Source separation

We wrote an extensive overview article about multichannel source separation and speech enhancement [14] and two book chapters about single-channel [72] and multichannel separation based on nonnegative matrix factorization [74].

7.2.1.1. Deep neural models for source separation and echo suppression

We pursued our research on the use of deep learning for multichannel source separation. In our previous work, which we summarized in a book chapter [73], we estimated the short-time spectra of the sound sources by a deep neural network and their spatial covariance matrices by a classical expectation-maximization (EM) algorithm and we derived the source signals by a multichannel Wiener filter. We also explored several variants of the multichannel Wiener filter, which turned out to result in better speech recognition performance on the CHiME-3 dataset [23]. We developed a new "end-to-end" approach which estimates both the short-time spectra and the spatial covariance matrices by a dedicated deep neural network architecture and which outperforms previously proposed approaches on CHiME-3. Arie Aditya Nugraha described the latter approach in his thesis, which he successfully defended. We started exploring the usage of deep neural networks for reducing the residual nonlinear echo after linear acoustic echo cancellation [80] and for separating multiple speakers from each other.

We also continued our work on music source separation, with the organization of the successful Signal Separation Evaluation Challenge (SiSEC 2016 [46]), as well as with national and international collaborations on this topic [34], [47], [58], [59], [60]. This research activity features several important research directions, described below.

7.2.1.2. Alpha-stable modeling of audio signals

Under the KAMoulox funding, we investigated the use of alpha-stable probabilistic models for source separation. As opposed to their more classical counterparts, these models feature very heavy tails, which allows to better account for the large dynamics found in audio signals. In close collaboration with national and international partners, we published several papers in international conferences on these topics. We demonstrated that alpha-stable processes allow to understand long-standing practices in speech enhancement [36]. More specifically, we showed that parameterized Wiener filters, dating back to the early 80s, can be understood as the optimal filtering strategy when sources are distributed with respect to alpha-stable distributions of different characteristic exponents. Interestingly, this gives a rationale for setting filtering parameters that were always manually tuned. Stable distributions also allow generalizing Wiener filtering for nonnegative sources [48], [49], and are interesting for robust multichannel separation [45], in the sense that they permit to compensate for model mismatch efficiently.

7.2.1.3. Scalable source localization

In the context of KAMoulox, we studied how probabilistic modeling of multichannel audio with alpha-stable distributions leads to models for microphone arrays that allow for scalable inference for the source positions [37], [38]. The core points of these methods are twofold. First, heaviness of the tails of alpha-stable distributions allows to efficiently model the marginal distribution of sources spectra. This is in sharp contrast with Gaussian distributions, that can only correctly represent audio signals adequately if each time-frequency point has its own distribution. On the contrary, while alpha-stable distributions give a high probability mass to small magnitudes, they also allow for the important deviations to be expected when the source is active. The advantage of such a model for marginal distributions over the whole time-frequency plane is to dramatically reduce the number of parameters and thus lead to much robust estimation methods. The second innovation brought in by the proposed localization method is to compute a summarized representation of the data, and to proceed to inference on this representation instead of using the -massive- original data.

7.2.1.4. Interference reduction

Under the DYCI2 schedule, we significantly extended our previous research on interference reduction for musical recordings. This task consists in reducing inter-microphone leakage in live recordings and has many applications in the audio engineering industry. This lead us to propose two important contributions on this

respect. First, we amended previous methods to correctly exploit the proposed probabilistic model: previous research indeed featured some ad-hoc and suboptimal steps. This was corrected and the corresponding extension proved to behave much better [30]. Second, we investigated whether the proposed methods can be generalized to process full-length recordings. This is indeed an important and challenging question, because full-length multitrack recordings are extremely large and cannot reasonably be processed with current methods. This line of research lead us to propose inferring some parameters on compressed representations, which is promising ongoing research.

7.2.2. Acoustic modeling

7.2.2.1. Noise-robust acoustic modeling

In many real-world conditions, the target speech signal is reverberated and noisy. We conducted an extensive evaluation of several approaches for speech recognition in varied reverberation conditions, including both established and newly proposed approaches [21].

Speech enhancement and automatic speech recognition (ASR) are most often evaluated in matched (or multicondition) settings where the acoustic conditions of the training data match (or cover) those of the test data. We conducted a systematic assessment of the impact of acoustic mismatches (noise environment, microphone response, data simulation) between training and test data on the performance of recent DNN-based speech enhancement and ASR techniques [22]. The results show that multi-condition training outperforms matched training on average, but training on a subset of noise environments only is preferable in a few specific cases [25]. This raises the question: what are the optimal training conditions given the task to be solved, the deep neural network architecture, and the test conditions? We provided a preliminary answer to this question by means of a discriminative importance weighting algorithm which aims to select the most useful training data in a rigorous optimization framework [64].

In order to motivate further work by the community, we created the series of CHiME Speech Separation and Recognition Challenges in 2011. Following the organization of the CHiME-3 Challenge in 2015, we edited a special issue [9] of *Computer Speech and Language*, which includes a detailed description of its outcomes [10]. We also published a book chapter that summarizes the outcomes of the whole series of challenges [70].

7.2.2.2. Environmental sounds

Following the recruitment of Romain Serizel in Fall 2016, our team has become more involved in the community on environmental sound recognition. In collaboration with Carnegie Mellon University (USA), we co-organized the first ever large-scale environmental sound recognition evaluation. This evaluation relied on the Audioset corpus released by Google and was part of the DCASE 2017 Challenge [24]. It focused on the problem of learning from weak labels for an application to smart cars.

We continued our work on acoustic scene classification. In particular, we focused on exploiting matrix factorization techniques for features learning. We extended previous work that used these learned features as an input to a linear classifier [11] to the deep learning framework [27], [28] and we proposed to jointly learn the deep-learning based classifier and the dictionary matrix [27]. A system based on this approach was submitted to DCASE challenge and was among the top 25% systems [28].

7.2.2.3. Speech/Non-speech detection

Automatic Speech Recognition (ASR) of multimedia content such as videos or multi-genre broadcasting requires a correct extraction of speech segments. We explored the efficiency of deep neural models for speech/non-speech segmentation. The first results, achieved in the MGB Challenge framework, show an improvement of the ASR word error rate compared to a Gaussian Mixture Model (GMM) based speech/non-speech segmenter.

7.2.2.4. Data selection

Training a speech recognition system needs audio data and their corresponding exact transcriptions. However, manual transcribing is expensive, labor intensive and error-prone. Some sources, such as TV broadcast, have subtitles. Subtitles are closed to the exact transcription, but not exactly the same. Some sentences might be paraphrased, deleted, changed in word order, etc. Building automatic speech recognition from inexact

subtitles may result in a poor model and low performance system. Therefore, selecting data is crucial to obtain highly efficient models. We study data selection methods based on phone matched error rate and average word duration [26]

7.2.2.5. Transcription systems

We designed a new automatic transcription system based on deep learning with an acoustic modeling done by TDNN-HMM and a language model rescoring using RNN. In the framework of the AMIS project, we developped automatic systems for the transcription of TV shows in English, in French and in Arabic [52] [51].

7.2.2.6. Speaker identification

We proposed supervised feature learning approaches for speaker identification that rely on nonnegative matrix factorization [61]. The approach integrates a recent method that relies on group nonnegative matrix factorization into a task-driven supervised framework for speaker identification [11]. The goal is to capture both the speaker variability and the session variability while exploiting the discriminative learning aspect of the task-driven approach.

7.2.3. Language modeling

7.2.3.1. Out-of-vocabulary proper name retrieval

The diachronic nature of broadcast news causes frequent variations in the linguistic content and vocabulary, leading to the problem of Out-Of-Vocabulary (OOV) words in automatic speech recognition. Most of the OOV words are found to be proper names whereas proper names are important for automatic indexing of audio-video content as well as for obtaining reliable automatic transcriptions. New proper names missed by the speech recognition system can be recovered by a dynamic vocabulary multi-pass recognition approach in which new proper names are added to the speech recognition vocabulary based on the context of the spoken content. We proposed a Neural Bag-of-Weighted Words (NBOW2) model which learns to assign higher weights to words that are important for retrieval of an OOV PN. [20]. We explored topic segmentation in ASR transcripts using bidirectional RNNs for change detection [62].

7.2.3.2. Adding words in a language model

We proposes new approaches to OOV proper noun probability estimation using Recurrent Neural Network Language Model (RNNLM). The proposed approaches are based on the notion of closest in-vocabulary words (list of brothers) to a given OOV proper noun. The probabilities of these words are used to estimate the probabilities of OOV proper nouns thanks to RNNLM [40].

7.2.3.3. Updating speech recognition vocabularies

In the framework of the AMIS project, the update of speech recognition vocabularies has been investigated using web data collected over a time period similar to that of the collected videos, for three languages: French, English and Arabic [42]. Results shows that a significant reduction of the amount of out-of-vocabulary words is observed for the three languages, and that, for a given vocabulary size, the percentage of out-of-vocabulary words is higher for Arabic than for the other languages.

7.2.3.4. Segmentation and classification of opinions

Automatic opinion/sentiment analysis is essential for analysing large amounts of text as well as audio/video data communicated by users. This analysis provides highly valuable information to companies, government and other entities, who want to understand the likes, dislikes and feedback of the users and people in general. We proposed a recurrent neural network model with bi-directional LSTM-RNN, to perform joint segmentation and classification of opinions [63].

7.2.3.5. Music language modeling

Similarly to speech, music involves several levels of information, from the acoustic signal up to cognitive quantities such as composer style or key, through mid-level quantities such as a musical score or a sequence of chords. The dependencies between mid-level and lower- or higher-level information can be represented through acoustic models and language models, respectively [79]. In the context of ANR DYCI2, we described a general framework for automatic music improvisation that encompasses three existing paradigms [56] and that relies on our previous work about combining a multi-dimensional probabilistic model encoding the musical experience of the system and a factor oracle encoding the local context of the improvisation. Inspired in particular by the regularity of the temporal structure of popular music pieces [19], we proposed a new polyphonic music improvisation approach that takes the structure of the musical piece at multiple time scales into account [32].

7.2.4. Speech generation

Work on Arabic speech synthesis was carried out within a CMCU PHC project with ENIT (École Nationale d'Ingénieurs de Tunis, Tunisia, cf. 9.4.2.1), using HMM and NN based approaches applied to Modern Standard Arabic language.

HMM-based speech synthesis system relies on a description of speech segments corresponding to phonemes, with a large set of features that represent phonetic, phonologic, linguistic and contextual aspects. When applied to Modern Standard Arabic, two specific phenomena have to be taken in account, the vowel quantity and the consonant gemination. This year, we studied thoroughly the modeling of these phenomena. Results of objective and subjective evaluations showed that the results are similar between the different approaches that have been studied [39]. Other similar experiments are on-going using neural-network-based synthesis.

A particular weakness point of HMM-based synthesis quality may be due to the prediction of prosodic features which is based on a decision tree approach. Neural network are known for their ability to model complex relationships. This year, we studied the modeling of phoneme duration with NN approaches. Predicted phoneme durations will then be included in the Modern Standard Arabic synthesis system.

In parallel, the neural network based approach has also been tested on the French language.

7.3. Uncertainty Estimation and Exploitation in Speech Processing

Participants: Vincent Colotte, Dominique Fohr, Denis Jouvet, Yves Laprie, Odile Mella, Emmanuel Vincent, Yassine Boudi, Mathieu Hu, Karan Nathwani.

7.3.1. Uncertainty and acoustic modeling

7.3.1.1. Uncertainty in noise-robust speech and speaker recognition

In many real-world conditions, the target speech signal overlaps with noise and some distortion remains after speech enhancement. The framework of uncertainty decoding assumes that this distortion has a Gaussian distribution and seeks to estimate its covariance matrix and propagate it through the acoustic model for robust ASR. We conducted an extensive experimental investigation of existing uncertainty estimation and propagation techniques using deep neural network acoustic models on two different datasets (CHiME-2 and CHiME-3) [53]. We also proposed a deep neural network-based uncertainty estimator and a consistent way of accounting for uncertainty in both the training and decoding stage [54]. Overall, we were the first to report a significant improvement using uncertainty estimation and propagation compared to a competitive deep neural network acoustic modeling baseline based on feature-domain maximum likelihood linear regression (fMLLR) features.

7.3.1.2. Uncertainty in other applications

Besides the above applications, we pursued our exploration of uncertainty modeling for robot audition and wind turbine control. In the first context, uncertainty arises about the location of acoustic sources and the robot is controlled to locate the sources as quickly as possible [55]. In his successfully defended thesis, Quan Van Nguyen also described a way of locating multiple sources. In the second context, uncertainty arises about the noise intensity of each wind turbine and the turbines are controlled to maximize electrical production under a maximum noise threshold [31].

7.3.2. Uncertainty and phonetic segmentation

In the framework of the LCHN CPER project (cf. 9.1.1), for studying prosodic correlates of discourse particles in French, phonetic boundaries of discourse particles and adjacent words have been checked and manually corrected; this shows that there is still a need for performance improvement of the automatic speech-text alignment process.

We also worked on speech-to-speech alignement, with the goal of obtaining a precise alignement between two speakers pronouncing the same sentence. This task is difficult due to the fact that the speakers may pronounce certain sounds in a different way, or they may insert or remove silences between words. We introduced explicit phoneme duration and insertion/deletion models for alignment and evaluated them on real data.

7.3.3. Uncertainty and prosody

The fundamental frequency is one of the prosodic features. Numerous approaches exist for the computation of F0. Most of them lead to good performance on good quality speech. The performance degradation with respect to noise level has been studied on reference databases, for several (about ten) F0 detection approaches. It was observed that for each algorithm, a large part of the errors are due to incorrect voiced/unvoiced decision [43]. A first set of experiments have been conducted for computing a confidence measure on the estimated F0 values through the use of neural network approaches [29].

Study of discourse particles in French has continued thanks to the support of the CPER LCHN project. So far a few French words frequently used as discourse particles have been studied. Several thousands occurrences have been extracted from the ESTER and the ORFEO speech corpora, and annotated as discourse particle or not. The pragmatic function of the discourse particles has also been annotated. Prosodic correlates of these words have been analyzed with respect to their function (discourse particle or not, as well as pragmatic function) [66], and some automatic classification processes have been investigated [41].

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. Orange

Company: Orange SA (France)

Duration: Nov 2016 – Nov 2019

Participants: Laureline Perotin, Romain Serizel, Emmanuel Vincent

Abstract: This CIFRE contract funds the PhD thesis of Laureline Perotin with Orange Labs. Our goal is to develop deep learning based speaker localization and speech enhancement algorithms for robust hands-free voice command. We are especially targetting difficult scenarios involving several simultaneous speakers.

8.1.2. Invoxia

Company: Invoxia SAS (France)

Duration: Mar 2017 – Mar 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.

8.1.3. Studio Maia

Company: Studio Maia SARL (France)

Other partners: Imaging Factory

Duration: Jul 2017 – Dec 2018

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

Abstract: This Inria Innovation Lab aims to develop a software suite for voice processing in the multimedia creation chain. The software is aimed at sound engineers and it will rely on the team's expertise in speech enhancement, robust speech and speaker recognition, and speech synthesis.

8.1.4. Samsung

Company: Samsung Electronics Co., Ltd (South Korea)

Duration: Jan – Nov 2017

Participants: Aditya Nugraha, Romain Serizel, Emmanuel Vincent

Abstract: This project aimed to transfer a modified version of dnnsep for hands-free voice command applications. We changed the type of multichannel filter used and modified the software so that it runs online in real time.

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 of the project 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. This year we have also developped a complete system for the transcription of English broadcast TV shows to participate to the MGB challenge.

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, Vishnu Varanasi, Emmanuel Vincent

Abstract: The goal of the project is to develop innovative technologies for health, and tools and strategies for personalized medicine.

MULTISPEECH will investigate acoustic monitoring using an array of microphones.

9.1.3. Dynalips

Project title: Control of the movements of the lips in the context of facial animation for an intelligible lipsync.

Duration: February 2017 - January 2018

Coordinator: Slim Ouni

Participants: Valerian Girard, Slim Ouni

Funding: SATT

Abstract: We propose in this project the development of tools of lipsync which from recorded speech will provide realistic mechanisms of animating the lips. These tools will be available to be integrated into existing 3D animation software and existing game engines. One objective is that these lipsync tools fit easily into the production pipeline in the field of 3D animation and video games. The goal of this maturation is to propose a product ready to be exploited in the industry whether by the creation of a start-up or by the distribution of licenses.

9.2. National Initiatives

9.2.1. 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, Denis Jouvet, Odile Mella, Slim Ouni

Abstract: METAL aims at improving the learning of languages (both written and oral components) through the development of new tools and the analysis of numeric traces associated with students' learning, in order to adapt to the needs and rythm of each learner.

MULTISPEECH is concerned by oral language learning aspects.

9.2.2. PIA2 ISITE LUE

Project acronym: ISITE LUE

Project title: Lorraine Université d'Excellence

Duration: starting in 2016

Coordinator: Univ. Lorraine

Participants: Ioannis Douros, Yves Laprie

Abstract: The initiative aims at developing and densifying the initial perimeter of excellence, within the scope of the social and economic challenges, so as to build an original model for a leading global engineering university, with a strong emphasis on technological research and education through research. For this, we have designed LUE as an "engine" for the development of excellence, by stimulating an original dialogue between knowledge fields.

MULTISPEECH is mainly concerned with challenge number 6: "Knowledge engineering", i.e., engineering applied to the field of knowledge and language, which represent our immaterial wealth while being a critical factor for the consistency of future choices. In 2016, this project has funded a new PhD thesis.

9.2.3. ANR ContNomina

Project acronym: ContNomina

Project title: Exploitation of context for proper names recognition in diachronic audio documents

Duration: February 2013 - March 2017

Coordinator: Irina Illina

Other partners: LIA, Synalp

Participants: Dominique Fohr, Irina Illina, Denis Jouvet, Odile Mella, Imran Sheikh

Abstract: The ContNomina project was focus on the problem of proper names in automatic audio processing systems by exploiting in the most efficient way the context of the processed documents. To do this, the project has addressed the statistical modeling of contexts and of relationships between contexts and proper names; the contextualization of the recognition module (through the dynamic adjustment of the lexicon and of the language model in order to make them more accurate and certainly more relevant in terms of lexical coverage, particularly with respect to proper names); and the detection of proper names (on the one hand, in text documents for building lists of proper names, and on the other hand, in the output of the recognition system to identify spoken proper names in the audio/video data).

MULTISPEECH contributes to speech recognition and proper names handling (prediction, introduction in models, ...)

9.2.4. ANR DYCI2

Project acronym: DYCI2 (http://repmus.ircam.fr/dyci2/)

Project title: Creative Dynamics of Improvised Interaction

Duration: March 2015 - February 2018

Coordinator: Ircam (Paris)

Other partners: Inria (Nancy), University of La Rochelle

Participants: Ken Deguernel, Nathan Libermann, Emmanuel Vincent

Abstract: The goal of this project is to design a music improvisation system which will be able to listen to the other musicians, improvise in their style, and modify its improvisation according to their feedback in real time.

MULTISPEECH is responsible for designing a system able to improvise on multiple musical dimensions (melody, harmony) across multiple time scales.

9.2.5. ANR JCJC KAMoulox

Project acronym: KAMoulox

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

Duration: January 2016 - January 2019

Coordinator: Antoine Liutkus

Participants: Mathieu Fontaine, Antoine Liutkus

Abstract: The objective is to develop the theoretical and applied tools required to embed audio denoising and separation tools in web-based audio archives. The applicative scenario is to deal with large audio archives, and more precisely with the notorious “Archives du CNRS — Musée de l’homme”, gathering about 50,000 recordings dating back to the early 1900s.

9.2.6. ANR ArtSpeech

Project acronym: ArtSpeech

Project title: Synthèse articulatoire phonétique

Duration: October 2015 - March 2019

Coordinator: Yves Laprie

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

Participants: Ioannis Douros, Benjamin Elie, Yves Laprie, Anastasiia Tsukanova

Abstract: The objective is to synthesize speech from text via the numerical simulation of the human speech production processes, i.e. the articulatory, aerodynamic and acoustic aspects. Corpus based approaches have taken a hegemonic place in text to speech synthesis. They exploit very good acoustic quality speech databases while covering a high number of expressions and of phonetic contexts. This is sufficient to produce intelligible speech. However, these approaches face almost insurmountable obstacles as soon as parameters intimately related to the physical process of speech production have to be modified. On the contrary, an approach which rests on the simulation of the physical speech production process makes explicitly use of source parameters, anatomy and geometry of the vocal tract, and of a temporal supervision strategy. It thus offers direct control on the nature of the synthetic speech.

Acquisition and processing of cineMRI, new developments of acoustic simulations concercing the production of fricatives and trills, and first workds in the implementation of coarticulation in articulatory synthesis are the main activities of this year.

9.2.7. ANR VOCADOM

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

Project title: Robust voice command adapted to the user and to the context for AAL

Duration: January 2017 - December 2020

Coordinator: CNRS - LIG (Grenoble)

Other partners: Inria (Nancy), Univ. Lyon 2 - GREPS, THEORIS (Paris)

Participants: Dominique Fohr, Sunit Sivasankaran, Emmanuel Vincent

Abstract: The goal of this project is to design a robust voice control system for smart home applications. We are responsible for the speech enhancement and robust automatic speech recognition bricks.

MULTISPEECH is responsible for wake-up word detection, overlapping speech separation, and speaker recognition.

9.2.8. FUI VoiceHome

Project acronym: VoiceHome

Duration: February 2015 - July 2017

Coordinator: VoiceBox Tachnologies France

Other partners: Orange, Delta Dore, Technicolor Connected Home, eSoftThings, Inria (Nancy), IRISA, LOUSTIC

Participants: Irina Illina, Karan Nathwani, Emmanuel Vincent

Abstract: The goal of this project was to design a robust voice control system for smart home and multimedia applications. We were responsible for the robust automatic speech recognition brick.

MULTISPEECH was responsible for robust automatic speech recognition by means of speech enhancement and uncertainty propagation.

9.2.9. MODALISA

Project acronym: MODALISA

Project title: Multimodality during Language Acquisition: Interaction between Speech Signal and gestures

Duration: January 2017 - December 2017

Coordinator: Christelle Dodane (Praxiling, UMR 5267, Montpellier)

Other partners: Slim Ouni

Participants: Slim Ouni

Funding: CNRS DEFI Instrumentation aux limites

Abstract: The objective of this project was to setup a multimodal platform allowing simultaneous visualization of gestural (motion capture system) and prosodic data during speech and more specifically during language acquisition.

Les contributions de MULTISPEECH concernent l'acquisition et le traitement des données multimodales grâce à la plateforme multimodale MultiMod.

9.3. European Initiatives

9.3.1. Collaborations in European Programs, Except FP7 & H2020

9.3.1.1. AMIS

Program: CHIST-ERA

Project acronym: AMIS

Project title: Access Multilingual Information opinionS

Duration: Dec 2015- Nov 2018

Coordinator: Kamel Smaïli

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

Participants: Dominique Fohr, Denis Jouvet, Odile Mella

Abstract: The idea of the project is to develop a multilingual help system of understanding without any human being intervention. What the project would like to do, is to help people understanding broadcasting news, presented in a foreign language and to compare it to the corresponding one available in the mother tongue of the user.

MULTISPEECH contributions concern mainly the speech recognition in French, English and Arabic videos.

9.3.2. Collaborations with Major European Organizations

Jon Barker: University of Sheffield (UK)

Robust speech recognition [22], [10], [9], [70]

9.4. International Initiatives

9.4.1. Inria International Partners

9.4.1.1. Informal International Partners

Shinji Watanabe, Johns Hopkins University (USA)

Robust speech recognition [22], [10], [9], [70]

9.4.2. Participation in Other International Programs

9.4.2.1. PHC UTIQUE - Arabic speech synthesis

PHC UTIQUE - Arabic speech synthesis, with ENIT (École Nationale d'Ingénieurs de Tunis, Tunisia)

Duration: 2015 - 2018.

Coordinators: Vincent Colotte (France) and Zied LACHIRI (Tunisia).

Participants: Vincent Colotte, Amal Houidhek, Denis Jouvet

Abstract: Modeling of a speech synthesis system for the Arabic language. This includes the use of an Arabic speech corpus, the selection of linguistic features relevant to an Arabic speech synthesis, as well as improving the quality of the speech signal generated by the system (prosodic and acoustic features).

MULTISPEECH co-supervises PhD students.

9.4.2.2. FIRAH - La famille face au handicap

Program: FIRAH, International Foundation of Applied Disability Research

Project title: La famille face au handicap : la gestion du stress parental des parents d'enfants souffrant du syndrome de Dravet

Duration: Jan 2017- Dec 2019

Coordinator: T. Leonova, University of Lorraine (Perseus)

Other partners: MHS-USR 3261 CNRS, Université de Lorraine, Associations Alliance Syndrome de Dravet (France) and Alliance Syndrome de Dravet (Suisse), Hopital de Hautepierre - Strasbourg University (France), Hopital Necker enfants malades - Paris Descartes University - INSERM U1129, Hôpital Robert Debré - Paris Diderot University- INSERM U1141, Hôpitaux Universitaires de Genève - Université de Genève (Suisse), Université catholique du Sacré Cœur - Rome (Italie), Quebec University (Canada), McMaster Children's Hospital - McMaster University - Hamilton (Canada), MIA518-AgroParisTech/INRA.

Participant: Agnès Piquard-Kipffer

Abstract: the aims of the project are, in a first step, to explore parental stress with Chidren with Dravet syndrom which combine infant epilepsy and autism and in a second step to create a training programm for professionnals of Education [68], [69]

In this project, MULTISPEECH is involved in finding the best ways to maximize the communication efficiency between the children and their families, using the methodology or the tools created by the Handicom project.

9.5. International Research Visitors

9.5.1. Visits of International Scientists

Ziteng Wang

Date: Sep 2016 – Sep 2017

Institution: Institute of Acoustics, Chinese Academy of Sciences (China)

Vishnuvardhan Varanasi

Date: Feb – Aug 2017

Institution: Indian Institute of Science, Kanpur (India)

Md Sahidullah

Date: Aug – Oct 2017

Institution: University of Eastern Finland (Finland)

9.5.2. Visits to International Teams

9.5.2.1. Research Stays Abroad

Antoine Liutkus was invited by Kazuyoshi Yoshii (RIKEN, Kyoto University) to work on multichannel extensions to his tensor-factorization methods, that would also allow for much easier inference. This led to a joint publication [47] about the resulting method.

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)

General co-chair, AVSP 2017 - 14th International Conference on Auditory-Visual Speech Processing (S. Ouni)

10.1.1.2. Member of the Organizing Committees

Co-organizer of the Task "Large-scale weakly supervised sound event detection for smart cars", DCASE 2017 Challenge on Detection and Classification of Acoustic Scenes and Events (E. Vincent)

Member of the organizing committee, 2017 IEEE Automatic Speech Recognition and Understanding Workshop, Okinawa, Japan, December 2017 (E. Vincent)

Member of the steering committee, Detection and Classification of Acoustic Scenes and Events (DCASE) challenge series (E. Vincent)

Co-organizer of AVSP 2017 - International Conference on Auditory-Visual Speech Processing (S. Ouni)

10.1.2. Scientific Events Selection

10.1.2.1. Chair of Conference Program Committees

Program chair, DCASE 2017 Workshop on Detection and Classification of Acoustic Scenes and Events (E. Vincent)

Review chair, IEEE Technical Committee on Audio and Acoustic Signal Processing, responsible for organizing the review of the 313 papers submitted to the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) in the general AASP domain (E. Vincent)

Proceedings co-chair (editor), AVSP 2017 - 14th International Conference on Auditory-Visual Speech Processing [76] (S. Ouni)

10.1.2.2. Member of the Conference Program Committees

AVSP 2017 - 14th International Conference on Auditory-Visual Speech Processing (S. Ouni)

ICNLSSP'2017 - International Conference on Natural Language, Signal and Speech Processing (D. Fohr, D. Jouvet, O. Mella)

Area chair, 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (E. Vincent)

ISSP 2017 - International Seminar On Speech Production (Y. Laprie)

10.1.2.3. Reviewer

ASRU'2017 - IEEE Automatic Speech Recognition and Understanding Workshop (D. Jouvet, I. Illina, E. Vincent)

AVSP 2017 - 14th International Conference on Auditory-Visual Speech Processing (S. Ouni)

DCASE'2017 - Workshop on Detection and Classification of Acoustic Scenes and Events (R. Serizel, E. Vincent)

EUSIPCO'2017 - European Signal Processing Conference (D. Jouvet, A. Liutkus)

GLU'2017 - International Workshop on Grounding Language Understanding (D. Jouvet)

GRETSI 2017 - Colloque du Groupe d'Etudes du Traitement du Signal et des Images (R. Serizel)

HSCMA'2017 - Joint Workshop on Hands-free Speech Communication and Microphone Arrays (E. Vincent)

ICASSP'2017 - IEEE International Conference on Acoustics, Speech and Signal Processing (D. Jouvet, A. Liutkus, R. Serizel, E. Vincent)

ICNLSSP'2017 - International Conference on Natural Language, Signal and Speech Processing (D. Jouvet, O. Mella, E. Vincent)

INTERSPEECH 2017 (A. Bonneau, D. Jouvet, I. Illina, Y. Laprie, S. Ouni, E. Vincent)

ISSP 2017 - International Seminar On Speech Production (Y. Laprie)

LVA/ICA'2017 - International Conference on Latent Variable Analysis and Signal Separation (A. Liutkus, E. Vincent)

PaPE 2017 - Phonetics and Phonology in Europe (A. Bonneau)

SLaTE'2017 - ISCA Workshop on Speech and Language Technology in Education (A. Bonneau, D. Jouvet)

WASPAA 2017 - IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (A. Liutkus, R. Serizel, E. Vincent)

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

ANAE Approche Neuropsychologique des Apprentissages chez l'enfant. Coordination of a special issue: N°147 : "Troubles de l'apprentissage du langage écrit et prise en charge multidisciplinaire: De la science à la salle de classe" (A. Piquard-Kipffer)

Computer Speech and Language, special issue on Multi-Microphone Speech Recognition in Everyday Environments (E. Vincent)

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

Speech Communication (D. Jouvet)

Speech Communication, special issue on Realism in Robust Speech and Language Processing (E. Vincent)

Traitement du signal (E. Vincent)

10.1.3.2. Reviewer - Reviewing Activities

Computer Speech and Language (D. Jouvet)

Computers in Biology and Medicine (R. Serizel)

IEEE Transactions on Audio, Speech and Language Processing (A. Liutkus, S. Ouni, R. Serizel)

IEEE Transactions on Emerging Topics in Computational Intelligence (R. Serizel)

IET Signal Processing (R. Serizel)

Journal of the Acoustical Society of America (B. Elie, Y. Laprie)

Jasa Express Letters (Y. Laprie, S. Ouni, R. Serizel)

Logopedics Phoniatrics Vocology (S. Ouni)

Speech Communication (V. Colotte, S. Ouni)

10.1.4. Invited Talks

Dyslexia-dysorthographia, « Dyslexie-Dysorthographie. Du repérage à la prise en charge ». Unaape, Le Chesnay, Jan 2017 (A. Piquard-Kipffer)

New paradigms in speech recognition, SIIE 2017, Feb 2017 (D. Fohr, I. Illina) [35]

Speech processing techniques for far-end spoken interactions in noisy environments, Sonos, Santa Barbara (US), March 2017 (R. Serizel)

A tutorial on probabilistic modeling for audio source separation, Kyoto University, March 2017 (A. Liutkus)

An articulatory model of the complete vocal tract from medical images, Electronic Speech Signal Processing 2017, Saarbrücken, March 2017 (Y. Laprie)

Language pathology, Séminaire "Dépistage des troubles des apprentissages" in EHESP, Rennes, March 2017 (A. Piquard-Kipffer)

Speech synthesis, Ecole Nationale d'Ingénieurs de Tunis (Tunisia), May 2017 (V. Colotte)

Speech recognition, Ecole Nationale d'Ingénieurs de Tunis (Tunisia), May 2017 (D. Jouvet)

Deep learning for distant-microphone enhancement and recognition — Expected and unexpected results, Audio Analytic, Cambridge (UK), June 2017 (E. Vincent)

Deep learning for speech and audio processing, Journées scientifiques Inria, Sophia-Antipolis, June 2017 (R. Serizel)

Deep learning for distant-microphone enhancement and recognition — Expected and unexpected results, Inria Rennes - Bretagne Atlantique (France), July 2017 (E. Vincent)

When mismatched training data outperform matched data, Erwin Schroedinger Institute Workshop on "Systematic approaches to deep learning methods for audio", Vienna (Austria), Sep 2017 (E. Vincent)

Rehaussement et reconnaissance robuste de la parole, Université Grenoble - Alpes (France), Nov 2017 (E. Vincent)

Reading predictors, "les habiletés associées à la lecture. Comment être le mieux outillé(e) pour apprendre à lire ?". Canopé, Nancy, Nov 2017 (A. Piquard-Kipffer)

Beginning of reading, "Les premiers apprentissages de la lecture". Conférences de circonscriptions 1 et 2. Longwy, Nov 2017 (A. Piquard-Kipffer)

10.1.5. Leadership within the Scientific Community

Elected chair, ISCA Special Interest Group on Robust Speech Processing (E. Vincent)

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

10.1.6. Scientific Expertise

Expertise of an ANR project proposal (D. Jouvet, Y. Laprie)

Expertise of an ERC project proposal (E. Vincent)

Expertise of the ÖAW - Austrian Academy of Sciences Fund (S. Ouni)

10.1.7. Research Administration

Vice Scientific Deputy of Inria Nancy - Grand Est from Sep 2017 (E. Vincent)

Elected Member of the board of the AM2I Scientific Pole - Université de Lorraine (E. Vincent)

Member of Comité Espace Transfert (E. Vincent)

Member of the Comipers of Inria Nancy - Grand Est until Aug 2017 (E. Vincent)

Member of the Comité de Centre of Inria Nancy - Grand Est until Aug 2017 (E. Vincent)

Vice-Chair of the Recruitment Jury for Junior Research Scientists, Inria Nancy - Grand Est (E. Vincent).

Member of the "Commission de développement technologique" (A. Bonneau)

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

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

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

DUT: I. Illina, Programming in Java, 150 hours, L1, University of Lorraine, France

DUT: I. Illina, Linux System, 65 hours, L1, University of Lorraine, France

DUT: I. Illina, Supervision of student projects and stages, 50 hours, L2, University of Lorraine, France

DUT: S. Ouni, Programming in Java, 24 hours, L1, University of Lorraine, France

DUT: S. Ouni, Web Programming, 24 hours, L1, University of Lorraine, France

DUT: S. Ouni, Graphical User Interface, 96 hours, L1, University of Lorraine, France

DUT: S. Ouni, Advanced Algorihms, 24 hours, L2, University of Lorraine, France

DUT: R. Serizel, Computer science basics, 90 hours, L1, University of Lorraine, France

DUT: R. Serizel, Introduction to office software applications, 18h, L2, University of Lorraine, France

DUT: R. Serizel, Multimedia and web applications, 20h, L1, University of Lorraine, France

DUT: R. Serizel, Digital image processing basics

Licence: V. Colotte, C2i - Certificat Informatique et Internet, 50h, L1, University of Lorraine, France

Licence: V. Colotte, System, 115h, L3, University of Lorraine, France

Licence: O. Mella, C2i - Certificat Informatique et Internet, 20h, L1, University of Lorraine, France

Licence: O. Mella, Introduction to Web Programming, 30h, L1, University of Lorraine, France

Licence: O. Mella, Computer Networking, 128h, L2-L3, University of Lorraine, France

Licence: O. Mella Supervision of student internships, 4 hours, L3, University of Lorraine, France

Licence: A. Piquard-Kipffer, Education Science, 32 hours, L1, Departement Orthophonie, University of Lorraine, France

Licence: A. Piquard-Kipffer, Learning to Read, 34 hours, L2, Departement Orthophonie, University of Lorraine, France

Licence: A. Piquard-Kipffer, Psycholinguistics, 12 hours, L2, Departement Orthophonie, University Pierre et Marie Curie-Paris, France

Licence: A. Piquard-Kipffer, Dyslexia, Dysorthographia, 22 hours, L3, Departement Orthophonie, University of Lor- raine, France

Master: A. Bonneau, Ecole d'audioprothèse (Phonetics), 16h, University of Lorraine, France

Master: A. Bonneau, Ecole d'orthophonie (Speech Manipulation with Praat), 2h, University of Lorraine, France

Master: V. Colotte, Introduction to Speech Analysis and Recognition, 18h, M1, University of Lorraine, France

Master: D. Jouvet, Modélisation sensorielle (partie reconnaissance de la parole), 12h, M2, University of Lorraine, France

Master: Y. Laprie, Master de Sciences Cognitives (Analyse, perception et reconnaissance de la parole), 30h, University of Lorraine, France

Master: O. Mella, Computer Networking, 60h, M1, University of Lorraine, France

Master: O. Mella, Introduction to Speech Analysis and Recognition, 12h, M1, University of Lorraine, France

Master: S. Ouni, Multimedia in Distributed Information Systems, 31 hours, M2, University of Lorraine, France

Master: A. Piquard-Kipffer Agnès, Dyslexia, Dysorthographia diagnosis, 4 hours, Departement Orthophonie, University of Lor- raine, France

Master: A. Piquard-Kipffer, Deafness & reading, 21 hours, Departement Orthophonie, University of Lorraine, France

Master: A. Piquard-Kipffer, French Language Didactics, 73 hours, ESPE, University of Lorraine, France

Master: A. Piquard-Kipffer, Special educational needs, 18 hours, ESPE, University of Lorraine, France

Master: A. Piquard-Kipffer, Psychology, 6 hours, University of Lorraine, France

Continuous training : O. Mella, Computer science courses for secondary school teachers (ISN courses), 10h, ESPE, University of Lorraine, France

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

Doctorat: A. Piquard-Kipffer, Language Pathology, 20 hours, University of Lorraine, France

Other: V. Colotte, Responsible for "Certificat Informatique et Internet" for the University of Lorraine, France (50000 students, 30 departments)

Other: S. Ouni, Responsible of Année Spéciale DUT, University of Lorraine, France

10.2.2. Supervision

PhD : Quan Nguyen, "Mapping of a sound environment by a mobile robot", University of Lorraine, November 3, 2017, Francis Colas and Emmanuel Vincent.

PhD : Aditya Nugraha, "Deep neural networks for source separation and noise-robust speech recognition", December 5, 2017, Antoine Liutkus and Emmanuel Vincent.

PhD in progress: Ken Deguernel, "Apprentissage de structures musicales en situation d'improvisation", March 2015, Emmanuel Vincent and Gérard Assayag (Ircam).

PhD in progress: Amal Houidhek, "Élaboration et analyse d'une base de parole arabe pour la synthèse vocale", December 2015, cotutelle, Denis Jouvet and Vincent Colotte (France) and Zied Mnasri (Tunisia).

PhD in progress: Imène Zangar, "Amélioration de la qualité de synthèse vocale par HMM pour la parole arabe", December 2015, codirection, Denis Jouvet and Vincent Colotte (France) and Zied Mnasri (Tunisia).

PhD in progress: Amine Menacer, "Traduction automatique de vidéos", May 2016, Kamel Smaïli and Denis Jouvet.

PhD in progress: Mathieu Fontaine, "Processus alpha-stable pour le traitement du signal", May 2016, Antoine Liutkus and Roland Badeau (Télécom ParisTech).

PhD in progress: Anastasiia Tsukanova, "Coarticulation modeling in articulatory synthesis", May 2016, Yves Laprie.

PhD in progress: Nathan Libermann, "Deep learning for musical structure analysis and generation", October 2016, Frédéric Bimbot (IRISA) and Emmanuel Vincent.

PhD in progress: Lauréline Perotin, "Séparation aveugle de sources sonores en milieu réverbérant", November 2016, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin (Orange).

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

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

PhD in progress: Guillaume Carbajal, "Apprentissage profond bout-en-bout pour le rehaussement de la parole", March 2017, Romain Serizel, Emmanuel Vincent, and AÉric Humbert (Invoxia).

PhD in progress: Sunit Sivasankaran, “Exploiting contextual information in the speech processing chain”, July 2017, Dominique Fohr and Emmanuel Vincent.

PhD in progress: Ioannis Douros, “Combining cineMRI and static MRI to analyze speech production”, July 2017, Pierre-André Vuissoz (IADI) and Yves Laprie.

PhD in progress: Lou Lee, “Du lexique au discours: les particules discursives en français”, October 2017, Yvon Keromnes and Mathilde Dargnat (ATILF) and Denis Jouvet.

10.2.3. Participation in HDR and PhD juries

Participation in Habilitation Jury for Damien Lolive (Université de Rennes 1, November 2017), Y. Laprie, reviewer.

Participation in PhD thesis Jury for Dionyssos Kounades-Bastian (Université Grenoble - Alpes, February 2017), E. Vincent, reviewer.

Participation in PhD thesis Jury for Emilio Molina Martínez (University of Málaga, Spain, March 2017), E. Vincent, reviewer.

Participation in PhD thesis Jury for Simon Durand (Télécom ParisTech, May 2017), E. Vincent, reviewer.

Participation in PhD thesis Jury for Mathieu Baqué (Université du Maine, June 2017), E. Vincent, reviewer.

Participation in PhD thesis Jury for Waad Ben Kheder (Université d’Avignon et des Pays du Vaucluse, July 2017), R. Serizel.

Participation in PhD thesis Jury for Waad Ben Kheder (Université d’Avignon et des Pays du Vaucluse, July 2017), D. Jouvet, reviewer.

Participation in PhD thesis Jury for Gabriel Bustamante (Université Fédérale Toulouse Midi-Pyrénées, September 2017), E. Vincent, reviewer.

Participation in PhD thesis Jury for Fangchen Feng (Université Paris - Sud, September 2017), E. Vincent, reviewer.

Participation in PhD thesis Jury for Gregory Gelly (Université Paris-Saclay, September 2017), D. Jouvet, reviewer.

Participation in PhD thesis Jury for Andrew Szabados (Université de Grenoble Alpes, November 2017), Y. Laprie, reviewer.

Participation in PhD thesis Jury for Natalia Tomashenko (Université du Mans, December 2017), D. Jouvet, reviewer.

Participation in PhD thesis Jury for Daniele Battaglino (EURECOM - Télécom ParisTech, December 2017), E. Vincent, reviewer.

Participation in PhD thesis Jury for Mohamed Bouaziz (Université d’Avignon, December 2017), I. Illina, reviewer.

Participation in PhD thesis Jury for Inaki Frenandez (Université de lorraine, December 2017), I. Illina.

10.2.4. Participation in other juries

Participation in CAFIPEMPF Jury - Master Learning Facilitator, Académie de Nancy-Metz & Université de Lorraine, April, May 2017, A. Piquard-Kipffer

Participation in the Competitive Entrance Examination into Speech-Language Pathology Departement, Université de Lorraine, June 2017, A. Piquard-Kipffer.

10.3. Popularization

Interview for “Quand les machines apprennent à parler l’humain — Le Zoom de la Rédaction”, *France Inter*, February 2, 2017 (E. Vincent)

Interview for “Les prochains défis de la reconnaissance vocale”, *Le Figaro*, April 11, 2017 (E. Vincent)

Interview for “2017 : Alexa, la voix d’Amazon”, *Les Échos*, August 31, 2017 (E. Vincent)

Interactions vocales, Grand-Est Numérique, September 2017 (R. Serizel)

Demonstrations at Fête de la Science, University of Lorraine, October 13, 2017 (G. Carbajal, L. Perotin, R. Serizel, E. Vincent)

Demonstrations at the Rencontre Inria Industrie “Les données et leurs applications”, October 18, 2017 (E. Vincent)

Interview for “Les assistants vocaux vont bousculer la radio”, *Le Monde*, October 19, 2017 (E. Vincent)

Interview for *France Culture*, November 2017 (E. Vincent)

Talk at “Fresh from the Labs”, Station F, Paris (France), November 30, 2017 (E. Vincent)

Demonstration at Journée des métiers, Collège Péguy, le Chesnay, March 2017 (A. Piquard-Kipffer).

Demonstration of Dynalips at “50 ans Inria”, Paris, November, 7-8, 2017 (S. Ouni)

11. Bibliography

Major publications by the team in recent years

[1] F. BAHJA, J. DI MARTINO, E. H. IBN ELHAJ, D. ABOUTAJDINE. *An overview of the CATE algorithms for real-time pitch determination*, in "Signal, Image and Video Processing", 2013 [*DOI :* 10.1007/s11760-013-0488-4], https://hal.inria.fr/hal-00831660

[2] J. BARKER, E. VINCENT, N. MA, H. CHRISTENSEN, P. GREEN. *The PASCAL CHiME Speech Separation and Recognition Challenge*, in "Computer Speech and Language", February 2013, vol. 27, n^{o} 3, pp. 621-633 [*DOI :* 10.1016/J.CSL.2012.10.004], https://hal.inria.fr/hal-00743529

[3] A. BONNEAU, D. FOHR, I. ILLINA, D. JOUVET, O. MELLA, L. MESBAHI, L. OROSANU. *Gestion d’erreurs pour la fiabilisation des retours automatiques en apprentissage de la prosodie d’une langue seconde*, in "Traitement Automatique des Langues", 2013, vol. 53, n^{o} 3, https://hal.inria.fr/hal-00834278

[4] D. JOUVET, D. FOHR. *Combining Forward-based and Backward-based Decoders for Improved Speech Recognition Performance*, in "InterSpeech - 14th Annual Conference of the International Speech Communication Association - 2013", Lyon, France, August 2013, https://hal.inria.fr/hal-00834282

[5] A. OZEROV, M. LAGRANGE, E. VINCENT. *Uncertainty-based learning of acoustic models from noisy data*, in "Computer Speech and Language", February 2013, vol. 27, n^{o} 3, pp. 874-894 [*DOI :* 10.1016/J.CSL.2012.07.002], https://hal.inria.fr/hal-00717992

[6] A. OZEROV, E. VINCENT, F. BIMBOT. *A General Flexible Framework for the Handling of Prior Information in Audio Source Separation*, in "IEEE Transactions on Audio, Speech and Language Processing", May 2012, vol. 20, n^{o} 4, pp. 1118 - 1133, 16, https://hal.archives-ouvertes.fr/hal-00626962

[7] A. PIQUARD-KIPFFER, B. CHRISTIAN. *Je peux voir les mots que tu dis ! Histoire d’un projet*, in "13ème édition du Festival du film de chercheur CNRS 2012", Nancy, France, June 2012, https://hal.inria.fr/hal-01263907

[8] A. PIQUARD-KIPFFER, L. SPRENGER-CHAROLLES. *Predicting reading level at the end of Grade 2 from skills assessed in kindergarten: contribution of phonemic discrimination (Follow-up of 85 French-speaking children from 4 to 8 years old)*, in "Topics in Cognitive Psychology", 2013, https://hal.inria.fr/hal-00833951

Publications of the year

Articles in International Peer-Reviewed Journals

[9] J. BARKER, R. MARXER, E. VINCENT, S. WATANABE. *Multi-microphone speech recognition in everyday environments*, in "Computer Speech and Language", July 2017, vol. 46, pp. 386-387 [*DOI :* 10.1016/J.CSL.2017.02.007], https://hal.inria.fr/hal-01483469

[10] J. BARKER, R. MARXER, E. VINCENT, S. WATANABE. *The third 'CHIME' speech separation and recognition challenge: Analysis and outcomes*, in "Computer Speech and Language", July 2017, vol. 46, pp. 605-626, https://hal.inria.fr/hal-01382108

[11] V. BISOT, R. SERIZEL, S. ESSID, G. RICHARD. *Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification*, in "IEEE/ACM Transactions on Audio, Speech and Language Processing", May 2017, vol. 25, n° 6, pp. 1216 - 1229, https://hal.archives-ouvertes.fr/hal-01362864

[12] B. ELIE, Y. LAPRIE. *Acoustic impact of the gradual glottal abduction on the production of fricatives: A numerical study*, in "Journal of the Acoustical Society of America", September 2017, vol. 142, n° 3, pp. 1303-1317 [*DOI :* 10.1121/1.5000232], https://hal.archives-ouvertes.fr/hal-01423206

[13] B. ELIE, Y. LAPRIE. *Simulating alveolar trills using a two-mass model of the tongue tip*, in "Journal of the Acoustical Society of America", 2017, vol. 142, n° 5, forthcoming, https://hal.archives-ouvertes.fr/hal-01525882

[14] S. GANNOT, E. VINCENT, S. MARKOVICH-GOLAN, A. OZEROV. *A consolidated perspective on multimicrophone speech enhancement and source separation*, in "IEEE/ACM Transactions on Audio, Speech and Language Processing", April 2017, vol. 25, n° 4, pp. 692–730, Added equation (108), https://hal.inria.fr/hal-01414179

[15] C. LECLERC, A. PIQUARD-KIPFFER, C. ROSIN, M. WERNET. *Inclusive education: a particular system of teaching with dyslexic and dysphasicchildren, in a specialized school*, in "ANAE - Approche Neuropsychologique des Apprentissages Chez L'enfant", October 2017, https://hal.inria.fr/hal-01635918

[16] T. LÉONOVA, A. PIQUARD-KIPFFER, A. JUMAGELDINOV, M. ROBERT, M. BEREBIN. *Inclusive education for students with specific language disorders: What schoolingaccording to country and language*, in "ANAE - Approche Neuropsychologique des Apprentissages Chez L'enfant", October 2017, https://hal.inria.fr/hal-01647486

[17] K. NATHWANI, E. VINCENT, I. ILLINA. *DNN Uncertainty Propagation using GMM-Derived Uncertainty Features for Noise Robust ASR*, in "IEEE Signal Processing Letters", January 2018, https://hal.inria.fr/hal-01680658

[18] A. PIQUARD-KIPFFER, T. LÉONOVA. *Scolarité et handicap : parcours de 170 jeunes dysphasiques ou dyslexiques- dysorthographiques âgés de 6 à 20 ans*, in "ANAE - Approche Neuropsychologique des Apprentissages Chez L'enfant", October 2017, https://hal.inria.fr/hal-01645096

[19] G. SARGENT, F. BIMBOT, E. VINCENT. *Estimating the structural segmentation of popular music pieces under regularity constraints*, in "IEEE/ACM Transactions on Audio, Speech, and Language Processing", 2017, https://hal.inria.fr/hal-01403210

[20] I. A. SHEIKH, D. FOHR, I. ILLINA, G. LINARES. *Modelling Semantic Context of OOV Words in Large Vocabulary Continuous Speech Recognition*, in "IEEE/ACM Transactions on Audio, Speech and Language Processing", January 2017, vol. 25, n^{o} 3, pp. 598 - 610 [*DOI :* 10.1109/TASLP.2017.2651361], https://hal.inria.fr/hal-01461617

[21] S. SIVASANKARAN, E. VINCENT, I. ILLINA. *A combined evaluation of established and new approaches for speech recognition in varied reverberation conditions*, in "Computer Speech and Language", July 2017, vol. 46, pp. 444-460, https://hal.inria.fr/hal-01461382

[22] E. VINCENT, S. WATANABE, A. A. NUGRAHA, J. BARKER, R. MARXER. *An analysis of environment, microphone and data simulation mismatches in robust speech recognition*, in "Computer Speech and Language", July 2017, vol. 46, pp. 535-557, https://hal.inria.fr/hal-01399180

[23] Z. WANG, E. VINCENT, R. SERIZEL, Y. YAN. *Rank-1 Constrained Multichannel Wiener Filter for Speech Recognition in Noisy Environments*, in "Computer Speech and Language", 2017, forthcoming, https://hal.inria.fr/hal-01634449

Invited Conferences

[24] A. MESAROS, T. HEITTOLA, A. DIMENT, B. ELIZALDE, A. SHAH, E. VINCENT, B. RAJ, T. VIRTANEN. *DCASE 2017 Challenge setup: Tasks, datasets and baseline system*, in "DCASE 2017 - Workshop on Detection and Classification of Acoustic Scenes and Events", Munich, Germany, November 2017, https://hal.inria.fr/hal-01627981

[25] E. VINCENT. *When mismatched training data outperform matched data*, in "Systematic approaches to deep learning methods for audio", Vienna, Austria, September 2017, https://hal.inria.fr/hal-01588876

International Conferences with Proceedings

[26] I. BADA, J. KARSTEN, D. FOHR, I. ILLINA. *Data Selection in the Framework of Automatic Speech Recognition*, in "ICNLSSP 2017 - International conference on natural language, signal and speech processing 2017", Casablanca, Morocco, Proceedings of ICNLSSP 2017, December 2017, pp. 1-5, https://hal.archives-ouvertes.fr/hal-01629340

[27] V. BISOT, R. SERIZEL, S. ESSID, G. RICHARD. *Leveraging deep neural networks with nonnegative representations for improved environmental sound classification*, in "IEEE International Workshop on Machine Learning for Signal Processing MLSP", Tokyo, Japan, September 2017, https://hal.archives-ouvertes.fr/hal-01576857

[28] V. BISOT, R. SERIZEL, S. ESSID, G. RICHARD. *Nonnegative Feature Learning Methods for Acoustic Scene Classification*, in "DCASE 2017 - Workshop on Detection and Classification of Acoustic Scenes and Events", Munich, Germany, November 2017, https://hal.inria.fr/hal-01636627

[29] B. DENG, D. JOUVET, Y. LAPRIE, I. STEINER, A. SINI. *Towards Confidence Measures on Fundamental Frequency Estimations*, in "IEEE International Conference on Acoustics, Speech and Signal Processing", New Orleans, United States, March 2017, https://hal.inria.fr/hal-01493168

[30] D. DI CARLO, K. DÉGUERNEL, A. LIUTKUS. *Gaussian framework for interference reduction in live recordings*, in "AES International Conference on Semantic Audio", Erlangen, Germany, June 2017, https://hal.inria.fr/hal-01515971

[31] B. DUMORTIER, E. VINCENT, M. DEACONU. *Recursive Bayesian estimation of the acoustic noise emitted by wind farms*, in "2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)", New Orleans, United States, March 2017, https://hal.inria.fr/hal-01428962

[32] K. DÉGUERNEL, J. NIKA, E. VINCENT, G. ASSAYAG. *Generating Equivalent Chord Progressions to Enrich Guided Improvisation : Application to Rhythm Changes*, in "SMC 2017 - 14th Sound and Music Computing Conference", Espoo, Finland, July 2017, 8 p. , https://hal.inria.fr/hal-01528559

[33] B. ELIE, Y. LAPRIE. *Glottal Opening and Strategies of Production of Fricatives*, in "Interspeech 2017", Stockholm, Sweden, August 2017, pp. 206-209 [*DOI :* 10.21437/INTERSPEECH.2017-1039], https://hal.archives-ouvertes.fr/hal-01574839

[34] D. FITZGERALD, Z. RAFII, A. LIUTKUS. *User Assisted Separation of Repeating Patterns in Time and Frequency using Magnitude Projections*, in "42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)", New Orleans, United States, March 2017, https://hal.inria.fr/hal-01515956

[35] D. FOHR, O. MELLA, I. ILLINA. *New Paradigm in Speech Recognition: Deep Neural Networks*, in "IEEE International Conference on Information Systems and Economic Intelligence", Marrakech, Morocco, April 2017, https://hal.archives-ouvertes.fr/hal-01484447

[36] M. FONTAINE, A. LIUTKUS, L. GIRIN, R. BADEAU. *Explaining the Parameterized Wiener Filter with Alpha-Stable Processes*, in "IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)", New Paltz, New York, United States, Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), October 2017, https://hal.archives-ouvertes.fr/hal-01548508

[37] M. FONTAINE, C. VANWYNSBERGHE, A. LIUTKUS, R. BADEAU. *Scalable Source Localization with Multichannel Alpha-Stable Distributions*, in "25th European Signal Processing Conference (EUSIPCO)", Kos, Greece, Proc. of 25th European Signal Processing Conference (EUSIPCO), August 2017, pp. 11-15, https://hal.archives-ouvertes.fr/hal-01531252

[38] M. FONTAINE, C. VANWYNSBERGHE, A. LIUTKUS, R. BADEAU. *Sketching for nearfield acoustic imaging of heavy-tailed sources*, in "13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017)", Grenoble, France, Latent Variable Analysis and Signal Separation 13th International Conference, LVA/ICA 2017, Grenoble, France, February 21-23, 2017, Proceedings, February 2017, vol. 10169, pp. 80-88 [*DOI :* 10.1007/978-3-319-53547-0_8], https://hal.archives-ouvertes.fr/hal-01401988

[39] *Best Paper*
A. HOUIDHEK, V. COLOTTE, Z. MNASRI, D. JOUVET, I. ZANGAR. *Statistical modelling of speech units in HMM-based speech synthesis for Arabic*, in "LTC 2017 - 8th Language & Technology Conference", Poznań, Poland, November 2017, pp. 1-5, https://hal.inria.fr/hal-01649034.

[40] I. ILLINA, D. FOHR. *Out-of-Vocabulary Word Probability Estimation using RNN Language Model*, in "8th Language & Technology Conference", Poznan, Poland, proceedings of LTC 2017, November 2017, https://hal.archives-ouvertes.fr/hal-01623784

[41] D. JOUVET, K. BARTKOVA, M. DARGNAT, L. LEE. *Analysis and Automatic Classification of Some Discourse Particles on a Large Set of French Spoken Corpora*, in "SLSP'2017, 5th International Conference on Statistical Language and Speech Processing", Le Mans, France, October 2017, https://hal.inria.fr/hal-01585567

[42] *Best Paper*
D. JOUVET, D. LANGLOIS, M. A. MENACER, D. FOHR, O. MELLA, K. SMAÏLI. *About vocabulary adaptation for automatic speech recognition of video data*, in "ICNLSSP'2017 - International Conference on Natural Language, Signal and Speech Processing", Casablanca, Morocco, December 2017, pp. 1-5, https://hal.inria.fr/hal-01649057.

[43] D. JOUVET, Y. LAPRIE. *Performance Analysis of Several Pitch Detection Algorithms on Simulated and Real Noisy Speech Data*, in "EUSIPCO'2017, 25th European Signal Processing Conference", Kos, Greece, August 2017, https://hal.inria.fr/hal-01585554

[44] Y. LAPRIE, B. ELIE, P.-A. VUISSOZ, A. TSUKANOVA. *Articulatory model of the epiglottis*, in "The 11th International Seminar on Speech Production", Tianjin, China, October 2017, https://hal.inria.fr/hal-01643227

[45] S. LEGLAIVE, U. SIMSEKLI, A. LIUTKUS, R. BADEAU, G. RICHARD. *Alpha-Stable Multichannel Audio Source Separation*, in "42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)", New Orleans, United States, Proc. 42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, March 2017, https://hal.archives-ouvertes.fr/hal-01416366

[46] A. LIUTKUS, F.-R. STÖTER, Z. RAFII, D. KITAMURA, B. RIVET, N. ITO, N. ONO, J. FONTECAVE. *The 2016 Signal Separation Evaluation Campaign*, in "13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017)", Grenoble, France, P. TICHAVSKÝ, M. BABAIE-ZADEH, O. J. MICHEL, N. THIRION-MOREAU (editors), LNCS - Lecture Notes in Computer Science, Springer, February 2017, vol. 10169, pp. 323 - 332 [*DOI :* 10.1007/978-3-319-53547-0_31], https://hal.inria.fr/hal-01472932

[47] A. LIUTKUS, K. YOSHII. *A diagonal plus low-rank covariance model for computationally efficient source separation*, in "IEEE international workshop on machine learning for signal processing (MLSP)", Tokyo, Japan, September 2017, https://hal.inria.fr/hal-01580733

[48] P. MAGRON, R. BADEAU, A. LIUTKUS. *Lévy NMF : un modèle robuste de séparation de sources non-négatives*, in "Colloque GRETSI", Juan-Les-Pins, France, Actes du XXVIème Colloque GRETSI, September 2017, https://hal.archives-ouvertes.fr/hal-01540484

[49] P. MAGRON, R. BADEAU, A. LIUTKUS. *Lévy NMF for Robust Nonnegative Source Separation*, in "IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2017)", New Paltz, NY, United States, IEEE, October 2017, https://hal.archives-ouvertes.fr/hal-01548488

[50] M. A. MENACER, D. LANGLOIS, O. MELLA, D. FOHR, D. JOUVET, K. SMAÏLI. *Is statistical machine translation approach dead?*, in "ICNLSSP 2017 - International Conference on Natural Language, Signal and Speech Processing", Casablanca, Morocco, ISGA, December 2017, pp. 1-5, https://hal.inria.fr/hal-01660016

[51] M. A. MENACER, O. MELLA, D. FOHR, D. JOUVET, D. LANGLOIS, K. SMAÏLI. *An enhanced automatic speech recognition system for Arabic*, in "The third Arabic Natural Language Processing Workshop - EACL 2017", Valencia, Spain, Arabic Natural Language Processing Workshop - EACL 2017, April 2017, https://hal.archives-ouvertes.fr/hal-01531588

[52] M. A. MENACER, O. MELLA, D. FOHR, D. JOUVET, D. LANGLOIS, K. SMAÏLI. *Development of the Arabic Loria Automatic Speech Recognition system (ALASR) and its evaluation for Algerian dialect*, in "ACLing 2017 - 3rd International Conference on Arabic Computational Linguistics", Dubai, United Arab Emirates, November 2017, pp. 1-8, https://hal.archives-ouvertes.fr/hal-01583842

[53] K. NATHWANI, J. A. MORALES-CORDOVILLA, S. SIVASANKARAN, I. ILLINA, E. VINCENT. *An extended experimental investigation of DNN uncertainty propagation for noise robust ASR*, in "5th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA 2017)", San Francisco, United States, March 2017, https://hal.inria.fr/hal-01446441

[54] K. NATHWANI, E. VINCENT, I. ILLINA. *Consistent DNN Uncertainty Training and Decoding for Robust ASR*, in "2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)", Okinawa, Japan, December 2017, https://hal.inria.fr/hal-01585956

[55] Q. V. NGUYEN, F. COLAS, E. VINCENT, F. CHARPILLET. *Long-term robot motion planning for active sound source localization with Monte Carlo tree search*, in "HSCMA 2017 - Hands-free Speech Communication and Microphone Arrays", San Francisco, United States, March 2017, https://hal.archives-ouvertes.fr/hal-01447787

[56] J. NIKA, K. DÉGUERNEL, A. CHEMLA–ROMEU-SANTOS, E. VINCENT, G. ASSAYAG. *DYCI2 agents: merging the "free", "reactive", and "scenario-based" music generation paradigms*, in "International Computer Music Conference", Shangai, China, October 2017, https://hal.archives-ouvertes.fr/hal-01583089

[57] S. OUNI, S. DAHMANI, V. COLOTTE. *On the quality of an expressive audiovisual corpus: a case study of acted speech*, in "The 14th International Conference on Auditory-Visual Speech Processing", Stockholm, Sweden, S. OUNI, C. DAVIS, A. JESSE, J. BESKOW (editors), KTH, August 2017, Proceedings on line: http://avsp2017.loria.fr/proceedings/, https://hal.inria.fr/hal-01596614

[58] F. PISHDADIAN, B. PARDO, A. LIUTKUS. *A multi-resolution approach to common fate-based audio separation*, in "42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)", New Orleans, United States, March 2017, https://hal.inria.fr/hal-01515951

[59] C. ROHLFING, J. E. COHEN, A. LIUTKUS. *Very Low Bitrate Spatial Audio Coding with Dimensionality Reduction*, in "42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)", New Orleans, United States, March 2017, https://hal.inria.fr/hal-01515954

[60] C. ROHLFING, A. LIUTKUS, J. M. BECKER. *Quantization-aware Parameter Estimation for Audio Upmixing*, in "42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)", New Orleans, United States, March 2017, https://hal.inria.fr/hal-01515955

[61] R. SERIZEL, V. BISOT, S. ESSID, G. RICHARD. *Supervised Group Nonnegative Matrix Factorisation With Similarity Constraints And Applications To Speaker Identification*, in "IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)", New Orleans, United States, March 2017, https://hal.inria.fr/hal-01484744

[62] I. SHEIKH, D. FOHR, I. ILLINA. *Topic segmentation in ASR transcripts using bidirectional rnns for change detection*, in "ASRU 2017 - IEEE Automatic Speech Recognition and Understanding Workshop", Okinawa, Japan, proceedings of IEEE ASRU 2017, December 2017, https://hal.archives-ouvertes.fr/hal-01599682

[63] I. A. SHEIKH, I. ILLINA, D. FOHR. *Segmentation and Classification of Opinions with Recurrent Neural Networks*, in "IEEE Information Systems and Economic Intelligence", Al Hoceima, Morocco, proceedings of IEEE SIIE, May 2017, https://hal.inria.fr/hal-01491182

[64] S. SIVASANKARAN, E. VINCENT, I. ILLINA. *Discriminative importance weighting of augmented training data for acoustic model training*, in "42th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2017)", New Orleans, United States, March 2017, Added missing sign in equations (2) and (3) + explanation about iteration 1 in Fig. 1, https://hal.inria.fr/hal-01415759

[65] A. TSUKANOVA, B. ELIE, Y. LAPRIE. *Articulatory Speech Synthesis from Static Context-Aware Articulatory Targets*, in "ISSP 2017 - 11th International Seminar on Speech Production", Tianjin, China, October 2017, https://hal.archives-ouvertes.fr/hal-01643487

National Conferences with Proceedings

[66] K. BARTKOVA, M. DARGNAT, D. JOUVET, L. LEE. *Annotation of discourse particles in French over a large variety of speech corpora*, in "ACor4French - Les corpus annotés du français, TALN'2017 - Traitement Automatique des Langues Naturelles", Orléans, France, June 2017, https://hal.inria.fr/hal-01585540

Conferences without Proceedings

[67] A. BONNEAU. *Acoustic correlates of L2 prosodic boundaries by German learners of French*, in "SLaP3 2017 - 3rd Workshop on Second Language Prosody", Bangor , United Kingdom, November 2017, 1 p. , https://hal.inria.fr/hal-01639515

[68] T. LÉONOVA, A. DE SAINT-MARTIN, R. NABBOUT, S. AUVIN, M. ROBERT, S. CAHAREL, N. COQUÉ, A. PIQUARD-KIPFFER. *L'anxiété et les symptômes dépressifs chez les parents d'enfants atteints de syndrome de Dravet*, in "58 ème Congrès de la societe francaise de psychologie", Nice, France, August 2017, https://hal.inria.fr/hal-01645104

[69] T. LÉONOVA, D. SARDIN, A. GOSSE, M. ROBERT, A. PIQUARD-KIPFFER, P. CLAUDON, S. CLAUDEL, S. CAHAREL. *Etre parent d'enfant atteint des troubles du spectre de l'autisme : Le stress parental à travers l'analyse interprétative phénoménologique*, in "14ème congrès international de recherche sur le handicap", Genève, Switzerland, September 2017, https://hal.inria.fr/hal-01645101

Scientific Books (or Scientific Book chapters)

[70] J. BARKER, R. MARXER, E. VINCENT, S. WATANABE. *The CHiME challenges: Robust speech recognition in everyday environments*, in "New era for robust speech recognition - Exploiting deep learning", Springer, November 2017, pp. 327-344, https://hal.inria.fr/hal-01383263

[71] S. ESSID, S. PAREKH, N. Q. K. DUONG, R. SERIZEL, A. OZEROV, F. ANTONACCI, A. SARTI. *Multiview approaches to event detection and scene analysis*, in "Computational Analysis of Sound Scenes and Events", T. VIRTANEN, M. D. PLUMBLEY, D. ELLIS (editors), Springer, 2017, pp. 243-276 [*DOI :* 10.1007/978-3-319-63450-0_9], https://hal.archives-ouvertes.fr/hal-01620341

[72] C. FÉVOTTE, E. VINCENT, A. OZEROV. *Single-channel audio source separation with NMF: divergences, constraints and algorithms*, in "Audio Source Separation", Springer, 2017, forthcoming, https://hal.inria.fr/hal-01631185

[73] A. A. NUGRAHA, A. LIUTKUS, E. VINCENT. *Deep neural network based multichannel audio source separation*, in "Audio Source Separation", Springer, 2017, forthcoming, https://hal.inria.fr/hal-01633858

[74] A. OZEROV, C. FÉVOTTE, E. VINCENT. *An introduction to multichannel NMF for audio source separation*, in "Audio Source Separation", Springer, 2017, forthcoming, https://hal.inria.fr/hal-01631187

[75] R. SERIZEL, V. BISOT, S. ESSID, G. RICHARD. *Acoustic Features for Environmental Sound Analysis*, in "Computational Analysis of Sound Scenes and Events", T. VIRTANEN, M. D. PLUMBLEY, D. ELLIS (editors), Springer, 2017, pp. 71-101 [*DOI :* 10.1007/978-3-319-63450-0_4], https://hal.archives-ouvertes.fr/hal-01575619

Books or Proceedings Editing

[76] S. OUNI, C. DAVIS, A. JESSE, J. BESKOW (editors). *The proceedings of the 14th International Conference on Auditory-Visual Speech Processing*, August 2017, https://hal.inria.fr/hal-01596625

[77] J. TROUVAIN, F. ZIMMERER, B. MÖBIUS, M. GOSY, A. BONNEAU (editors). *Segmental, prosodic and fluency features in phonetic learner corporaSpecial issue of the International Journal of Learner Corpus Research 3:2*, Segmental, prosodic and fluency features in phonetic learner corpora, John Benjamins Publishing Company, December 2017, vol. 3, n^{o} 2, 176 p. [*DOI :* 10.1075/IJLCR.3.2], https://hal.inria.fr/hal-01670975

Research Reports

[78] M. CADOT, A. LELU, M. ZITT. *Benchmarking 17 clustering methods*, LORIA, June 2017, https://hal.archives-ouvertes.fr/hal-01532894

Scientific Popularization

[79] K. DÉGUERNEL, N. LIBERMANN, E. VINCENT. *La musique comme une langue*, March 2017, Commission française pour l'enseignement des mathématiques, livret "Mathématiques et langages - Panorama du thème", https://hal.inria.fr/hal-01485209

Patents and standards

[80] G. CARBAJAL, R. SERIZEL, E. VINCENT, E. HUMBERT. *Procédé de suppression d'écho résiduel dans un signal acoustique*, October 2017, n^{o} 1760200, https://hal.inria.fr/hal-01638050

References in notes

[81] A. PIQUARD-KIPFFER. *Storytelling with a digital album that use an avatar as narrator*, in "XVIèmes rencontres internationales en orthophonie - Orthophonie et technologies innovantes ", PARIS, France, XVIèmes rencontres internationales en orthophonie - Orthophonie et technologies innovantes, December 2016, https://hal.inria.fr/hal-01403204