IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2016

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:

- 3.1.4. - Uncertain data
- 3.4.6. - Neural networks
- 3.4.8. - Deep learning
- 5.1.7. - Multimodal interfaces
- 5.7.2. - Music
- 5.7.3. - Speech
- 5.7.4. - Analysis
- 5.7.5. - Synthesis
- 5.8. - Natural language processing
- 5.9.1. - Sampling, acquisition
- 5.9.2. - Estimation, modeling
- 5.9.3. - Reconstruction, enhancement
- 5.9.5. - Sparsity-aware processing
- 5.10.2. - Perception
- 5.11.2. - Home/building control and interaction
- 6.2.4. - Statistical methods
- 6.3.1. - Inverse problems
- 6.3.5. - Uncertainty Quantification
- 8.2. - Machine learning
- 8.3. - Signal analysis

Other Research Topics and Application Domains:

- 4.3.3. - Wind energy
- 8.1.2. - Sensor networks for smart buildings
- 8.4. - Security and personal assistance
- 9.1.1. - E-learning, MOOC
- 9.2.1. - Music, sound
- 9.2.2. - Cinema, Television
- 9.4.1. - Computer science
- 9.4.2. - Mathematics
- 9.4.5. - Data science
- 9.5.8. - Linguistics
- 9.5.10. - Digital humanities

1. Members

Research Scientists

Denis Jouvet [Team leader, Inria, Senior Researcher, HDR]

Anne Bonneau [CNRS, Researcher]
Dominique Fohr [CNRS, Researcher]
Yves Laprie [CNRS, Senior Researcher, HDR]
Antoine Liutkus [Inria, Researcher]
Emmanuel Vincent [Inria, Senior Researcher, HDR]

Faculty Members

Vincent Colotte [Univ. Lorraine, Associate Professor]
Irène Illina [Univ. Lorraine, Associate Professor, HDR]
Odile Mella [Univ. Lorraine, Associate Professor]
Slim Ouni [Univ. Lorraine, Associate Professor, HDR]
Agnès Piquard-Kipffer [ESPE, Univ. Lorraine, Associate Professor]
Romain Serizel [Univ. Lorraine, Associate Professor, from Sep 2016]

Engineers

Ismaël Bada [CNRS, from Aug 2016]
Sara Dahmani [Inria]
Valérian Girard [Univ. Lorraine, until Jun 2016; then CNRS]
Karan Nathwani [Inria, from Jun 2016]
Aghilas Sini [CNRS, until Nov 2016]
Sunit Sivasankaran [Inria]

PhD Students

Ken Deguernel [Inria]
Baldwin Dumortier [Inria]
Mathieu Fontaine [Inria, from May 2016]
Amal Houidhek [École Nationale d'Ingénieurs de Tunis, Tunisie]
Yang Liu [Univ. Lorraine, from Oct 2016]
Aditya Nugraha [Inria]
Luiza Orosanu [Inria, until Feb 2016]
Imran Sheikh [Univ. Lorraine]
Anastasiia Tsukanova [Univ. Lorraine, from May 2016]
Imene Zangar [École Nationale d'Ingénieurs de Tunis, Tunisie]

Post-Doctoral Fellows

Mohamed Bouallegue [Univ. Lorraine, until Sep 2016]
Benjamin Elie [CNRS]
Sucheta Ghosh [CNRS]
Juan Andres Morales Cordovilla [Inria, until Mar 2016]

Visiting Scientists

Sebastian Gonzalez Mora [Univ. Chile, Jan 2016]
Benjamin Martinez Elizalde [Carnegie Mellon University, from May 2016 until Aug 2016]
Dayana Ribas [CENATAV, from Sep 2016 until Dec 2016]
Ziteng Wang [Institute of Acoustics, Chinese Academy of Sciences, from Sep 2016]

Administrative Assistants

Antoinette Courrier [CNRS]
Sylvie Musilli [Univ. Lorraine]
Hélène Zganic [Inria]

Others

Lucas Antonelli [Inria, Intern, from Apr 2016 until Jun 2016]
Imen Ben Othmane [ESTI, Univ. de Carthage, Tunisia, Visiting PhD Student, until Jun 2016]
Clement Bordes [Inria, Intern, from Jun 2016 until Sep 2016]
Freha Boumazza [Uuiv. Hassiba Benbouali de Chlef, Algérie, Visiting PhD Student, Mar 2016]
Anna Currey [Univ. Lorraine, Intern, from Feb 2016 until Jun 2016]

Siddharth Dalmia [Univ. Lorraine, Intern, until Jul 2016]
Narjes Daoud [ESTI, Univ. de Carthage, Tunisia, Intern, from Mar 2016 until Aug 2016]
Boyuan Deng [Univ. Lorraine, Intern, from Feb 2016 until Jul 2016]
Diego Di Carlo [Inria, Intern, from Nov 2016]
Bertrand Muller [Inria, Intern, from Apr 2016 until Jun 2016]

2. Overall Objectives

2.1. Overall Objectives

MULTISPEECH is a joint project between Inria, CNRS and University of Lorraine, hosted in the LORIA laboratory (UMR 7503). 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 [7], 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

We ranked 1st ex aqueo for the "Professionally produced music recordings" task of the 2016 Signal Separation Evaluation Campaign (SiSEC) [39].

6. New Software and Platforms

6.1. ASTALI

Automatic Speech-Text Alignment Software
KEYWORD: Speech-text alignment
FUNCTIONAL DESCRIPTION

ASTALI is a software for aligning a speech signal with its corresponding orthographic transcription (given in simple text file for short audio signals or in .trs files as generated by transcriber for longer speech signals). Using a phonetic lexicon and automatic grapheme-to-phoneme converters, all the possible sequences of phones corresponding to the text are generated. Then, using acoustic models, the tool finds the best phone sequence and provides the boundaries at the phone and at the word levels. ASTALI is available through a web application, which makes the service easy to use, without requiring any software downloading. This year, the integration of the web application on the ORTOLANG platform has been finalized.

- Participants: Dominique Fohr, Odile Mella, Antoine Chemardin, Valérian Girard and Denis Jouvet
- Contact: Dominique Fohr
- URLs: https://www.ortolang.fr/market/tools/astali; http://astali.loria.fr/; and http://ortolang108.inist.fr/astali/

6.2. 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

dnnsep is a new software that combines deep neural networks and multichannel signal processing for speech enhancement and separation of musical recordings.

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

6.3. JSnoori

FUNCTIONAL DESCRIPTION

JSnoori is written in Java and uses signal processing algorithms developed within the WinSnoori software with the double objective of being a platform independent signal visualization and manipulation tool, and also for designing exercises for learning the prosody of a foreign language. JSnoori can be used directly or via scripts written in Jython. This year, several approaches for computing the fundamental frequency have been added; and, JSnoori is now available through the ORTOLANG platform.

- Participants: Yves Laprie, Slim Ouni, Aghilas Sini and Ilef Ben Farhat
- Contact: Yves Laprie

6.4. 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. The audio stream is first split into homogeneous segments that are decoded using the most adequate acoustic model with a large vocabulary continuous speech recognition engine. In this new software, the recognition engine is based on the Kaldi toolkit, and uses Deep Neural Network - DNN - based acoustic models. An extra processing pass is run in order to rescore the n-best hypotheses with a higher order language model.

- Participants: Odile Mella, Dominique Fohr and Denis Jouvet
- Contact: Dominique Fohr
- URL: Available online on the Allgo platform: https://allgo.inria.fr/app/loriasts_kaldi

6.5. PLAVIS

Sofware for audio-visual and multimodal data acquisition and processing
FUNCTIONAL DESCRIPTION

Within the ADT PLAVIS (cf. 9.2.12), we have developed a software for 3D audiovisual data acquisition and synthesis. The system incorporates an animation module of the talking head to reconstruct the animated face along with audio. The acquisition software handles one or several acquisition systems: motion-capture (Kinect-like), Vicon or EMA systems. The various acquisition channels are synchronized. The animation technique can exploit multimodal data to define blendshapes that controls the face; the advantage of using blendshapes is to be able to transfer the animation from one 3D human model to another. A semi-automatic acoustic boundary correction process is integrated in the corpus building process. The text-to-speech processing is driven by the Soja software.

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

6.6. SOJA

Speech Synthesis platform in JAva
FUNCTIONAL DESCRIPTION

SOJA is a software for Text-To-Speech synthesis (TTS) which relies on a non uniform unit selection algorithm. 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. During 2016, the part of code in C was reduced to go to a full-Java software in the future. The natural language processing can now be restarted from any step. This functionality is useful for instance during corpus processing when using semi-automatic boundaries correction.

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

6.7. VisArtico

Visualization of EMA Articulatory data
FUNCTIONAL DESCRIPTION

VisArtico is a user-friendly software which allows visualizing EMA data acquired by an articulograph (AG500, AG501 or NDI Wave). This visualization software has been designed so that it can directly use the data provided by the articulograph to display the articulatory coil trajectories, synchronized with the corresponding acoustic recordings. Moreover, VisArtico not only allows viewing the coil trajectories but also enriches the visual information by indicating clearly and graphically the data for the tongue, lips and jaw. In addition, it is possible to insert images (MRI or X-Ray, for instance) to compare the EMA data with data obtained through other acquisition techniques. It is possible to generate a movie for any articulatory-acoustic sequence. During 2016, we have made a new version of VisArtico where the 3D view is now based on OpenGL. This allows a better quality rendering. It is possible to make measurement between sensors to compute the distance. Finally, we added the possibility to display the fundamental frequency on the spectrogram.

- Participants: Slim Ouni, Loic Mangeonjean, Ilef Ben Farhat and Bertrand Muller
- Contact: Slim Ouni
- URL: http://visartico.loria.fr

6.8. Xarticulators

KEYWORD: Medical imaging
FUNCTIONAL DESCRIPTION

The Xarticulators software is intended to delineate contours of speech articulators in X-ray 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 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 X-ray or 2D-MRI films.
During 2016, we developed a new version of the articulatory model which incorporates a more realistic model of the epiglottis and lips.

- Contact: Yves Laprie

6.9. Platforms

6.9.1. Platform MultiMod : Multimodal Acquisition Data Platform

FUNCTIONAL DESCRIPTION

Within a LORIA exploratory project (cf. 9.2.13), we have set up an acquisition hardware platform to acquire multimodal data in speech communication context. The system is composed of the articulograph Carstens AG501 (which was acquired as part of the EQUIPEX ORTOLANG - cf. 9.2.1), 4 Vicon cameras (a motion capture system), an Intel RealSense which is a depth camera (acquired as part of the project COREXp - cf. 9.1.1), a video camera and a microphone. 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, the system has been used to acquire multimodal data for the MCC project (cf. 9.4.2.1) and a first exploratory expressive multimodal corpus [40].

- 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: Yves Laprie, Slim Ouni, Vincent Colotte, Anne Bonneau, Agnès Piquard-Kipffer, Denis Jouvet, Odile Mella, Dominique Fohr, Benjamin Elie, Sucheta Ghosh, Anastasiia Tsukanova, Yang Liu, Sara Dahmani, Valérian Girard, Aghilas Sini.

7.1.1. Articulatory modeling

7.1.1.1. Acoustic simulations

The acoustic simulations play a central role in articulatory synthesis and should enable the production of all classes of sounds in a realistic manner. The production of voiced fricatives relies on a partial closure of the glottis which simultaneously creates an airflow which generates turbulence downwards from the constriction and the vibration of the vocal folds. Our acoustic simulation framework [14] has been extended to incorporate a glottal chink [29] in a self-oscillating vocal fold model. The glottis is then made up of two main separated components: a self-oscillating part and a constantly open chink. This feature allows the simulation of voiced fricatives, thanks to a self-oscillating model of the vocal folds to generate the voiced source, and the glottal opening that is necessary to generate the frication noise.

The acoustic propagation paradigm is appropriately chosen so that it can deal with complex geometries and a time-varying length of the vocal tract. Temporal scenarios for the dynamic shapes of the vocal tract and the glottal configurations were derived from the simultaneous acquisition of X-ray or MRI images and audio recording. Copy synthesis of a few French sentences [30], [31], [53] shows the accuracy of the simulation framework to reproduce acoustic cues of phrase-level utterances containing most of French phone (sound) classes while considering the real geometric shape of the speaker. For this purpose the articulatory model has been extended to offer a better precision of the epiglottis and of lips.

7.1.1.2. Acquisition of articulatory data

The acquisition of dynamic data is a key objective since speech production gestures involve the anticipation of the articulatory targets of the coming sounds. Cine-MRI represents an invaluable tool since it can image the whole vocal tract. However, speech requires a sampling frequency above 30 Hz to capture interesting information. Compressive sampling relies on partially collecting data in the Fourier space of the images acquired via MRI. The combination of compressed sensing technique, along with homodyne reconstruction, enables the missing data to be recovered [32]. The good reconstruction is guaranteed by an appropriate design of the sampling pattern. It is based on a pseudo-random Cartesian scheme, where each line is partially acquired for use of the homodyne reconstruction, and where the lines are pseudo-randomly sampled: central lines are constantly acquired and the sampling density decreases as the lines are far from the center.

7.1.1.3. Markerless articulatory acquisition techniques

With the spread of depth cameras (kinect-like systems), many researchers consider using these systems to track the movement of some speech articulators as lips and jaw. We are considering using this kind of system if it is suitable for speech production studies. For this reason, we have assessed the precision of markerless acquisition techniques when used to acquire articulatory data for speech production studies [19]. Two different markerless systems have been evaluated and compared to a marker-based one. The main finding is that both markerless systems provide reasonable results during normal speech and the quality is uneven during fast articulated speech. The quality of the data is dependent on the temporal resolution of the markerless system.

7.1.2. Expressive acoustic-visual synthesis

7.1.2.1. Expressive speech

A comparison between emotional and neutral speech was conducted using a small database containing utterances recorded in six emotional types (anger, fear, sadness, disgust, surprise and joy) as well as in a neutral pronunciation. The prosodic analysis focused on the main prosodic parameters such as vowel duration, energy and fundamental frequency (F0) level, and pause occurrences. The values of prosodic parameters were compared among the various emotional styles, as well as between emotional style and neutral style utterances. Moreover, the structuration of the sentences, in the various emotional styles, was particularly studied through a detailed analysis of pause occurrences and their length, and of the length of prosodic groups [23].

7.1.2.2. Expressive acoustic and visual speech

Concerning expressive audiovisual speech synthesis, a case study of a semi-professional actor who uttered a set of sentences for 6 different emotions in addition to neutral speech was conducted. Our purpose is to identify the main characteristics of audiovisual expressions that need to be integrated during synthesis to provide believable emotions to the virtual 3D talking head. We have recorded concurrently audio and motion capture data. The acoustic and the visual data have been analyzed. The main finding is that although some expressions are not well identified, some expressions were well characterized and tied in both acoustic and visual space [40]. The acquisition of the corpus was done with the platform software PLAVIS (cf. 9.2.12).

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

7.1.3.1. Categorization of sounds for native speech

We examined the schooling experiences of 166 young people with disabilities, aged from 6 to 20 years old. These children and teenagers had specific language impairment : SLI (severe language impairment), dyslexia, dysorthographia. The phonemic discrimination, phonological and phonemic analysis difficulties faced in their childhoods had raised reading difficulties which constituted a major obstacle, which the pupils did not overcome. Consequently, this led them to repeat one or more grades. This rate is 18 times higher than the French average. The importance of this cycle of learning can be better understood through this data, which could also enable, if not overcoming the handicap, to at least improving their learning possibilities [64].

7.1.3.2. Digital books for language impaired children

Three digital albums for language impaired children were designed within the Handicom (ADT funded by Inria). These three prototypes focus on the importance of multimodal speech combining written words and visual clues: a 3D avatar telling the stories and coding oral language in LPC (french cued speech) for hearing impaired children. Eight speech and language therapists used one of these albums (the digital prototype *Nina fête son anniversaire* !) with 8 children who are aged 5 years: 4 hearing impaired children, 2 children with SLI and 2 children with autism. The training they experienced with these children showed that the use of the digital book can foster some capacities involved in language learning [41].

7.1.3.3. Analysis of non-native pronunciations

The IFCASL corpus is a French-German bilingual phonetic learner corpus designed, recorded and annotated in the IFCASL project (cf. 9.2.6). It incorporates data for a language pair in both directions, i.e. in our case French learners of German, and German learners of French. In addition, the corpus is complemented by two sub-corpora of native speech by the same speakers. The corpus has been finalized, and provides spoken data by about 100 speakers with comparable productions, annotated and segmented at the word and phone levels, with more than 50% of manually checked and corrected data [51].

We investigated the correct placement of lexical (German) or post-lexical (French) accents [52]. French and German differ with respect to the representation and implementation of prominence. French can be assumed to have no prominence represented in the mental lexicon and accents are regularly assigned post-lexically on the last full vowel of an accentual group. In German, prominence is considered to be represented lexically. This difference may give rise to interferences when German speakers learn French and French speakers learn German. Results of a judgment task (conducted with 3 trained phoneticians) of native and nonnative productions of French learners of German and German learners of French, all of them beginners, show that both groups have not completely acquired the correct suprasegmental structures in the respective L2 [1], since both groups are worse concerning the correct placement of prominence than the native speakers. Furthermore, the results suggest that the native pattern is one of the most important factors for wrong prominence placements in the foreign language, e.g., if the prominence placement of L1 and L2 coincide, speakers produce the smallest amount of errors. Finally, results indicate that visual display of accented syllables increases the likelihood of a correct accent placement.

7.1.3.4. Implementation of acoustic feedback for devoicing of final fricatives

In view of implementing acoustic feedback in foreign language learning we analyzed acoustic cues which could explain that final fricatives are perceived as voiced or unvoiced. The ratio of unvoiced frames in the consonantal segment and also the ratio between consonantal duration and vowel duration were measured. As expected, we found that beginners face more difficulties to produce voiced fricatives than advanced learners. Also, the production becomes easier for the learners, especially for beginners, if they practice repetition after a native speaker. We use these findings to design and develop feedback via speech analysis/synthesis technique TD-PSOLA using the learner's own voice and voiced fricatives uttered by French speakers [36]. We selected fully voiced exemplars and evaluated whether the presence of an additional schwa fosters the perception of voicing by native French speakers.

7.2. Statistical Modeling of Speech

Participants: Antoine Liutkus, Emmanuel Vincent, Irène Illina, Dominique Fohr, Denis Jouvet, Vincent Colotte, Ken Deguernel, Mathieu Fontaine, Amal Houidhek, Aditya Nugraha, Imran Sheikh, Imene Zangar, Mohamed Bouallegue, Sunit Sivasankaran.

7.2.1. Source separation

7.2.1.1. Deep neural models for source separation

We pursued our research on the use of deep learning for multichannel source separation [18]. Our technique exploits both the spatial properties of the sources as modeled by their spatial covariance matrices and their spectral properties as modeled by a deep neural network. The model parameters are alternately estimated in an expectation-maximization (EM) fashion. We used this technique for music separation in the context of the 2016 Signal Separation Evaluation Campaign (SiSEC) [39]. We also used deep learning to address the fusion of multiple source separation techniques and found it to perform much better than the variational Bayesian model averaging techniques previously investigated [17].

We wrote an article about music source separation for the general public [59].

7.2.1.2. α-stable modeling of audio signals

The alpha-harmonizable model has recently been proposed by A. Liutkus et al. [66] as the only available probabilistic framework to account for signal processing methods manipulating fractional spectrograms instead of more traditional power spectrograms. Indeed, they generalize the classical Gaussian formulation and permit to handle large uncertainties or signal dynamics, which are both common in audio.

[1] L2 indicates the non-native language, whereas L1 indicates the native language

Our work on this topic this year has notably focused on its extension to the multichannel setting, which is important for music processing and source localization. Since inference in multivariate alpha-stable distribution is a very intricate issue, the approach undertaken has focused on analysing the multichannel signals through the joint analysis of multiple scalar projections on the real line. This results in an original algorithm called PROJET that combines computational tractability with the inherent robustness of alpha-stable models [15], [34].

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. In order to motivate further work by the community, we created an international evaluation campaign on that topic in 2011: the CHiME Speech Separation and Recognition Challenge. After three successful editions [11], [55], we organized the fourth edition in 2016. We also summarized the speech distortion conditions in real scenarios for speech processing applications [42] and collected a French corpus for distant-microphone speech processing in real homes [24].

Speech enhancement and automatic speech recognition (ASR) are most often evaluated in matched (or multi-condition) 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 [21]. The results show that most algorithms perform consistently on real and simulated data and are barely affected by training on different noise environments. This suggests that DNNs generalize more easily than previously thought.

7.2.2.2. Environmental sounds

We explored acoustic modeling for the classification of environmental sound events and sound scenes and submitted our system to the DCASE 2016 Challenge [33].

7.2.3. Linguistic 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 [47]. The goal of this work is to model the semantic and topical context of new proper names in order to retrieve OOV words which are relevant to the spoken content in the audio document. Probabilistic topic models [44] and word embeddings from neural network models are explored for the task of retrieval of relevant proper names. Neural network context models trained with an objective to maximise the retrieval performance are proposed. A Neural Bag-of-Words (NBOW) model trained to learn context vector representations at a document level is shown to outperform the generic representations. The proposed Neural Bag-of-Weighted-Words (NBOW2) model learns to assign a degree of importance to input words and has the ability to capture task specific key-words [46] [45]. Experiments on automatic speech recognition of French broadcast news videos demonstrate the effectiveness of the proposed models. Further evaluation of the NBOW2 model on standard text classification tasks, including movie review sentiment classification and newsgroup topic classification, shows that it learns interesting information about the task and gives the best classification accuracies among the bag-of-words models.

7.2.3.2. Adding words in a language model

Out-of-vocabulary (OOV) words can pose a particular problem for automatic speech recognition of broadcast news. The language models (LMs) of ASR systems are typically trained on static corpora, whereas new words (particularly new proper nouns) are continually introduced in the media. Additionally, such OOVs are often content-rich proper nouns that are vital to understanding the topic. We explore methods for dynamically adding OOVs to language models by adapting the n-gram language model used in our ASR system. We propose two strategies: the first one relies on finding in-vocabulary (IV) words similar to the OOVs, where word embeddings are used to define similarity. Our second strategy leverages a small contemporary corpus to estimate OOV probabilities. The models we propose yield improvements in perplexity over the baseline; in addition, the corpus-based approach leads to a significant decrease in proper noun error rate over the baseline in recognition experiments [26].

7.2.3.3. 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. We published two articles that summarize our work on the System & Contrast model for the characterization of the mid-term and long-term structure of music [12] and on the structural segmentation of popular music pieces using a regularity constraint that naturally stems from this model [20], [58]. We also proposed a new model for automatic music improvisation that combines a multi-dimensional probabilistic model encoding the musical experience of the system and a factor oracle encoding the local context of the improvisation [27].

7.2.4. Speech generation by statistical methods

Work on HMM-based Arabic speech synthesis was carried out within a CMCU PHC project with ENIT (Engineer school at Tunis-Tunisia; cf. 9.4.2.2). A first version of the system, based on the HTS toolkit (HMM-based Speech Synthesis System), is now working; and the study of the impact of some parameters is ongoing.

In parallel, the HTS system is also applied to the French language.

7.3. Uncertainty Estimation and Exploitation in Speech Processing

Participants: Emmanuel Vincent, Odile Mella, Dominique Fohr, Denis Jouvet, Baldwin Dumortier, Juan Andres Morales Cordovilla, Karan Nathwani, Ismaël Bada.

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 in order to exploit it for subsequent feature extraction and decoding. A number of uncertainty estimators have been proposed in the literature, which are typically based on fixed mathematical approximations or heuristics. We finalized our work on a principled variational Bayesian approach to uncertainty estimation and showed its benefit w.r.t. other estimators for speech and speaker recognition [9]. We also pursued our work on the propagation of uncertainty in deep neural network acoustic models.

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 [38]. 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 [62].

7.3.2. Uncertainty and phonetic segmentation

7.3.2.1. Speech-text alignment

We have continued our work on determining more accurate phonetic boundaries with two new approaches based on DNN. The first approach proposes to find phonetic boundaries directly from the parameterized speech signal using an LSTM (Long Short-Term Memory) neural network. The aim of the second approach is twofold: provide confidence measures for evaluating speech-text alignment outputs and refine these outputs. One of these studies was done with the Synalp team of LORIA in the framework of the project ORFEO (cf. 9.2.5). The achieved confidence measure outperforms a confidence score (based on acoustic posterior probability) derived from a state-of-the-art text-to-speech aligner [43].

Within the IFCASL project (cf. 9.2.6), we have also developed a speech-text alignment system for German which will be integrated into the ASTALI software.

7.3.3. Uncertainty and prosody

The study of discourse particles that was initiated last year, has continued in the framework of the CPER LCHN (cf. 9.1.2). A larger set of words and expressions that can be used either as normal lexical words or as discourse particles (as for example *quoi* (what), *voilà* (there it is), ...) has been considered. For each of these words/expressions and for each speech corpus that was aligned in the ORFEO project (cf. 9.2.5), a subset of about one hundred occurrences were selected. Thanks to the CPER LCHN support, a part of these occurrences have been annotated as "discourse particle" or "non discourse particle". Detailed analysis is in progress, with respect to the function (discourse particle or not), the type of speech corpus, and the associated prosodic features.

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. Studies have also been initiated for computing a confidence measure on the estimated F0 values through the use of neural network approaches.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. Venathec

Company: Venathec SAS

Other partners: ACOEM Group, GE Intelligent Platforms (contracted directly with Venathec)

Duration: June 2014 - August 2017

Supported by: Bpifrance

Abstract: The project aims to design a real-time control system for wind farms that will maximize energy production while limiting sound nuisance. This will leverage our know-how on audio source separation and uncertainty modeling and propagation.

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. COREXp

Project acronym: COREXp

Project title: Acquisition, Processing and Analysis of a Corpus for the Synthesis of Expressive Audiovisual Speech

Duration: December 2014 - December 2016

Coordinator: Slim Ouni

Cofunded by Inria and Région Lorraine

Abstract: The main objective of this project was the acquisition of a bimodal corpus of a considerable size (several thousand sentences) to study the expressiveness and emotions during speech (for example, how to decode facial expressions that are merged with speech signals). The main purpose was to acquire, process and analyze the corpus and to study the expressiveness; the results will be used for the expressive audiovisual speech synthesis system.

9.1.2. CPER LCHN

Project acronym: CPER LCHN

Project title: CPER "Langues, Connaissances et Humanités Numériques"

Duration: 2015-2020

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

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.

9.1.3. CPER IT2MP

Project acronym: CPER IT2MP

Project title: CPER "Innovation Technologique Modélisation et Médecine Personalisée"

Duration: 2015-2020

Coordinator: Faiez Zannad (Inserm-CHU-UL)

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.

9.1.4. SATT Dynalips

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

Duration: May 2016 - December 2017

Coordinator: Slim Ouni

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. EQUIPEX ORTOLANG

Project acronym: ORTOLANG [2]

[2] http://www.ortolang.fr

Project title: Open Resources and TOols for LANGuage

Duration: September 2012 - December 2016 (phase I)

Coordinator: Jean-Marie Pierrel, ATILF (Nancy)

Other partners: LPL (Aix en Provence), LORIA (Nancy), Modyco (Paris), LLL (Orléans), INIST (Nancy)

Abstract: The aim of ORTOLANG was to propose a network infrastructure offering a repository of language data (corpora, lexicons, dictionaries, etc.) and tools and their treatment that are readily available and well-documented. This will enable a real mutualization of analysis research, of modeling and automatic treatment of the French language. This will also facilitate the use and transfer of resources and tools set up within public laboratories towards industrial partners, in particular towards SME which often cannot develop such resources and tools for language treatment due to the costs of their realization. Moreover, this will promote the French language and local languages of France by sharing knowledge which has been acquired by public laboratories.

Several teams of the LORIA laboratory contribute to this Equipex, mainly with respect to providing tools for speech and language processing. MULTISPEECH contributes with text-speech alignment and speech visualization tools.

9.2.2. 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.

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.

9.2.3. PIA2 ISITE LUE

Project acronym: ISITE LUE

Project title: Lorraine Université d'Excellence

Duration: starting in 2016

Coordinator: Univ. Lorraine

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.

9.2.4. 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

Abstract: The ContNomina project focuses 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 addresses 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).

9.2.5. ANR ORFEO

Project acronym: ORFEO [3]

Project title: Outils et Ressources pour le Français Écrit et Oral

Duration: February 2013 - February 2016

Coordinator: Jeanne-Marie DEBAISIEUX (Université Paris 3)

Other partners: ATILF, CLLE-ERSS, ICAR, LIF, LORIA, LATTICE, MoDyCo

Abstract: The main objective of the ORFEO project is the constitution of a corpus for the study of contemporary French.

In this project, we were concerned by the automatic speech-text alignment at the word and phoneme levels for audio files from several corpora gathered by the project. These corpora orthographically transcribed with Transcriber contain mainly spontaneous speech, recorded under various conditions with a large SNR range and a lot of overlapping speech and anonymised speech segments. For the forced speech-text alignment phase, we applied our 2-step methodology (the first step uses a detailed acoustic model for finding the pronunciation variants; then, in the second step a more compact model is used to provide more temporally accurate boundaries).

9.2.6. ANR-DFG IFCASL

Project acronym: IFCASL

Project title: Individualized feedback in computer-assisted spoken language learning

Duration: March 2013 - December 2016

Coordinator: Jürgen Trouvain (Saarland University)

Other partners: Saarland University (COLI department)

Abstract: The main objective of IFCASL is to investigate learning of oral French by German speakers, and oral German by French speakers at the phonetic level.

A French-German learner corpus was designed and recorded. French speakers were recorded in Nancy, whereas German speakers were recorded in Saarbrücken. An automatic speech-text alignment process was applied on all the data. Then, the French speech data (native and non-native) were manually checked and annotated in France, and the German speech data (native and non-native) were manually checked and annotated in Germany. The corpora are currently used for analyzing non-native pronunciations, and studying feedback procedures.

[3] http://www.agence-nationale-recherche.fr/en/anr-funded-project/?tx_lwmsuivibilan_pi2[CODE]=ANR-12-CORP-0005

9.2.7. ANR DYCI2

Project acronym: DYCI2 [4]

Project title: Creative Dynamics of Improvised Interaction

Duration: March 2015 - February 2018

Coordinator: Ircam (Paris)

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

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.

9.2.8. 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

Abstract: 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.

9.2.9. 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)

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.

Measurements of glottis opening during the production of fricatives via EPGG (ElectroPhotoGlottoGraphy), the design of acoustic experiments with a replica of the vocal tract and the design of dynamic acquisition with MRI were the main activities of this first year.

9.2.10. FUI RAPSODIE

Project acronym: RAPSODIE

Project title: Automatic Speech Recognition for Hard of Hearing or Handicapped People

Duration: March 2012 - February 2016

Coordinator: eRocca (Mieussy, Haute-Savoie)

[4] http://repmus.ircam.fr/dyci2/

Other partners: CEA (Grenoble), Inria (Nancy), CASTORAMA (France)

Abstract: The goal of the project was to realize a portable device to help a hard-of-hearing person to communicate with other people. To achieve this goal the portable device needs to access a speech recognition system, adapted to this task. Another application of the device is environment vocal control for handicapped persons.

In this project, MULTISPEECH was involved in optimizing the speech recognition models for the envisaged task, and in finding the best way of presenting the speech recognition results in order to maximize the communication efficiency between the hard-of-hearing person and the speaking person.

9.2.11. FUI VoiceHome

Project acronym: VoiceHome

Duration: February 2015 - July 2017

Coordinator: onMobile

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

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

9.2.12. ADT Plavis

Project acronym: Plavis

Project title: Platform for acquisition and audiovisual speech synthesis

Duration: January 2015 - December 2016

Coordinator: Vincent Colotte

Abstract: The objective of this project was to develop a platform acquisition and audiovisual synthesis system (3D animation of the face synchronously with audio). The main purpose was to build a comprehensive platform for acquisition and processing of audiovisual corpus (selection, acquisition and acoustic processing, 3D visual processing and linguistic processing). The acquisition was performed using a motion-capture system (Kinect-like), a Vicon system, and an electromagnetic articulography (EMA) system.

9.2.13. LORIA exploratory project

Project title: Acquisition and processing of multimodal corpus in the context of interactive human communication

Duration: June 2015 - May 2016

Coordinator: Slim Ouni

Abstract: The aim of this project was the study of the various mechanisms involved in multimodal human communication that can be oral, visual, gestural and tactile. This project focused on the identification and acquisition of a very large corpus of multimodal data from multiple information sources and acquired in the context of interaction and communication between two people or more.

9.2.14. SYNABE

Project acronym: SYNABE

Project title: Articulatory data synchronization for studying stuttering

Duration: January 2016 - December 2016

Coordinator: Fabrice Hirsch (Praxiling, UMR 5267, Montpellier)

Other partners: S. Ouni

Funding: CNRS DEFI Instrumentation aux limites

Abstract: The objective of this project is to use simultaneously three hardware allowing having information on the subglottic (respiratory belt), glottic (electroglottograph) and supraglottic (articulograph) levels during the production of the speech in order to know the timing of the gestures during speech. This system will be used to study the motor coordination between the three levels mentioned in the stuttering and normo-fluent words. We will propose a new typology of normal and pathological disfluencies.

Our main contribution concerned the articulatory data acquisition using the articulograph AG501.

9.3. European Initiatives

9.3.1. Collaborations with Major European Organizations

Jon Barker: University of Sheffield (UK)

Robust speech recognition [11], [55]

9.4. International Initiatives

9.4.1. Inria International Partners

9.4.1.1. Informal International Partners

Jonathan Le Roux, Shinji Watanabe, John R. Hershey: Mitsubishi Electric Research Labs (MERL, Boston, USA)

Robust speech recognition [11], [55]

Dayana Ribas Gonzalez, Ramón J. Calvo: CENATAV (Habana, Cuba)

Robust speaker recognition [42]

9.4.2. Participation in Other International Programs

9.4.2.1. STIC-AmSud - multimodal communication corpus

STIC-AmSud: MCC - Multimodal Communication Corpus. A collaboration: Argentina, Chile and France (01/2015-12/2016)

Project acronym: MCC

Project title: Multimodal Communication Corpus

Duration: January 2015 - December 2016

International Coordinator: S. Ouni

National Coordinators: Nancy Hitschfeld (Depto. de Ciencias de la Computación (DCC), Universidad de Chile) - Chile; and, Juan Carlos Gomez (Centro Internacional Franco Argentino de Ciencias de la Información y de Sistemas (CIFASIS), UNR, CONICET) - Argentina

Abstract: The project aims to collect a multimodal speech corpus containing synchronized audio-visual data recorded from talking individuals. The corpus will incorporate several communication modes which appear in the communication among humans, such as the acoustic signal, facial movements and body gestures during speech. During 2016, a complete corpus of 8 speakers (4 French and 4 Spanish) has been acquired and processed. The corpus will be distributed using the Ortolang platform.

9.4.2.2. PHC UTIQUE - HMM-based Arabic speech synthesis

PHC UTIQUE - HMM-based Arabic speech synthesis, with ENIT (Engineer school at Tunis-Tunisia)

Duration: 2015 - 2018.

Coordinators: Vincent Colotte (France) and Noureddine Ellouze (Tunisia).

Abstract: Development of an HMM-based speech synthesis system for the Arabic language. This includes the development of an Arabic speech corpus, the selection of linguistic features relevant to Arabic HMM-based speech synthesis, as well as improving the quality of the speech signal generated by the system.

9.5. International Research Visitors

9.5.1. Visits of International Scientists

Sebastian Gonzales Mora

Date: Jan 2016

Faculty de Cs. Físicas y Matemáticas, University of Chile

Benjamin Martinez Elizalde

Date: May 2016 - Aug 2016

Institution: Carnegie Mellon University (USA)

Dayana Ribas Gonzalez

Date: Sep 2016 - Dec 2016

Institution: CENATAV (Cuba)

Ziteng Wang

Date: Sep 2016 - Sep 2017

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

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. General Chair, Scientific Chair

General co-chair, 4th CHiME Speech Separation and Recognition Challenge (E. Vincent)

General co-chair, 4th International Workshop on Speech Processing in Everyday Environments, San Francisco, USA, September 2016 (E. Vincent)

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

Chair, SiSEC 2016, Signal Separation Evaluation Challenge (A. Liutkus)

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

Chair, Challenges Subcommittee, IEEE Technical Committee on Audio and Acoustic Signal Processing (E. Vincent)

10.1.1.2. Member of the Organizing Committees

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)

10.1.1.3. Member of the Conference Program Committees

Area chair for Analysis of Speech and Audio Signal, INTERSPEECH'2016 (D. Jouvet)

Area chair for Audio and Speech Source Separation, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (E. Vincent)

10.1.1.4. Reviewer

CHiME 2016 - Speech Separation and Recognition Challenge (E. Vincent)

EUSIPCO 2016 - European Signal Processing Conference (D. Jouvet)

ICECS 2016 - Conference on Environmental and Computer Science (R. Serizel)

INTERSPEECH 2016 (A. Bonneau, S.Ouni, E. Vincent, I. Illina, Y. Laprie)

IROS 2016 - International Conference on Intelligent Robots and Systems (E. Vincent)

IVA 2016 - International Conference on Intelligent Virtual Agents (S. Ouni)

JEP 2016 - Journées d'Etudes sur la Parole (D. Jouvet, A. Bonneau, E. Vincent, D. Fohr, I. Illina, O. Mella, Y. Laprie)

SLT 2016 - IEEE Spoken Language Technology Workshop (E. Vincent)

10.1.2. Journal

10.1.2.1. Member of the Editorial Boards

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

Speech Communication (D. Jouvet)

Traitement du signal (E. Vincent)

International Journal of Learner Corpus Research, special issue on "Investigating segmental, prosodic and fluency features in spoken learner corpora" (A. Bonneau, Guest Editor)

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

10.1.2.2. Reviewer - Reviewing Activities

Computer Speech and Language (D. Jouvet, S. Ouni, E. Vincent)

Digital Signal Processing (E. Vincent)

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

IEEE Transactions on Signal Processing (A. Liutkus)

IEEE Signal Processing Letters (A. Liutkus, R. Serizel)

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

JASA Express Letters (Y. Laprie)

Speech Communication (E. Vincent, Y. Laprie)

10.1.3. Invited Talks

Séparation de sources: quand l'acoustique rencontre le machine learning, keynote talk, 13e Congrès Français d'Acoustique (E. Vincent) [22]

Speech recognition, Ecole Nationale d'Ingénieurs de Tunis, May 2016 (D. Jouvet)

Les corpus acoustiques et langagiers pour la reconnaissance de la parole, Seminar about Big Data, LIA-LINOS (Laboratoire International Associé), May 2016 (O. Mella)

Parole audiovisuelle : pour faciliter la communication parlée, Praxiling, Université de Montpellier 3, October 2016 (S. Ouni)

10.1.4. 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)

Member IEEE Technical Committee on Audio and Acoustic Signal Processing (A. Liutkus)

10.1.5. Scientific Expertise

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

Expertise of a project for the Research Foundation Flanders – FWO (S. Ouni)

10.1.6. Research Administration

Elected Member of the board of the AM2I Scientific Pole - Université de Lorraine (Y. Laprie)

Member of the HCERES visiting committee for LIUM (D. Jouvet)

Chairman of selection committee for the position of Assistant Professor (ESSTIN 0762, May 2016), Y. Laprie.

Member of a selection committee (Université d'Avignon, May 2016), E. Vincent

Member of a selection committee (Université du Maine, May 2016), D. Jouvet

Member of a selection committee (Télécom ParisTech, June 2016), E. Vincent

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

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, Introduction to computer tools, 108h, L1, University of Lorraine – IUT Nancy Charlemagne, France

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, 28h, L1, University of Lorraine, France

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

Licence: O. Mella, Computer Networking, 80h, L2-L3, University of Lorraine, France

Licence: A. Piquard-Kipffer, Education Sciences, 36h, L1, France

Licence: A. Piquard-Kipffer, Reading and Writing, 27h, L2, Département Orthophonie, University of Lorraine, France

Licence: A. Piquard-Kipffer, Psycholinguistics, 6 hours, L2 Département Orthophonie, University Pierre et Marie Curie-Paris, France

Licence: A. Piquard-Kipffer, Reading and Writing assessment, 10h, L3, Département Orthophonie, University of Lorraine, France

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

Master: Y. Laprie, Analyse, perception et reconnaissance de la parole, 32 hours, M1, University of Lorraine, France

Master: O. Mella, Computer Networking, 74h, 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, Dyslexia, 25 hours, M1, Département Orthophonie, University of Lorraine, France

Master: A. Piquard-Kipffer, Reading and writing, 6 hours, M1, Département Orthophonie, University Pierre et Marie Curie-Paris, France

Master: A. Piquard-Kipffer, Deaf People and Reading, 15 hours, M2 Département Orthophonie, University of Lorraine, France

Master: A. Piquard-Kipffer, Psychology, 40 hours, M2, ESPE, University of Lorraine, France

Master: A. Piquard-Kipffer, French Language Didactics, 80 hours, M2, ESPE, University of Lorraine, France

Engineer school: V. Colotte, Conception and developpement in XML, 20h, Bac+3, Telecom Nancy, France

Ecole d'audioprothèse : A. Bonneau, Phonetics, 16 h, University of Lorraine

Doctorat: A. Piquard-Kipffer, Language Pathology - speech and language screening, 15 hours, EHESP, University of Sorbonne- Paris Cité, France

Adults: O. Mella, Computer science courses for seconday school teachers (Informatique et Sciences du Numérique courses) (21h), ESPE of Academy Nancy-Metz, 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: Imran Sheikh, "Exploiting Semantic and Topic Context to Improve Recognition of Proper Names in Diachronic Audio Documents", November 2016, Irina Illina, Dominique Fohr and Georges Linares.

PhD in progress: Baldwin Dumortier, "Contrôle acoustique d'un parc éolien", September 2014, Emmanuel Vincent and Madalina Deaconu.

PhD in progress: Quan Nguyen, "Mapping of a sound environment by a mobile robot", November 2014, Francis Colas and Emmanuel Vincent.

PhD in progress: Aditya Nugraha, "Deep neural networks for source separation and noise-robust speech recognition", January 2015, 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.

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

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

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 and Emmanuel Vincent.

PhD in progress: Yang Liu, "Merging acquisition and processing of cineMRI of the vocal tract", October 2016, Pierre-André Vuissoz and Yves Laprie.

10.2.3. Participation in HDR and PhD juries

Participation in PhD thesis Jury for David Guennec (Université Rennes 1, September 2016), Y. Laprie.

Participation in PhD thesis Jury for Ugo Marchand (Université Paris 6, November 2016), E. Vincent, reviewer.

Participation in PhD thesis Jury for Joachim Flocon-Cholet (Université Rennes 1, June 2016), E. Vincent, reviewer.

Participation in PhD thesis Jury for Aly Magassouba (Université Rennes 1, December 2016), E. Vincent, reviewer.

Participation in PhD thesis Jury for Diandra Fabre (Université Grenoble Alpes, December 2016), S. Ouni, reviewer.

Participation in PhD thesis Jury for Ivana Didirková (Université Montpellier 3, December 2016), Y. Laprie, reviewer.

10.2.4. Participation in other juries

Chairman of Scientific « Baccalauréat », specialty Earth Sciences (Académie de Nancy-Metz and Université de Lorraine, July 2016), A. Piquard-Kipffer.

Participation in the Competitive Entrance Examination into Speech-Language Pathology Department (University of Lorraine, June 2016), A. Piquard-Kipffer.

10.3. Popularization

Demonstration at Village Sciences LORIA, March 2016 (K. Deguernel, E. Vincent, S. Ouni).

Demonstration at Forum des métiers, Collège Peguy, Le Chesnay, March 2016 (A. Piquard-Kipffer).

Demonstration at EHESP-University of Sorbonne- Paris Cité, March 2016 (A. Piquard-Kipffer).

Demonstration at LORIA's 40th Anniversary, June 2016 (K. Deguernel, E. Vincent).

"Démixer la musique", Interstices, January 2016 (A. Liutkus and E. Vincent).

Intervention lors d'une action pour la Maison pour la Science en Lorraine au service des professeurs (A. Bonneau)

Démonstration lors de la journée Rencontre Inria Industrie sur le thème « Ed-Techs au service de e-Education », December 2016 (D. Jouvet)

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° 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° 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° 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° 4, pp. 1118 - 1133, 16, https://hal.archives-ouvertes.fr/hal-00626962

[7] 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

Doctoral Dissertations and Habilitation Theses

[8] I. SHEIKH. *Exploiting Semantic and Topic Context to Improve Recognition of Proper Names in Diachronic Audio Documents*, Université de Lorraine, November 2016, https://hal.archives-ouvertes.fr/tel-01400694

Articles in International Peer-Reviewed Journals

[9] K. ADILOĞLU, E. VINCENT. *Variational Bayesian Inference for Source Separation and Robust Feature Extraction*, in "IEEE Transactions on Audio Speech and Language Processing", June 2016 [*DOI :* 10.1109/TASLP.2016.2583794], https://hal.inria.fr/hal-00726146

[10] M. ARON, M.-O. BERGER, E. KERRIEN, B. WROBEL-DAUTCOURT, B. POTARD, Y. LAPRIE. *Multimodal acquisition of articulatory data: Geometrical and temporal registration*, in "Journal of the Acoustical Society of America", 2016, vol. 139, n° 2, 13 p. [*DOI :* 10.1121/1.4940666], https://hal.inria.fr/hal-01269578

[11] J. BARKER, R. MARXER, E. VINCENT, S. WATANABE. *The third 'CHIME' speech separation and recognition challenge: Analysis and outcomes*, in "Computer Speech and Language", October 2016, https://hal.inria.fr/hal-01382108

[12] F. BIMBOT, E. DERUTY, G. SARGENT, E. VINCENT. *System & Contrast : A Polymorphous Model of the Inner Organization of Structural Segments within Music Pieces*, in "Music Perception", 2016, 41 p. , https://hal.inria.fr/hal-01188244

[13] M. CADOT, Y. LAPRIE. *Extraction d'un modèle articulatoire à partir d'une analyse tri-directionnelle de cinéradiographies d'un locuteur*, in "Revue des Nouvelles Technologies de l'Information", 2016, vol. Fouille de Données Complexes, n° RNTI-E-31, pp. 73-92, https://hal.archives-ouvertes.fr/hal-01346987

[14] B. ELIE, Y. LAPRIE. *Extension of the single-matrix formulation of the vocal tract: consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink*, in "Speech Communication", September 2016, vol. 82, pp. 85-96 [*DOI :* 10.1016/J.SPECOM.2016.06.002], https://hal.archives-ouvertes.fr/hal-01199792

[15] D. FITZGERALD, A. LIUTKUS, R. BADEAU. *Projection-based demixing of spatial audio*, in "IEEE Transactions on Audio, Speech and Language Processing", May 2016, https://hal.inria.fr/hal-01260588

[16] 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", December 2016, https://hal.inria.fr/hal-01414179

[17] X. JAUREGUIBERRY, E. VINCENT, G. RICHARD. *Fusion methods for speech enhancement and audio source separation*, in "IEEE Transactions on Audio, Speech and Language Processing", April 2016, https://hal.archives-ouvertes.fr/hal-01120685

[18] A. A. NUGRAHA, A. LIUTKUS, E. VINCENT. *Multichannel audio source separation with deep neural networks*, in "IEEE/ACM Transactions on Audio, Speech, and Language Processing", June 2016, vol. 24, n° 10, pp. 1652-1664 [*DOI :* 10.1109/TASLP.2016.2580946], https://hal.inria.fr/hal-01163369

[19] S. OUNI, S. DAHMANI. *Is markerless acquisition of speech production accurate ?*, in "Journal of the Acoustical Society of America", May 2016, vol. 139, n° 6, https://hal.inria.fr/hal-01315579

[20] 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

[21] 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", November 2016, https://hal.inria.fr/hal-01399180

Invited Conferences

[22] E. VINCENT. *Séparation de sources: quand l'acoustique rencontre le machine learning*, in "13e Congrès Français d'Acoustique", Le Mans, France, April 2016, https://hal.inria.fr/hal-01398720

International Conferences with Proceedings

[23] K. BARTKOVA, D. JOUVET, E. DELAIS-ROUSSARIE. *Prosodic Parameters and Prosodic Structures of French Emotional Data*, in "Speech Prosody 2016", Boston, United States, Speech Prosody 2016, May 2016, https://hal.inria.fr/hal-01293516

[24] N. BERTIN, E. CAMBERLEIN, E. VINCENT, R. LEBARBENCHON, S. PEILLON, É. LAMANDÉ, S. SIVASANKARAN, F. BIMBOT, I. ILLINA, A. TOM, S. FLEURY, E. JAMET. *A French corpus for distant-microphone speech processing in real homes*, in "Interspeech 2016", San Francisco, United States, September 2016, https://hal.inria.fr/hal-01343060

[25] M. CADOT, A. BONNEAU. *Du fichier audio à l'intonation en Français :Graphes pour l'apprentissage de 3 classes intonatives*, in "Fouille de données complexes (FDC@EGC2016)", Reims, France, Proceedings of FDC@EGC2016, January 2016, https://hal.archives-ouvertes.fr/hal-01292121

[26] A. CURREY, I. ILLINA, D. FOHR. *Dynamic adjustment of language models for automatic speech recognition using word similarity* , in "IEEE Workshop on Spoken Language Technology (SLT 2016)", San Diego, CA, United States, proceeding of IEEE Workshop on Spoken Language Technology, December 2016, https://hal.archives-ouvertes.fr/hal-01384365

[27] K. DÉGUERNEL, E. VINCENT, G. ASSAYAG. *Using Multidimensional Sequences For Improvisation In The OMax Paradigm*, in "13th Sound and Music Computing Conference", Hamburg, Germany, August 2016, https://hal.inria.fr/hal-01346797

[28] B. ELIE, G. CHARDON. *Robust tonal and noise separation in presence of colored noise, and application to voiced fricatives*, in "22nd International Congress on Acoustics (ICA)", Buenos Aires, Argentina, September 2016, https://hal.archives-ouvertes.fr/hal-01372313

[29] B. ELIE, Y. LAPRIE. *A glottal chink model for the synthesis of voiced fricatives*, in "International Conference on Acoustics, Speech and Signal Processing (ICASSP)", Shanghai, China, IEEE, March 2016, https://hal.archives-ouvertes.fr/hal-01314308

[30] B. ELIE, Y. LAPRIE. *Copy synthesis of phrase-level utterances*, in "EUSIPCO2016", Budapest, Hungary, August 2016, https://hal.archives-ouvertes.fr/hal-01278462

[31] B. ELIE, Y. LAPRIE. *Copy synthesis of running speech based on vocal tract imaging and audio recording*, in "22nd International Congress on Acoustics (ICA)", Buenos Aires, Argentina, September 2016, https://hal.archives-ouvertes.fr/hal-01372310

[32] B. ELIE, Y. LAPRIE, P.-A. VUISSOZ, F. ODILLE. *High spatiotemporal cineMRI films using compressed sensing for acquiring articulatory data*, in "EUSIPCO2016", Budapest, Hungary, August 2016, https://hal.archives-ouvertes.fr/hal-01372320

[33] B. ELIZALDE, A. KUMAR, A. SHAH, R. BADLANI, E. VINCENT, B. RAJ, I. LANE. *Experiments on the DCASE Challenge 2016: Acoustic scene classification and sound event detection in real life recording*, in "DCASE2016 Workshop on Detection and Classification of Acoustic Scenes and Events", Budapest, Hungary, September 2016, https://hal.inria.fr/hal-01354007

[34] D. FITZGERALD, A. LIUTKUS, R. BADEAU. *PROJET - Spatial Audio Separation Using Projections*, in "41st International Conference on Acoustics, Speech and Signal Processing (ICASSP)", Shanghai, China, IEEE, 2016, https://hal.archives-ouvertes.fr/hal-01248014

[35] 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, Proc. 13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017), February 2017, https://hal.archives-ouvertes.fr/hal-01401988

[36] S. GHOSH, C. FAUTH, A. SINI, Y. LAPRIE. *L1-L2 Interference: The case of final devoicing of French voiced fricatives in final position by German learners*, in "Interspeech 2016", San Francisco, United States,

September 2016, vol. 2016, pp. 3156 - 3160 [*DOI :* 10.21437/INTERSPEECH.2016-954], https://hal.inria.fr/hal-01397176

[37] 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

[38] V. Q. NGUYEN, F. COLAS, E. VINCENT, F. CHARPILLET. *Localizing an Intermittent and Moving Sound Source Using a Mobile Robot*, in "International Conference on Intelligent Robots and Systems (IROS)", Deajeon, South Korea, October 2016, https://hal.archives-ouvertes.fr/hal-01354006

[39] A. A. NUGRAHA, A. LIUTKUS, E. VINCENT. *Multichannel music separation with deep neural networks*, in "European Signal Processing Conference (EUSIPCO)", Budapest, Hungary, Proceedings of the 24th European Signal Processing Conference (EUSIPCO), August 2016, pp. 1748-1752, https://hal.inria.fr/hal-01334614

[40] S. OUNI, V. COLOTTE, S. DAHMANI, S. AZZI. *Acoustic and Visual Analysis of Expressive Speech: A Case Study of French Acted Speech*, in "Interspeech 2016", San Francisco, United States, ISCA, November 2016, vol. 2016, pp. 580 - 584 [*DOI :* 10.21437/INTERSPEECH.2016-730], https://hal.inria.fr/hal-01398528

[41] 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

[42] D. RIBAS, E. VINCENT, J. R. CALVO. *A study of speech distortion conditions in real scenarios for speech processing applications*, in "2016 IEEE Workshop on Spoken Language Technology", San Diego, United States, December 2016, https://hal.inria.fr/hal-01377638

[43] G. SERRIÈRE, C. CERISARA, D. FOHR, O. MELLA. *Weakly-supervised text-to-speech alignment confidence measure*, in "International Conference on Computational Linguistics (COLING)", Osaka, Japan, Proceedings of the 26th International Conference on Computational Linguistics (COLING), December 2016, https://hal.archives-ouvertes.fr/hal-01378355

[44] I. SHEIKH, I. ILLINA, D. FOHR, G. LINARES. *Document Level Semantic Context for Retrieving OOV Proper Names*, in "2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", shanghai, China, Proceeding of IEEE ICASSP 2016, IEEE, March 2016, pp. 6050-6054 [*DOI :* 10.1109/ICASSP.2016.7472839], https://hal.archives-ouvertes.fr/hal-01331716

[45] I. SHEIKH, I. ILLINA, D. FOHR, G. LINARES. *Improved Neural Bag-of-Words Model to Retrieve Out-of-Vocabulary Words in Speech Recognition*, in "INTERSPEECH 2016", San Francisco, United States, Proceedings of INTERSPEECH 2016, September 2016, vol. 2016 [*DOI :* 10.21437/INTERSPEECH.2016-1219], https://hal.archives-ouvertes.fr/hal-01384488

[46] I. SHEIKH, I. ILLINA, D. FOHR, G. LINARES. *Learning Word Importance with the Neural Bag-of-Words Model*, in "ACL, Representation Learning for NLP (Repl4NLP) workshop", Berlin, Germany, Proceedings of ACL 2016, August 2016, https://hal.archives-ouvertes.fr/hal-01331720

[47] I. SHEIKH, I. ILLINA, D. FOHR. *How Diachronic Text Corpora Affect Context based Retrieval of OOV Proper Names for Audio News*, in "LREC 2016", Portoroz, Slovenia, proceedings of LREC 2016, May 2016, https://hal.archives-ouvertes.fr/hal-01331714

[48] A. J. R. SIMPSON, G. ROMA, E. M. GRAIS, R. D. MASON, C. HUMMERSONE, A. LIUTKUS, M. D. PLUMBLEY. *Evaluation of Audio Source Separation Models Using Hypothesis-Driven Non-Parametric Statistical Methods*, in "European Signal Processing Conference", Budapest, Hungary, EURASIP, August 2016, https://hal.inria.fr/hal-01410176

[49] 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, https://hal.inria.fr/hal-01415759

[50] F.-R. STÖTER, A. LIUTKUS, R. BADEAU, B. EDLER, P. MAGRON. *Common Fate Model for Unison source Separation*, in "41st International Conference on Acoustics, Speech and Signal Processing (ICASSP)", Shanghai, China, Proceedings of the 41st International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, https://hal.archives-ouvertes.fr/hal-01248012

[51] J. TROUVAIN, A. BONNEAU, V. COLOTTE, C. FAUTH, D. FOHR, D. JOUVET, J. JÜGLER, Y. LAPRIE, O. MELLA, B. MÖBIUS, F. ZIMMERER. *The IFCASL Corpus of French and German Non-native and Native Read Speech*, in "LREC'2016, 10th edition of the Language Resources and Evaluation Conference", Portorož, Slovenia, Proceedings LREC'2016, May 2016, https://hal.inria.fr/hal-01293935

[52] F. ZIMMERER, A. BONNEAU, B. ANDREEVA. *Influence of L1 prominence on L2 production: French and German speakers*, in "Speech Prosody 2016", Boston, United States, May 2016, vol. 2016, pp. 370 - 374 [*DOI :* 10.21437/SPEECHPROSODY.2016-76], https://hal.inria.fr/hal-01399974

National Conferences with Proceedings

[53] B. ELIE, Y. LAPRIE, P.-A. VUISSOZ. *Acquisition temps-réel de données articulatoires par IRM : application à la synthèse par copie*, in "13ème Congrès Français d'Acoustique (CFA 2016)", Le Mans, France, SFA, April 2016, https://hal.archives-ouvertes.fr/hal-01314313

Conferences without Proceedings

[54] F. ZIMMERER, J. TROUVAIN, A. BONNEAU. *Methods of investigating vowel interferences of French learners of German*, in "New Sounds 2016", Aarhus, Denmark, June 2016, https://hal.inria.fr/hal-01400005

Scientific Books (or Scientific Book chapters)

[55] 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, October 2016, https://hal.inria.fr/hal-01383263

[56] M. CADOT. *Recoder les variables pour obtenir un modèle implicatif optimal*, in "L'Analyse Statisqtique Implicative", R. GRAS (editor), Cépaduès, December 2016, https://hal.archives-ouvertes.fr/hal-01398229

Research Reports

[57] P. MAGRON, R. BADEAU, A. LIUTKUS. *Generalized Wiener filtering for positive alpha-stable random variables*, Télécom ParisTech, June 2016, https://hal.archives-ouvertes.fr/hal-01340797

[58] G. SARGENT, F. BIMBOT, E. VINCENT. *Supplementary material to the article: Estimating the structural segmentation of popular music pieces under regularity constraints*, IRISA-Inria, Campus de Beaulieu, 35042 Rennes cedex ; Inria Nancy, équipe Multispeech, September 2016, https://hal.inria.fr/hal-01368683

Scientific Popularization

[59] A. LIUTKUS, E. VINCENT. *Démixer la musique*, in "Interstices", January 2016, https://hal.inria.fr/hal-01350450

[60] A. PIQUARD-KIPFFER. *Faire voir une histoire : Louis et son incroyable chien Noisette*, in "Les Cahiers Pédagogiques", February 2016, vol. Hors série numérique N°42, 7 p. , https://hal.inria.fr/hal-01191878

Patents and standards

[61] S. OUNI, G. GRIS. *Dispositif de traitement d'image*, January 2016, n^{o} 15 52058, Le rapport de recherche reconnait la brevetabilité, https://hal.inria.fr/hal-01294028

Other Publications

[62] B. DUMORTIER, E. VINCENT, M. DEACONU, P. CORNU. *Efficient optimisation of wind power under acoustic constraints*, November 2016, working paper or preprint, https://hal.inria.fr/hal-01393125

[63] B. ELIE, Y. LAPRIE. *Acoustic impact of the glottal chink on the production of fricatives: A numerical study*, December 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01423206

[64] A. PIQUARD-KIPFFER, T. LÉONOVA. *Parcours scolaire de 166 dysphasiques et/ou dyslexiques-dysorthographiques âgés de 6 à 20 ans en situation de handicap : Schooling experiences of 166 dysphasic or dyslexics-dysorthographic children, aged from 6 to 20 in a handicap situation*, July 2016, working paper or preprint, https://hal.inria.fr/hal-01402986

[65] A. PIQUARD-KIPFFER, O. MELLA, J. MIRANDA, D. JOUVET, L. OROSANU. *Terminal portable de communication et affichage de la reconnaissance vocale. Enjeux et rapports à l'écrit. Étude préliminaire auprès d'adultes déficients auditifs*, March 2016, 15p. p. , In M.Frisch (Eds) Le réseau Idéki : Didactiques, métiers de l'humain et Intelligence collective. Nouveaux espaces et dispositifs en question. Nouveaux horizons en éducation, formation et en recherche. L'harmattan, Collection I.D, https://hal.inria.fr/hal-01239910

References in notes

[66] A. LIUTKUS, R. BADEAU. *Generalized Wiener filtering with fractional power spectrograms*, in "2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", IEEE, 2015, pp. 266–270