IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2015

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:

3.4.8. - 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
6.2.4. - Statistical methods
6.3.1. - Inverse problems
8.2. - Machine learning

Other Research Topics and Application Domains:

9.1.1. - E-learning, MOOC
9.4.1. - Computer science
9.5.8. - Linguistics

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, Researcher, HdR]

Faculty Members

Vincent Colotte [Univ. Lorraine, Associate Professor]
Joseph Di Martino [Univ. Lorraine, Associate Professor]
Irina Illina [Univ. Lorraine, Associate Professor, HdR]
Odile Mella [Univ. Lorraine, Associate Professor]
Slim Ouni [Univ. Lorraine, Associate Professor, HdR]
Agnès Piquard-Kipffer [ESPE (École Supérieure du Professorat et de l'Éducation), Univ. Lorraine, Associate Professor]

Engineers

Ilef Ben Farhat [CNRS]
Julie Busset [CNRS, until May 2015]

Antoine Chemardin [CNRS]
Sara Dahmani [Inria]
Valérian Girard [Univ. Lorraine, from May 2015]
Aghilas Sini [CNRS]
Sunit Sivasankaran [Inria, from Mar. 2015]

PhD Students

Ken Deguernel [Inria, from Mar. 2015]
Baldwin Dumortier [Inria]
Amal Houidhek [École Nationale d'Ingénieurs de Tunis, Tunisie, from Nov. 2015]
Xabier Jaureguiberry [Institut Telecom, Paris, until Jun. 2015]
Aditya Nugraha [Inria]
Luiza Orosanu [Inria]
Imran Sheikh [Univ. Lorraine]
Nathan Souviraà-Labastie [Univ. Rennes 1, until Nov. 2015]
Dung Tran [Inria, until Nov. 2015]
Imene Zangar [École Nationale d'Ingénieurs de Tunis, Tunisie, from Nov. 2015]

Post-Doctoral Fellows

Mohamed Bouallegue [Univ. Lorraine, from Oct. 2015]
Benjamin Elie [Inria, until Feb. 2015; CNRS, since Feb. 2015]
Thibaut Fux [Inria, until Aug. 2015]
Sucheta Ghosh [Inria, from Sep. 2015]
Emad Girgis [Inria, until Jun. 2015]
Juan Andres Morales Cordovilla [Inria, from Mar. 2015]

Visiting Scientists

Andrea Bandini [Univ. of Bologna, Italy, until Apr. 2015]
Pablo Antonio Cabanas Molero [Univ. of Jaén, Spain, from May 2015 until Jul. 2015]

Administrative Assistants

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

Others

Adrien Anxionnat [Univ. Lorraine, from May until Jul. 2015]
Soumaya Azzi [Polytech Clermont-Ferrand, from May until Sep. 2015]
Abdelmalek Ben Ali [École Centrale de Nantes, from Mar. until Aug. 2015]
Imen Ben Othmane [ESTI, Univ. de Carthage, Tunisia, from May until Jul. 2015]
Quentin Claudel [Univ. Lorraine, from Apr. until Jun. 2015]
Jen-Yu Liu [National Taiwan University, Taiwan, from Apr. until Sep. 2015]
Yann Prono [Univ. Lorraine, from Apr. until Jun. 2015]
Thomas Trompette [Univ. Lorraine, from May until Jun. 2015]
Julie Bonmatin [Univ. Lorraine, from Mar. until Aug. 2015]
Freha Boumazza [Uuiv. Hassiba Benbouali de Chlef, Algérie, from Jan. until Apr. 2015]
Siddharth Dalmia [Birla Institute of Technology and Science, India, from May. until Jul. 2015]

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 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 providing the availability of corresponding speech corpora.

3. Research Program

3.1. Introduction

MULTISPEECH is structured along three research directions that are associated to the previously described challenges: explicit modeling of speech, statistical modeling of speech, and uncertainty in speech processing.

3.2. 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.

3.2.1. 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.

3.2.2. 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 was 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 adaption of the system to available data and to various conditions.

3.2.3. 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 feedbacks 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.

3.3. Statistical Modeling of Speech

Whereas the first research direction deals with the physical aspects of speech and its explicit modeling, this second research direction is concerned by investigating statistical models for speech data. Acoustic models are used to represent the pronunciation of the sounds or other acoustic events such as noises. 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.

3.3.1. 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, that will integrate into the recently developed FASST [6] and KAM software frameworks.

3.3.2. 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 the 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 prononciation 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 disfluences and of spontaneous speech prononciation 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 miss-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.

3.3.3. Speech generation by statistical methods

Voice conversion consists in building a function that transforms a given voice into another one. MULTISPEECH applies voice conversion techniques to enhance pathological voices that result from vocal folds problems, especially esophageal voice or pathological whispered voice. In addition to the statistical aspects of the voice conversion approaches, signal processing is critical for good quality speech output. As the fundamental frequency is chaotic in the case of esophageal speech, the excitation spectrum must be predicted or corrected. Voice conversion techniques are also of interest for text-to-speech synthesis systems as they aim at making possible the generation of new voice corpora (other kind of voice, or same voice with different kind of emotion). Also, in the context of acoustic feedback in foreign language learning, voice modification approaches will be 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.

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 speaker), however, the quality is not as good as that of the concatenation-based speech synthesis. MULTISPEECH will focus on an 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.

3.4. Uncertainty Estimation and Exploitation in Speech Processing

This axis focuses on the uncertainty associated to 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 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 on the results.

3.4.1. 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 on 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 is 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 on the acoustic model parameters and the acoustic scores themselves.

3.4.2. 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 on 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.

3.4.3. Uncertainty and prosody

Prosody information is also investigated as a means for structuring speech data (determining sentence boundaries, punctuation. . .) possibly in addition with 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 on the duration of the phones (see uncertainty on 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

Speech technologies provide ways of helping people in handicapped situations or of improving their autonomy. An application is related to the tuning of speech recognition technology for providing a means of communication between a speaking person and a hard-of-hearing or a deaf person, through an adequate display of the recognized words and syllables, which takes also into account the reliability of the recognized items.

Another application aims at improving pathological voices. In this context, the goal is typically to transform the pathological voice signal in order to make it more intelligible. Ongoing work deals with esophageal voices, i.e., substituted voice learned by a laryngectomized patient who has lost his/her vocal cords after surgery. Voice conversion techniques will be studied further to enhance such voice signals, in order to produce clean and intelligible speech signals in replacement of the pathological voice.

A third application aims at improving the autonomy of elderly or disabled people, and fit with smartrooms. 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 elder 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 out of 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 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 2nd among 9 teams for the ”Professionally produced music recordings” task of the 2015 Signal Separation Evaluation Campaign (SiSEC) [75].

We ranked 4th among 25 teams and as the best European team for the 3rd CHiME Speech Separation and Recognition Challenge [55].

5.1.1. Awards

Baldwin Dumortier received the best poster prize at EWEA 2015 (European Wind Energy Association 2015 Annual Event) [31].

Best paper award at SIIE 2015 (6th International Conference on Information Systems and Economic Intelligence) [34].

BEST PAPER AWARD:

[34]
D. FOHR, I. ILLINA. *Neural Networks for Proper Name Retrieval in the Framework of Automatic Speech Recognition*, in "IEEE International Conference on Information Systems and Economic Intelligence", hammamet, Tunisia, 2015, https://hal.archives-ouvertes.fr/hal-01184957

6. New Software and Platforms

6.1. ANTS - Automatic News Transcription System

FUNCTIONAL DESCRIPTION: ANTS 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 (Julius, Sphinx or Kaldi). Further processing passes are run in order to apply unsupervised adaptation processes on the features and/or on the model parameters, or to use Speaker Adaptive Training based models. Latest version include DNN (Deep Neural Network) acoustic modeling.

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

6.2. ASTALI - Automatic 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 potential 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. The web application makes the service easy to use, without requiring any software downloading. Also, the software is currently under integration in the EQUIPEX ORTOLANG platform.

- Participants: Dominique Fohr, Odile Mella, Antoine Chemardin and Denis Jouvet
- Contact: Dominique Fohr
- URL: http://astali.loria.fr/

6.3. JCorpusRecorder

FUNCTIONAL DESCRIPTION: JCorpusRecorder is a software for the recording of audio corpora. It provides an easy tool to record with a microphone. The audio input gain is controlled during the recording. From a list of sentences, the output is a set of wav files automatically renamed according to textual information given in input (nationality, speaker language, gender, ...). An easy to use tagging allows for displaying a textual/visual/audio context for guiding the speaker, along with the text of the sentence to pronounce. Several text encodings are enabled (allowing for instance Chinese texts). The sentences can be presented in a random order. The last version can record up to 8 synchronous channels (8 channels under Linux and 2 channels under Windows). The software is developed in Java, and is currently used for the recording of sentences in several projects.

- Contact: Vincent Colotte

6.4. 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. Thus JSnoori currently focuses the calculation of F0, the forced alignment of non native English uttered by French speakers and the correction of prosody parameters (F0, rhythm and energy). Several tools have been incorporated to segment and annotate speech. A complete phonetic keyboard is available, several levels of annotation can be used (phonemes, syllables and words) and forced alignment can exploit pronunciation variants. In addition, JSnoori offers real time F0 calculation which can be useful from a pedagogical point of view. Besides the traditional graphic interface, JSnoori can now be used via scripts written in Jython.

- Participants: Yves Laprie, Slim Ouni, Julie Busset, Aghilas Sini and Ilef Ben Farhat
- Contact: Yves Laprie
- URL: http://www.loria.fr/~laprie/WinSnoori/

6.5. 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 coils 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. The last version of VisArtico can handle multimodal data, not articulatory data only. In fact, it is possible to visualize motion capture data from Vicon or kinect-like systems (PrimeSense and RealSense). It is possible to generate video from the visualized trajectories. A derived version from VisArtico is also used in the ADT Plavis as a tool to visualize and process the audiovisual data. The software is used by more than 170 researchers around the world.

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

6.6. 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. 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 constructing a velum model and incorporating it into the area functions.

- Contact: Yves Laprie

6.7. Platform EMA - Electromagnetic Articulography Acquisition

FUNCTIONAL DESCRIPTION: Since the purchase of the articulograph AG500 in 2007, we have built a strong experience with respect to the acquisition technique and we have developed an acquisition protocol. The platform has been improved by acquiring the latest articulograph AG501 funded by the EQUIPEX ORTOLANG project. The AG501 allows tracking the movement of 24 sensors at reasonable high frequency (250Hz) up to a very high frequency (1250Hz). In addition, we have continued improving VisArtico (cf. 6.5), a powerful tool to visualize articulatory data acquired using an articulograph. This year we have used the system to acquire articulatory data for the tongue, jaw and lips to study stuttering speech disorder (informal collaboration with F. Hirsch, Praxiling (UMR 5267)). We have also used the EMA platform to acquire motion capture data for the lips, to be used in the context of audiovisual speech synthesis [82].

- Contact: Slim Ouni

6.8. Platform MRI - Magnetic Resonance Imaging

KEYWORDS: Health - Medical imaging

FUNCTIONAL DESCRIPTION: Magnetic Resonance Imaging (MRI) takes an increasing place in the investigation of speech production because it provides a complete geometrical information of the vocal tract. We thus initiated a cooperation with the IADI laboratory (Imagerie Adaptive Diagnostique et Interventionnelle) at Nancy Hospital, which studies in particular magnetic resonance imaging. This year the work focused on the development of compressed sensing algorithms and the reconstruction of good quality images to acquire cineMRI at a sampling rate between 25 and 60 Hz. The algorithms were implemented on the 3T GE research MRI machine of the Nancy Hospital.

- Contact: Yves Laprie

7. New Results

7.1. Explicit Modeling of Speech Production and Perception

Participants: Yves Laprie, Slim Ouni, Vincent Colotte, Anne Bonneau, Agnès Piquard-Kipffer, Emmanuel Vincent, Denis Jouvet, Julie Busset, Benjamin Elie, Andrea Bandini, Ilef Ben Farhat, Sara Dahmani, Valérian Girard.

7.1.1. Articulatory modeling

7.1.1.1. Acoustic simulations

The acoustic simulation plays a key role in articulatory synthesis since it generates the acoustic signal from the instantaneous geometry of the vocal tract. This year we extended the single-matrix formulation to enable self-oscillation models of vocal folds, including glottal chinks, to be connected to the vocal tract. It also integrates the case of a local division of the main air path into two lateral channels, as it may occur during the production of lateral approximants. Extensions give rise to a reformulation of the acoustic conditions at the glottis, and at the upstream connection of bilateral channels. Numerical simulations validate the simulation framework. In particular the presence of a zero around 4 kHz due to the presence of bilateral channels around both sides of the tongue for the sound /l/ is confirmed by the simulations. These results agree with those obtained via independent techniques. Simulations of static vowels reveal that the behavior of the vocal folds is qualitatively similar whether they are connected to the single-matrix formulation or to the classic reflection type line analog model.

7.1.1.2. Acquisition of articulatory data

Magnetic resonance imaging (MRI) is a technique which provides very good static images of the vocal tract. However, it cannot be used directly to acquire dynamic images of the vocal tract which would enable a better comprehension of articulatory phenomena and the development of better coarticulation models. We thus have a cooperation with the IADI (Imagerie Adaptatative Diagnostique et Interventionnelle) INSERM laboratory in Nancy Hospital intended to develop cineMRI [86], [87] (see. 6.8).

7.1.1.3. Articulatory models

An articulatory model of the velum [66], [65] was developed in order to complete an articulatory model already comprising other articulators. The velum contour was delineated and extracted from a thousand of X-ray images corresponding to short sentences in French. A principal component analysis was applied in order to derive the main deformation modes. The first component corresponds to the opening and comes with a shape modification linked to the apparition of a bulb in the upper part of the velum when it rises. The area function of the oral tract is modified so as to incorporate the velum movements. This model was connected with acoustic simulations in order to synthesize sentences containing French nasal vowels and consonants.

7.1.2. Expressive acoustic-visual synthesis

During this year, we have focused on the development of the acquisition infrastructure necessary to acquire audiovisual data. Mainly, we have developed several methods that allow acquiring acoustic and visual data synchronously. The visual data can originate from the Articulograph, Vicon or Intel RealSense devices. This heterogeneity of the data needs developing techniques to merge precisely the data in one unique reference. Synchronization techniques have also been developed for this purpose. We have evaluated the precision of the acquisition of such systems [61]. The combination of more than one motion capture technique aims to use the best quality data for each part of the face: (1) EMA (articulograph) for the lips, to have high precise measurement of the shape of the mouth that is related to speech and (2) kinect-like or Vicon system for the upper part of the face, that model mainly expressions.

We have acquired a small expressive audiovisual speech corpus of two actors: based on motion capture data (Vicon) and acoustic data. The content of the corpus is composed of six basic emotions (joy, sadness, anger, surprise, disgust and fear). This corpus will be used to investigate the characterization of emotions in audiovisual speech in the visual space and in the acoustic space.

We have also developed an algorithm to animate the 3D model of human face from a limited number of markers. The animation is very efficient and provides realistic animation results [82]. The 3D face will be used with the audiovisual system.

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 investigated the schooling of a population of 166 students from primary to intermediate and secondary schools. These children and teenagers had specific language impairment: SLI (severe language impairment), dyslexia, dysorthographia. Since their childhood, they faced phonemic discrimination, phonological and phonemic analysis difficulties. We observed that they had trouble learning to read and more generally they experienced learning difficulties. Consequently, this lead them to repeat one or more grades, whereas in France, repetition is prohibited within each cycle and very limited between cycles.

7.1.3.2. Analysis of non-native pronunciations

Thanks to the detailed manual annotation of the French-German learner corpus that was carried out at the phonetic level in the IFCASL project (cf. 9.1.2), it was possible to investigate non-native pronunciation variants. The analysis revealed that German learners of French have most problems with obstruents in word-final position, whereas French learners of German show complex interferences with the vowel contrasts for length and quality [41]. Also, the correct pronunciation rate of the sounds, for several phonetic classes, was analyzed with respect to the learner's level, and compared to native pronunciations. One outcome is that different sound classes show different correct rates over the proficiency levels; and, for the German data, the frequently occurring syllabic [=n] is a prime indicator of the proficiency level.

We analyzed the realizations of French voiced fricatives by German non-native and French native speakers, in final position of an accentual group, a position where German fricatives are devoiced [27], [28]. Three speaker levels (from beginners to advanced) and different boundary types (depending on whether the fricative is followed by a pause, a schwa, or is directly followed by the first phoneme of the subsequent group) were considered. A set of cues, among which periodicity and fricative duration, have been analyzed. Results argue in favor of an influence of L1 (German) final devoicing on non-native realizations and show a strong interdependence between voicing, speakers' level, prosodic boundaries. The influence of orthography also strongly influenced voicing results.

We also investigated the realization of the short/long German contrast by French learners through three methods [60]. All these methods - phonetic annotation, perceptual experiment and acoustic analysis - used the same database (the IFCASL corpus). Depending on the method the results shed light on slightly different aspects of the same process, the interference of the French phonetic and phonological systems on the production of the German L2 vowels. Whereas the first method (phonetic annotation) revealed that especially rounded vowels are problematic in the long/short distinction, we could show with the second method (a perceptual experiment) that particularly the [o:]/[O] distinction seems to be hard to produce for French learners. The third method (an acoustical analysis) corroborated this finding and added acoustic details on duration and formants. The results of the studies can be used to create individualized training and feedback for foreign language learners, aimed at reducing their accent in L2.

7.2. Statistical Modeling of Speech

Participants: Antoine Liutkus, Emmanuel Vincent, Irina Illina, Dominique Fohr, Denis Jouvet, Joseph Di Martino, Vincent Colotte, Ken Deguernel, Amal Houidhek, Xabier Jaureguiberry, Aditya Nugraha, Luiza Orosanu, Imran Sheikh, Nathan Souviraà-Labastie, Dung Tran, Imene Zangar, Mohamed Bouallegue, Thibaut Fux, Emad Girgis, Juan Andres Morales Cordovilla, Sunit Sivasankaran, Freha Boumazza.

7.2.1. Source separation

Audio source separation is an inverse problem, which requires the user to guide the separation process using prior models for the source spectra and their spatial covariance matrices. We studied the impact of deterministic subspace constraints [14] over the spatial covariance matrices and pursued our work on the separation of multichannel mixtures guided by multiple, deformed reference signals such as repeated excerpts of the same music or repeated versions of the same sentence uttered by different speakers [17], [56]. Other models we have been working on include those based on local regularities of the spectral representations of musical sources (KAM, [52], [43], [51]). We also validated the positive impact of speech enhancement based on the FASST toolbox on speaker recognition [53].

As a new research direction, we extended the Gaussian framework for source separation to the family of α-stable stochastic processes [42]. This extension notably opens the path to new and robust parameters estimation algorithms for source separation [16], [67], that should be less prone to local minima. Current research notably comprises multichannel stable processes.

In parallel, we started yet another research track on the use of deep learning for source separation [24]. We proposed a new multichannel enhancement technique that 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 [75]. The model parameters are alternately estimated in an expectation-maximization (EM) fashion. We used this technique for music separation and speech enhancement in the context of the 2015 Signal Separation Evaluation Campaign (SiSEC) and the 3rd CHiME Speech Separation and Recognition Challenge, respectively [55]. 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 [81].

Finally, we pursued our long-lasting efforts on the evaluation of audio source separation by co-organizing the 2015 Signal Separation Evaluation Campaign (SiSEC) [69] and writing a position paper about the scaling up of dataset sizes [21].

The ANR young researcher project KAMoulox (2016-2019 - cf. 9.1.5), that has just been accepted will deal with large audio archives, and more precisely with the "Archives du CNRS — Musée de l'homme" that gather a large set of old and noisy audio recordings (cf. 4.4). 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.

7.2.2. Acoustic modeling

We explored the use of an auxiliary function technique for fast training of neural networks [58]. We did not apply this technique to deep neural network acoustic models yet.

In the framework of using speech recognition for helping communication with deaf or hard-of-hearing people, robustness of the acoustic modeling was investigated. Studies were related to improving robustness with respect to speech signal level and environment noise through multicondition training and enhanced set of acoustic features (noise robust features or standard features after spectral noise subtraction) [37].

7.2.3. Linguistic modeling

7.2.3.1. Out-of-vocabulary proper name retrieval

Recognition of proper names (PN) is a challenging task in information retrieval in large audio/video databases. Proper names are semantically rich and are usually key to understanding the information contained in a document. Within the ContNomina project (cf. 9.1.3), we focus on increasing the vocabulary coverage of a speech transcription system by automatically retrieving proper names from contemporary text documents. We proposed methods that dynamically augment the automatic speech recognition system vocabulary, using lexical and temporal features in diachronic documents (documents that evolve over the time). Our work uses temporal context modeling to capture the lexical information surrounding proper names so as to retrieve out-of-vocabulary (OOV) proper names and increase the automatic speech recognition vocabulary.

We proposed new methods to retrieve OOV PNs relevant to an audio news document by using probabilistic topic models. We addressed retrieval of rare OOV PNs, which further improves the recall. Our proposed lexical context model improves the mean average precision of OOV PN retrieval [62]. We also proposed a two step approach for recognition of OOV PNs in an audio document. The first step retrieves OOV PNs relevant to an audio document using probabilistic topic models; and the second step uses a phonetic search for the target OOV PNs using a k-differences approximate string matching algorithm [63]. In [64], we discuss two specific phenomena, word frequency bias and loss of specificity, which affect the retrieval of OOV PNs using Latent Dirichlet Allocation (LDA) topic models. We studied different entity-topic models, which are extensions of LDA designed to learn relations between words, topics and PNs. We showed that our proposed methods of rare OOV PN and lexical context re-ranking improve the recall and the mean average precision for the LDA and the entity-topic models.

For OOV retrieval, we proposed the continuous space word representation using neural networks. This continuous vector representation (word embeddings) is learned from large amounts of unstructured text data. To model semantic and lexical context of proper names, different strategies of local context modeling were proposed [34], [33]. We studied OOV PN retrieval using temporal versus topic context modeling, different word representation spaces for word-level and document-level context modeling, and combinations of retrieval results [38]. We extended the previously proposed neural networks for word embedding models: the word vector representation proposed by Mikolov is enriched by an additional non-linear transformation. This model allows to better take into account lexical and semantic word relationships [39].

7.2.3.2. Adding words in a language model

A novel approach was proposed to add some new words in an existing n-gram language model, based on a similarity measure between the new words to be added and words already present in the language model [47]. Based on a small set of sentences containing the new words and on a set of n-gram counts containing the known words (known for the current language model), we search for known words which have the most similar neighbor distribution (of the few preceding and few following neighbor words) to the new words. The similar words are determined through the computation of KL divergences on the distribution of neighbor words. The n-gram parameter values associated to the similar words are then used to define the n-gram parameter values of the new words.

7.2.3.3. Selecting data for training a language model

Large vocabulary language models for speech transcription are usually trained from large amounts of textual data collected from various sources, which are more or less related to the target task. Selecting data that matches the target task was investigated in this context [46], this leads to a small reduction of the perplexity, and a smaller size of the resulting language model.

7.2.3.4. 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 pursued our pioneering work on music language modeling, with a particular focus on the modeling of long-term structure [12]. We also assessed the applicability of our prior work on joint modeling of note and chord sequences to new corpora of improvised jazz music, with the difficulty that these corpora are very small.

7.2.4. Speech generation by statistical methods

7.2.4.1. Pathological voice transformation

With respect to pathological voice processing, a competing approach to signal processing techniques consists in recognizing the pathological voice in order to transform it in a text version that can be re-synthesized. Such an approach is currently being experimented, and preliminary results are quite encouraging [15].

7.2.4.2. HMM-based synthesis

This year, we started working on HMM-based synthesis in the framework of a CMCU PHC project with ENIT (Engineer school at Tunis-Tunisia; cf. 9.3.2.2). Two topics will be explored by two PhD students. The first topic deals with the building of an Arabic corpora along with the analysis of linguistic features which are relevant for the HMM-based synthesis of the Arabic language. The second topic deals with improving the quality of the HMM-based synthesis system. In parallel, we started applying the HTS system (HMM-based Speech Synthesis System) to the French language.

7.3. Uncertainty Estimation and Exploitation in Speech Processing

Participants: Emmanuel Vincent, Odile Mella, Dominique Fohr, Denis Jouvet, Agnès Piquard-Kipffer, Baldwin Dumortier, Luiza Orosanu, Dung Tran, Sucheta Ghosh, Antoine Chemardin, Aghilas Sini.

7.3.1. Uncertainty and acoustic modeling

7.3.1.1. Noise-robust speech recognition

In many real-world conditions, the target speech signal overlaps with noise and some distortion remains after speech enhancement. 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 two successful editions in 2011 and 2013, we organized the third edition in 2015 [25].

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 made a conceptual breakthrough by proposing to learn the estimator from data using a non-parametric estimator and discriminative training [18], [59]. With GMM-HMM acoustic models, we obtained on the order of 30% relative word error rate reduction with respect to conventional decoding (without uncertainty), that is about twice as much as the reduction achieved by the best single uncertainty estimator. We also started working on the propagation of uncertainty in deep neural network acoustic models [19] and on its use for noise-robust speaker recognition [54].

7.3.1.2. Other applications

Besides the above applications, we started exploring applications of uncertainty modeling to robot audition [23] and control of wind turbines [31]. 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. 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.

7.3.2. Uncertainty and speech recognition

In the framework of using speech recognition for helping communication with deaf or hard-of-hearing people in the FUI project RAPSODIE (cf. 9.1.7), the best way for displaying the speech transcription results has been investigated. To our knowledge there is no suitable, validated and currently available display of the output of automatic speech recognizer for hard-of-hearing persons, in terms of size, colors and choice of the written symbols. The difficulty comes from the fact that speech transcription results contain recognition errors, which may impact the understanding process. Although the speech recognition system does not know the errors it makes, through the computation of confidence measures, the speech recognizer estimates if a word or a syllable is rather correctly recognized or not; hence such information can be used to adjust the display of the transcription results. Different ways were investigated for displaying the speech recognition results which take also into account the reliability of the recognized items. In this qualitative study, 10 persons have been interviewed to find the best way of displaying the speech transcription results. All the participants are deaf with different levels of hearing loss and various modes of communication [50].

7.3.3. Uncertainty and phonetic segmentation

Within the framework of the IFCASL project (cf. 9.1.2), a speech corpus of native and non-native speech for the French-German language pair was designed and recorded. Besides beeing used for analyzing non-native phenomena (cf. 7.1.3.2), this corpus will be used for developing and assessing automatic algorithms that will provide diagnosis on the learner mispronunciations [78]. Therefore, the automatic alignments of the audio files corresponding to the French and German speakers uttering French sentences (4100 audio files) were manually checked and corrected by a group of seven French annotators (the German data were handled by the German partner). We analyzed with CoALT the inter-annotator agreement with respect to an expert annotator for boundary shifts, insertions and deletions as well as devoicing diacritic [45]. The accuracy of the phone boundaries on non-native speech were investigated with respect to the HMM acoustic models used. The best performance (smallest amount of non-native phone segments whose boundaries are shifted by more than 20 ms compared to the manual boundaries) was obtained by combining each French native HMM model with an automatically selected German native HMM model [35].

Within the ANR ORFEO project (cf. 9.1.6), we addressed the problem of the alignment of spontaneous speech. The audio files processed in the ORFEO project were recorded under various conditions with a large SNR range and contain extra speech phenomena and overlapping speech. We trained several sets of acoustic models and tested different methods to adapt them to the various audio files [36]. Moreover in the framework of the EQUIPEX ORTOLANG (cf. 9.1.1), a web application, ASTALI (cf. 6.2), was developped in order to align 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).

In conventional speech-text alignments, a 10 ms frame shift is usually used for the acoustic analysis which leads to a minimum duration of 30 ms for each phone segment. Such duration constraint may not fit with actual sound duration in fast speaking rate. To overcome such contraint, a 5 ms frame shift can be used. Statistics on pronunciations variants estimated on large speech corpora have shown that when the conventional 10 ms frame shift is used, the frequency of the longest pronunciation variants gets underestimated [26]. Moreover, the analysis of some pronunciation variant frequencies have shown that some final consonantal cluster completely disappear at high speaking rates [40].

7.3.4. Uncertainty and prosody

Detection of sentence modality (question vs. affirmation) has been investigated using linguistic and prosodic features. Best results are achieved when the classifier uses all the available information [48], that is both linguistic and prosodic features. A detailed analysis has also shown that small errors in the determination of the sentence boundaries are not critical [49].

Speech-text alignments have been used to extract speech segments containing words and expressions that can be used either as normal lexical words or as discourse particles (as for example *quoi*, *voilà*, ...). The prosodic features for these words and expressions were extracted and analyzed [30]; automatic identification of the word function (discourse particle or not) from these prosodic features was also investigated.

In the context of the EQUIPEX ORTOLANG (cf. 9.1.1), several algorithms for computing the fundamental frequency have been implemented in the JSnoori software. These features can be computed directly from the GUI interface or through Python scripts. Future work will focus on improving the quality and robustness of the fundamental frequency estimation, and on determining the reliability of the estimations.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. MAIA

Company: Studio MAIA

Duration: September 2014 - August 2015

Supported by: Bpifrance

Abstract: A pre-study contract was signed to investigate speech processing tools that could eventually be transferred as plugins for audio mixing software. Prosody modification, noise reduction, and voice conversion are of special interest.

8.1.2. 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. National Initiatives

9.1.1. EQUIPEX ORTOLANG

Project acronym: ORTOLANG [1]

Project title: Open Resources and TOols for LANGuage

Duration: September 2012 - May 2016 (phase I, signed in January 2013)

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 is 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 text-speech alignment and speech visualization tools.

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

9.1.2. ANR-DFG IFCASL

Project acronym: IFCASL

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

Duration: March 2013 - February 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.

The work involved the design and recording of a French-German learner corpus. French speakers were recorded in Nancy, wheras 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.

9.1.3. ANR ContNomina

Project acronym: ContNomina

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

Duration: February 2013 - July 2016

Coordinator: Irina Illina, MULTISPEECH

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).

9.1.4. ANR DYCI2

Project acronym: DYCI2 [2]

Project title: Creative Dynamics of Improvised Interaction

Duration: March 2015 - February 2018 (signed in October 2014)

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.

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

9.1.5. ANR JCJC KAMoulox

Project acronym: KAMoulox

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

Duration: January 2016 - January 2019 (signed in October 2015)

Coordinator: Antoine Liutkus, MULTISPEECH

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.

9.1.6. 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 are 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).

9.1.7. FUI RAPSODIE

Project acronym: RAPSODIE [4]

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

Duration: March 2012 - February 2016 (signed in December 2012)

Coordinator: eRocca (Mieussy, Haute-Savoie)

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

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

In this project, MULTISPEECH is involved for optimizing the speech recognition models for the envisaged task, and contributes also to 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.

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

[4] http://erocca.com/rapsodie

9.1.8. 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.

9.1.9. ADT Plavis

Project acronym: Plavis

Project title: Platform for acquisition and audiovisual speech synthesis

Duration: January 2015 - December 2016

Coordinator: Vincent Colotte, MULTISPEECH

Abstract: The objective of this project is to develop a platform acquisition and audiovisual synthesis system (3D animation of the face synchronously with audio). The main purpose is to build a comprehensive platform for acquisition and processing of audio-visual corpus (selection, acquisition and acoustic processing, 3D visual processing and linguistic processing). The acquisition is performed using a motion capture system (Kinect-like) or from Vicon system or EMA system. We also propose to develop a 3D audiovisual synthesis system text to audio and 3D information of a talking head. The system will incorporate an animation module of the talking head to reconstruct the face animated with audio. During the first year of the project, we are setting up and testing the acquisition techniques that will be used. We have developed several tools to acquire the audiovisual data and to process it. A synchronization step was developed.

9.1.10. ADT VisArtico

Project acronym: VisArtico

Project title: Software for Processing, analysis and articulatory data visualization

Duration: November 2013 - October 2015

Coordinator: Slim Ouni, MULTISPEECH

Abstract: The Technological Development Action (ADT) Inria Visartico aims at developing and improving VisArtico, an articulatory vizualisation software (see 6.5). In addition to improving the basic functionalities, several articulatory analysis and processing tools are being integrated.

9.1.11. 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: S. Ouni, MULTISPEECH

Cofunded by Inria and Région Lorraine

Abstract: The main objective of this project is 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 signal). The main purpose is to acquire, process and analyze the corpus and to study the expressiveness; the results will be used for the expressive audiovisual speech synthesis system.

9.1.12. LORIA exploratory project

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

Duration: June 2015 - May 2016

Coordinator: S. Ouni, MULTISPEECH

Abstract : The aim of this project is the study of the various mechanisms involved in multimodal human communication that can be oral, visual, gestural and tactile. This project focuses 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. We will set up and integrate hardware and software acquisition. Thereafter, we will acquire and structure the multimodal data.

9.2. European Initiatives

9.2.1. Collaborations with major european organizations

Jon Barker: University of Sheffield (UK)

Robust speech recognition [25].

9.3. International Initiatives

9.3.1. Inria international partners

9.3.1.1. Informal international partners

Nobutaka Ono: National Institute for Informatics (NII, Tokyo, Japan)

Machine learning and source separation [14], [58], [69] (former Inria associate team).

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

Source separation [19], [21], [24].

Bryan Pardo, Northwestern University (Evanston, IL, USA)

Audio source separation [52].

Derry Fitzgerald, Nimbus Center, Cork Institute of Technology (Ireland)

Audio source separation [43], [67].

Taylan Cemgil, Bosphorus University (Istambul, Turkey)

Multimodal data analysis [44] and source separation [16].

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

Robust speaker recognition [53], [54].

9.3.2. Participation in other international programs

9.3.2.1. 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

National Coordinators: Juan Carlos GÓMEZ (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.

9.3.2.2. 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 corpora, 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.

9.4. International Research Visitors

9.4.1. Visits of international scientists

9.4.1.1. Internships

Liu Jen-Yu

Date: Apr 2015 - Sep 2015

Institution: NTU (Taiwan)

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific events organisation

10.1.1.1. General chair, scientific chair

General co-chair, 3rd CHiME Speech Separation and Recognition Challenge, Scottsdale, USA, December 2015 (E. Vincent)

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)

Co-chair, special session on Audio for Robots – Robots for Audio, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (E. Vincent)

Co-chair, Workshop-satellite event of ICPhS2015. Phonetic Learner Corpora. Glasgow, GB, 08/12/15 (A. Bonneau)

10.1.1.2. Special session organisation

Special session Structuring Multimedia Streams, International Conference on Information Systems & Economic Intelligence, Hammamet, Tunisia, February 2015 (I. Illina, D. Fohr)

Co-organiser Workshop-satellite event of ICPhS2015. Phonetic Learner Corpora. Glasgow, GB, 08/12/15 (A. Bonneau)

10.1.1.3. Member of organizing committees

Member of the organizing committee, FAAVSP - The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (S. Ouni)

10.1.2. Scientific events selection

10.1.2.1. Chair of conference program committees

Program chair, 12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), Liberec, Czech Republic, August 2015 (E. Vincent) [71]

10.1.2.2. Member of conference program committees

Area chair for Analysis of Speech and Audio Signal, INTERSPEECH'2015 (D. Jouvet)

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

IROS'2015 - 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (D. Jouvet)

ICNLSP'2015 - International Conference on Natural Language and Speech Processing (D. Jouvet)

FAAVSP - The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (S. Ouni)

10.1.2.3. Reviewer

EUSIPCO'2015 - European Signal Processing Conference (D. Jouvet)

ICNLSP'2015 - International Conference on Natural Language and Speech Processing (D. Jouvet)

INTERSPEECH'2015 (I. Illina, A. Bonneau, S. Ouni, Y. Laprie)

ICASSP'2016 - 41th International Conference on Speech, Acoustic and Signal Processing (D. Jouvet, S. Ouni, Y. Laprie)

FAAVSP - The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (S. Ouni)

10.1.3. Journal

10.1.3.1. Member of editorial boards

Speech Communication (D. Jouvet)

Traitement du signal (E. Vincent)

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

Eurasip Journal on Audio, speech and music processing (Y. Laprie)

10.1.3.2. Reviewer - reviewing activities

IEEE Transactions on Audio, Speech, and Language Processing (D. Jouvet)

Journal of Phonetics (Y. Laprie)

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

IEEE Signal Processing Letters (Y. Laprie)

10.1.4. Invited talks

Keynote, 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (E. Vincent) [85]

Lecture, LVA/ICA 2015 Summer School on Latent Variable Analysis and Signal Separation (E. Vincent) [84]

Invited talk, Expressive Virtual Actors workshop, Nov 2015, Grenoble, France (S. Ouni) [22]

Reconnaissance de la parole, application aux personnes sourdes et malentendantes. Inria Scientific Days 2015. Nancy-Inria Grand-Est, June 19th 2015 (A. Piquard-Kipffer)

Dyslexie-dysorthographie et situation de handicap. Journées des dys, AFON. Vandœuvre-Lès-Nancy, Faculty of Medicine, October 20th 2015 (A. Piquard-Kipffer)

Invited talk, Contribution of the acoustic cues to the non-native accent, Spring ASA meeting Pittsburgh (Y. Laprie) [20]

10.1.5. Leadership within the scientific community

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

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

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

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

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

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

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

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

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

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

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

Licence: Joseph Di Martino, Programming in C 1, 67 hours, L1, University of Lorraine, France

Licence: Joseph Di Martino, Programming in C 2, 85 hours, L2, University of Lorraine, France

Licence: Joseph Di Martino, Programming in C/C++, 45 hours, L3, University of Lorraine, France

Licence: Odile Mella, C2i - Certificat Informatique et Internet, 28h, L1, University of Lorraine, France

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

Licence: Odile Mella, Computer Networking, 81h, L2-L3, University of Lorraine, France

Licence: Agnès Piquard-Kipffer, Psycholinguistics, 30 hours, L1, University of Lorraine, France

Licence: Agnès Piquard-Kipffer, Reading and Writing, 24 hours, L2, Département Orthophonie, University of Lorraine, France

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

Master: Vincent Colotte, Supervising student in research internship, M2, University of Lorraine, France

Master: Yves Laprie, Analysis, Perception and recognition of speech, 30 hours, M1, University of Lorraine,

Master: Odile Mella, Computer Networking, 66h, M1, University of Lorraine, France

Master: Odile Mella, Supervising students in company internship, M2, University of Lorraine, France

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

Master: Agnès Piquard-Kipffer, Dyslexia, 25 hours, Département Orthophonie, University of Lorraine, France

Master: Agnès Piquard-Kipffer, Deaf People and Reading, 9 hours, Département Orthophonie, University of Lorraine, France

Master: Agnès Piquard-Kipffer, Psycholinguistics, 12 hours, Département Orthophonie, University Pierre et Marie Curie-Paris, France

Master: Agnès Piquard-Kipffer, Psychology, 70 hours, ESPE, University of Lorraine, France

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

Master: Agnès Piquard-Kipffer, Psychology, 6 hours, University Blaise Pascal, Clermont-Ferrand, France

Ecole Audioprothèse: Anne Bonneau, Phonétique, 16 hours, Université de Lorraine, France

School of engineers: Vincent Colotte, XML, 20h, Telecom Nancy, France

Doctorat: Piquard-Kipffer Agnès, Language Pathology, 15 hours, EHESP, University of Sorbonne-Paris Cité, France

Adults: Odile Mella, Computer science courses for seconday school teachers (Informatique et Sciences du Numérique courses) (10h), ESPE of Academy Nancy-Metz, University of Lorraine, France

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

10.2.2. Supervision

PhD: Xabier Jaureguiberry, "Fusion pour la séparation de sources audio", Télécom ParisTech, 16 June 2015, Gaël Richard and Emmanuel Vincent [8].

PhD : Dung Tran, "Uncertainty learning for noise robust ASR", University of Lorraine, 20 November 2015, Emmanuel Vincent and Denis Jouvet [11].

PhD: Nathan Souviraà-Labastie, "Détection de motifs audio pour la séparation de sources guidée. Application aux bandes-son de films.", University Rennes 1, 23 November 2015, Frédéric Bimbot and Emmanuel Vincent [10].

PhD : Luiza Orosanu, "Speech recognition for helping communication for deaf or hard-of-hearing people", University of Lorraine, 11 December 2015, Denis Jouvet [9].

PhD in progress: Imran Sheikh, "OOV proper name retrieval", February 2014, Irina Illina 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", November 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", November 2015, Denis Jouvet and Vincent Colotte (France) and Zied Mnasri (Tunisia).

10.2.3. Participation in HDR and PhD juries

Participation in Habilitation thesis Jury for Pavel Král (University of West Bohemia, May 2015), D. Jouvet, reviewer.

Participation in PhD thesis Jury for Grégor Dupuy (Université du Maine, July 2015), D. Jouvet.

Participation in PhD thesis Jury for Sarah Flora Samson Juan (Université de Grenoble, July 2015), D. Jouvet, reviewer.

PhD thesis reviewer for Marc Evrard (Unviversité de Paris Sud, September 2015), Y. Laprie.

PhD thesis reviewer for Heikki Kallasjoki (Aalto University, Finland, October 2015), E. Vincent.

Participation in PhD thesis Jury for Adela Barbulescu (Université de Grenoble, November 2015), S. Ouni.

10.2.4. Participation in other juries

Chairman of Scientific « Baccalauréat », specialty Earth Sciences (Académie de Nancy-Metz and Université de Lorraine, June 2015), A. Piquard-Kipffer.

Participation in PGCE (french CAFIPEMF) Juries for Master Learning Facilitator (DSDEN 54, April, June 2015), A. Piquard-Kipffer.

Participation in the Competitive Entrance Examination into Speech-Language Pathology Departement (University of Lorraine, June 2015), A. Piquard-Kipffer.

10.2.5. Participation to external committees

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

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

Titular member of the National Council of Universities (CNU section 61), E. Vincent.

Member of the "Conseil de secteur IAEM" at the Faculté de sciences et technologie- Université de Lorraine, V. Colotte

Elected member, and member of the board of the "Conseil du pôle AM2I" Université de Lorraine, Y. Laprie

10.2.6. Participation to internal committees

Titular member of the "Comité de Centre Inria", E. Vincent

Member of the Comipers, E. Vincent

Member of the "Comité Espace Transfert", E. Vincent

Member of the "Commission locale de dévelopement durable", D. Fohr

Head of the "Commission des utilisateurs des moyens informatiques" (CUMI), D. Fohr

Member of the "Bureau de la Commission de Mention Informatique", I. Illina

Member of the "Commission de développement technologique", A. Bonneau

Elected member of the "Conseil de Laboratoire" (LORIA), S. Ouni

Appointed member of the "Conseil de Laboratoire" (LORIA), Y. Laprie

Appointed member at "Conseil de la Fédération Charles Hermite", S. Ouni

10.3. Popularization

Demonstration at Forum des métiers, Collège Peguy, Le Chesnay, March 2015 (A. Piquard-Kipffer).

Demonstration at Village Sciences LORIA, April 2015 (S. Ouni, J. Di Martino).

Demonstration at Science & You, Nancy, June 2015 (K. Deguernel, E. Vincent).

Demonstration to students of the Master Erasmus Mundu Program DESEM, Inria, July 2015 (Y. Laprie).

Demonstration at Fête de la Science, Artem, October 2015 (A. Liutkus, N. Souviraà-Labastie, E. Vincent).

Demonstration to students of ENS Cachan, Inria, December 2015 (S. Ouni, E. Vincent).

Les troubles Dys : la dyslexie-dysorthographie. Conf'curieuses, Museum & Aquarium of Nancy, Grand Nancy & University of Lorraine. Nancy, January 15th 2015 (A. Piquard-Kipffer).

“Séparation de sources audio”, Interstices, to appear (A. Liutkus and E. Vincent).

“Loria : un articulographe pour étudier la production de la parole”, Eureka Lorraine, 21/10/2015 (S. Ouni).

11. Bibliography

Major publications by the team in recent years

[1] F. BAHJA, J. DI MARTINO, E. H. IBN ELHAJ, D. ABOUTAJDINE. *An overview of the CATE algorithms for real-time pitch determination*, in "Signal, Image and Video Processing", 2013 [*DOI :* 10.1007/S11760-013-0488-4], https://hal.inria.fr/hal-00831660

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

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

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

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

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

[7] A. PIQUARD-KIPFFER, 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] X. JAUREGUIBERRY. *Fusion for audio source separation*, TELECOM ParisTech ; Inria Nancy, équipe Multispeech, June 2015, https://hal.archives-ouvertes.fr/tel-01189560

[9] L. OROSANU. *Speech recognition as a communication aid for deaf and hearing impaired people*, Université de Lorraine, December 2015, https://hal.inria.fr/tel-01251128

[10] N. SOUVIRAÀ-LABASTIE. *Audio motif spotting for guided source separation. Application to movie soundtracks*, Université de Rennes 1, November 2015, https://hal.inria.fr/tel-01245318

[11] D. T. TRAN. *Uncertainty learning for noise robust ASR*, Universite de Lorraine, November 2015, https://hal.inria.fr/tel-01246481

Articles in International Peer-Reviewed Journals

[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. , To appear - http://mp.ucpress.edu/, https://hal.inria.fr/hal-01188244

[13] M. CHAMI, M. IMMASSI, J. DI MARTINO. *An architectural comparison of signal reconstruction algorithms from short-time Fourier transform magnitude spectra*, in "International Journal of Speech Technology", September 2015, vol. 18, n^o^ 3, 9 p. [*DOI :* 10.1007/s10772-015-9281-9], https://hal.inria.fr/hal-01184625

[14] N. ITO, E. VINCENT, T. NAKATANI, N. ONO, S. ARAKI, S. SAGAYAMA. *Blind suppression of nonstationary diffuse noise based on spatial covariance matrix decomposition*, in "Journal of Signal Processing Systems", May 2015, vol. 79, n^o^ 2, pp. 145-157, https://hal.inria.fr/hal-01020255

[15] O. LACHHAB, J. DI MARTINO, E. H. IBN ELHAJ, A. HAMMOUCH. *A preliminary study on improving the recognition of esophageal speech using a hybrid system based on statistical voice conversion*, in "SpringerPlus", October 2015 [*DOI :* 10.1186/s40064-015-1428-2], https://hal.inria.fr/hal-01221503

[16] U. SIMSEKLI, A. LIUTKUS, T. CEMGIL. *Alpha-Stable Matrix Factorization*, in "IEEE Signal Processing Letters", September 2015, 5 p. , https://hal.inria.fr/hal-01194354

[17] N. SOUVIRAÀ-LABASTIE, A. OLIVERO, E. VINCENT, F. BIMBOT. *Multi-channel audio source separation using multiple deformed references*, in "IEEE Transactions on Audio, Speech and Language Processing", June 2015, vol. 23, n^o^ 11, pp. 1775-1787, https://hal.inria.fr/hal-01070298

[18] D. T. TRAN, E. VINCENT, D. JOUVET. *Nonparametric uncertainty estimation and propagation for noise robust ASR*, in "IEEE/ACM Transactions on Audio, Speech and Language Processing", November 2015, vol. 23, n^o^ 11, pp. 1835-1846 [*DOI :* 10.1109/TASLP.2015.2450497], https://hal.inria.fr/hal-01114329

Invited Conferences

[19] A. H. ABDELAZIZ, S. WATANABE, J. R. HERSHEY, E. VINCENT, D. KOLOSSA. *Uncertainty propagation through deep neural networks*, in "Interspeech 2015", Dresden, Germany, September 2015, https://hal.inria.fr/hal-01162550

[20] Y. LAPRIE. *Contribution of the acoustic cues to the non-native accent*, in "169th meeting: Acoustical Society of America", Pittsburgh, United States, May 2015, https://hal.inria.fr/hal-01188770

[21] J. Le Roux, E. Vincent, J. R. Hershey, D. P. Ellis. *MICbots: collecting large realistic datasets for speech and audio research using mobile robots*, in "IEEE 2015 International Conference on Acoustics, Speech and Signal Processing (ICASSP)", Brisbane, Australia, April 2015, https://hal.inria.fr/hal-01116822

[22] S. Ouni. *Toward Realistic Expressive Audiovisual Speech Synthesis*, in "Expressive Virtual Actors workshop", Grenoble, France, G. Bailly, R. Ronfard (editors), Gipsa-Lab, November 2015, This event is a special session in the series supported by the ERC Expressive grant, https://hal.inria.fr/hal-01243644

[23] E. Vincent, A. Sini, F. Charpillet. *Audio source localization by optimal control of a mobile robot*, in "IEEE 2015 International Conference on Acoustics, Speech and Signal Processing (ICASSP)", Brisbane, Australia, April 2015, https://hal.inria.fr/hal-01103949

[24] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, B. Schuller. *Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR*, in "12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)", Liberec, Czech Republic, August 2015, https://hal.inria.fr/hal-01163493

International Conferences with Proceedings

[25] J. Barker, R. Marxer, E. Vincent, S. Watanabe. *The third 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines*, in "2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)", Scottsdale, AZ, United States, December 2015, https://hal.inria.fr/hal-01211376

[26] K. Bartkova, D. Jouvet. *Impact of frame rate on automatic speech-text alignment for corpus-based phonetic studies*, in "ICPhS'2015 - 18th International Congress of Phonetic Sciences", Glasgow, United Kingdom, August 2015, https://hal.inria.fr/hal-01183637

[27] A. Bonneau. *Realizations of French voiced fricatives by German learners as a function of speaker level and prosodic boundaries*, in "18th International Congress of Phonetic Sciences, ICPhS 2015", Glasgow, United Kingdom, University of Glasgow, August 2015, 5 p. , https://hal.inria.fr/hal-01186062

[28] A. Bonneau, M. Cadot. *German non-native realizations of French voiced fricatives in final position of a group of words*, in "Interspeech 2015", Dresde, Germany, Möller, S., Ney, H., Moebius, B., Nöth, E., September 2015, https://hal.inria.fr/hal-01186068

[29] M. Cadot, A. Bonneau. *Pourquoi et comment transformer des variables quantitatives en catégorielles ? Application à l'intonation de la langue française*, in "ASI-8", Radès, France, November 2015, https://hal.archives-ouvertes.fr/hal-01188477

[30] M. Dargnat, K. Bartkova, D. Jouvet. *Discourse Particles In French: Prosodic Parameters Extraction and Analysis*, in "International Conference on Statistical Language and Speech Processing", Budapest, Hungary, November 2015, https://hal.inria.fr/hal-01184197

[31] B. Dumortier, E. Vincent, M. Deaconu. *Acoustic Control of Wind Farms*, in "EWEA 2015 - European Wind Energy Association", Paris, France, November 2015, https://hal.inria.fr/hal-01233730

[32] 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

[33] D. FOHR, I. ILLINA. *Continuous Word Representation using Neural Networks for Proper Name Retrieval from Diachronic Documents*, in "Interspeech 2015", Dresden, Germany, September 2015, https://hal.archives-ouvertes.fr/hal-01184951

[34] *Best Paper*
D. FOHR, I. ILLINA. *Neural Networks for Proper Name Retrieval in the Framework of Automatic Speech Recognition*, in "IEEE International Conference on Information Systems and Economic Intelligence", hammamet, Tunisia, 2015, https://hal.archives-ouvertes.fr/hal-01184957.

[35] D. FOHR, O. MELLA. *Detection of Phone Boundaries for Non-Native Speech using French-German Models*, in "Workshop on Speech and Language Technology in Education", Leipzig, Germany, September 2015, https://hal.archives-ouvertes.fr/hal-01185195

[36] D. FOHR, O. MELLA, D. JOUVET. *De l'importance de l'homogénéisation des conventions de transcription pour l'alignement automatique de corpus oraux de parole spontanée*, in "8es Journées Internationales de Linguistique de Corpus (JLC2015)", Orléans, France, September 2015, https://hal.inria.fr/hal-01183352

[37] T. FUX, D. JOUVET. *Evaluation of PNCC and extended spectral subtraction methods for robust speech recognition*, in "EUSIPCO 2015 - 23rd European Signal Processing Conference", Nice, France, August 2015, https://hal.inria.fr/hal-01183645

[38] I. ILLINA, D. FOHR. *Different word representations and their combination for proper name retrieval from diachronic documents*, in "IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)", Scottsdale, United States, December 2015, https://hal.inria.fr/hal-01201533

[39] I. ILLINA, D. FOHR. *Neural Networks Revisited for Proper Name Retrieval from Diachronic Documents*, in "LTC Language & Technology Conference", Poznan, Poland, November 2015, pp. 120-124, https://hal.archives-ouvertes.fr/hal-01240480

[40] D. JOUVET, K. BARTKOVA. *Acoustical Frame Rate and Pronunciation Variant Statistics*, in "International Conference on Statistical Language and Speech Processing", Budapest, Hungary, November 2015, https://hal.inria.fr/hal-01184195

[41] D. JOUVET, A. BONNEAU, J. TROUVAIN, F. ZIMMERER, Y. LAPRIE, B. MÖBIUS. *Analysis of phone confusion matrices in a manually annotated French-German learner corpus*, in "Workshop on Speech and Language Technology in Education", Leipzig, Germany, September 2015, https://hal.inria.fr/hal-01184186

[42] A. LIUTKUS, R. BADEAU. *Generalized Wiener filtering with fractional power spectrograms*, in "40th International Conference on Acoustics, Speech and Signal Processing (ICASSP)", Brisbane, Australia, IEEE, April 2015, https://hal.archives-ouvertes.fr/hal-01110028

[43] A. LIUTKUS, D. FITZGERALD, Z. RAFII. *Scalable audio separation with light kernel additive modelling*, in "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", Brisbane, Australia, IEEE, April 2015, https://hal.inria.fr/hal-01114890

[44] A. LIUTKUS, U. ŞIMŞEKLI, T. CEMGIL. *Extraction of Temporal Patterns in Multi-rate and Multi-modal Datasets*, in "International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)", Liberec, Czech Republic, August 2015, https://hal.inria.fr/hal-01170932

[45] O. MELLA, D. FOHR, A. BONNEAU. *Inter-annotator agreement for a speech corpus pronounced by French and German language learners*, in "Workshop on Speech and Language Technology in Education", Leipzig, Germany, ISCA Special Interest Group (SIG) on Speech and Language Technology in Education, September 2015, https://hal.archives-ouvertes.fr/hal-01185194

[46] F. MEZZOUDJ, D. LANGLOIS, D. JOUVET, A. BENYETTOU. *Textual Data Selection for Language Modelling in the Scope of Automatic Speech Recognition*, in "International Conference on Natural Language and Speech Processing", Alger, Algeria, October 2015, https://hal.inria.fr/hal-01184192

[47] L. OROSANU, D. JOUVET. *Adding new words into a language model using parameters of known words with similar behavior*, in "International Conference on Natural Language and Speech Processing", Alger, Algeria, October 2015, https://hal.inria.fr/hal-01184194

[48] L. OROSANU, D. JOUVET. *Combining lexical and prosodic features for automatic detection of sentence modality in French*, in "International Conference on Statistical Language and Speech Processing", Budapest, Hungary, November 2015, https://hal.inria.fr/hal-01184196

[49] L. OROSANU, D. JOUVET. *Detection of sentence modality on French automatic speech-to-text transcriptions*, in "International Conference on Natural Language and Speech Processing", Alger, Algeria, October 2015, https://hal.inria.fr/hal-01184193

[50] A. PIQUARD-KIPFFER, O. MELLA, J. MIRANDA, D. JOUVET, L. OROSANU. *Qualitative investigation of the display of speech recognition results for communication with deaf people*, in "6th Workshop on Speech and Language Processing for Assistive Technologies", Dresden, Germany, SIG-SLPAT, September 2015, 7 p. , https://hal.inria.fr/hal-01183349

[51] T. PRÄTZLICH, R. BITTNER, A. LIUTKUS, M. MÜLLER. *Kernel additive modeling for interference reduction in multi-channel music recordings*, in "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", Brisbane, Australia, April 2015, https://hal.inria.fr/hal-01116686

[52] Z. RAFII, A. LIUTKUS, B. PARDO. *A simple user interface system for recovering patterns repeating in time and frequency in mixtures of sounds*, in "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", Brisbane, France, April 2015, https://hal.inria.fr/hal-01116689

[53] D. RIBAS, E. VINCENT, J. R. CALVO. *Full multicondition training for robust i-vector based speaker recognition*, in "Interspeech 2015", Dresden, Germany, September 2015, https://hal.inria.fr/hal-01158774

[54] D. RIBAS, E. VINCENT, J. R. CALVO. *Uncertainty propagation for noise robust speaker recognition: the case of NIST-SRE*, in "Interspeech 2015", Dresden, Germany, September 2015, 5 p. , https://hal.inria.fr/hal-01158775

[55] S. SIVASANKARAN, A. A. NUGRAHA, E. VINCENT, J. A. MORALES CORDOVILLA, S. DALMIA, I. ILLINA, A. LIUTKUS. *Robust ASR using neural network based speech enhancement and feature simulation*, in "2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)", Arizona, United States, December 2015, https://hal.inria.fr/hal-01204553

[56] N. SOUVIRAÀ-LABASTIE, E. VINCENT, F. BIMBOT. *Music separation guided by cover tracks: designing the joint NMF model*, in "40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015", Brisbane, Australia, April 2015, https://hal.archives-ouvertes.fr/hal-01108675

[57] F. R. STOTER, 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, IEEE, 2016, https://hal.archives-ouvertes.fr/hal-01248012

[58] D. T. TRAN, N. ONO, E. VINCENT. *Fast DNN training based on auxiliary function technique*, in "ICASSP 2015 - 40th IEEE International Conference on Acoustics, Speech and Signal Processing", Brisbane, Queensland, Australia, April 2015, https://hal.inria.fr/hal-01107809

[59] D. T. TRAN, E. VINCENT, D. JOUVET. *Discriminative uncertainty estimation for noise robust ASR*, in "40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015", Brisbane, Queensland, Australia, April 2015, https://hal.inria.fr/hal-01103969

[60] F. ZIMMERER, J. TROUVAIN, A. BONNEAU. *One corpus, one research question, three methods "German vowels produced by French speakers"*, in "Worshop on Phonetic learner corpora. Satellite meeting of ICPhS 2015", Glasgow, United Kingdom, Trouvain, J., Zimmerer, F., Gosy, M., Bonneau, A., August 2015, https://hal.inria.fr/hal-01186078

[61] A. BANDINI, S. OUNI, P. COSI, S. ORLANDI, C. MANFREDI. *Accuracy of a markerless acquisition technique for studying speech articulators. In Interspeech 2015*, in "Interspeech 2015", Dresden, Germany, ISCA (editor), September 2015, https://hal.inria.fr/hal-01189000

[62] I. SHEIKH, I. ILLINA, D. FOHR, G. LINARÈS. *OOV Proper Name Retrieval using Topic and Lexical Context Model*, in "IEEE International Conference on Acoustics, Speech and Signal Processing", Brisbane, Australia, 2015, https://hal.archives-ouvertes.fr/hal-01184963

[63] I. SHEIKH, I. ILLINA, D. FOHR. *Recognition of OOV Proper Names in Diachronic Audio News*, in "IEEE International Conference on Information Systems and Economic Intelligence", Hammamet, Tunisia, 2015, https://hal.archives-ouvertes.fr/hal-01184958

[64] I. SHEIKH, I. ILLINA, D. FOHR. *Study of Entity-Topic Models for OOV Proper Name Retrieval*, in "Interspeech 2015", Dresden, Germany, September 2015, https://hal.archives-ouvertes.fr/hal-01184955

Conferences without Proceedings

[65] Y. LAPRIE, B. ELIE, A. TSUKANOVA. *2D Articulatory Velum Modeling Applied to Copy Synthesis of Sentences Containing Nasal Phonemes*, in "International Congress of Phonetic Sciences", Glasgow, United Kingdom, August 2015, https://hal.inria.fr/hal-01188738

[66] Y. LAPRIE. *An articulatory model of the velum developed from cineradiographic data*, in "169th Meeting: Acoustical Society of America", Pittsburgh, United States, May 2015, https://hal.inria.fr/hal-01188760

[67] A. LIUTKUS, D. FITZGERALD, R. BADEAU. *Cauchy Nonnegative Matrix Factorization*, in "IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)", New Paltz, NY, United States, October 2015, https://hal.inria.fr/hal-01170924

[68] A. LIUTKUS, T. OLUBANJO, E. MOORE, M. GHOVANLOO. *Source Separation for Target Enhancement of Food Intake Acoustics from Noisy Recordings*, in "IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)", New Paltz, NY, United States, October 2015, https://hal.inria.fr/hal-01174886

[69] N. ONO, Z. RAFII, D. KITAMURA, N. ITO, A. LIUTKUS. *The 2015 Signal Separation Evaluation Campaign*, in "International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)", Liberec, France, Latent Variable Analysis and Signal Separation, August 2015, vol. 9237, pp. 387-395 [*DOI :* 10.1007/978-3-319-22482-4_45], https://hal.inria.fr/hal-01188725

[70] A. PIQUARD-KIPFFER. *Terminal portable de communication et affichage de la reconnaissance vocale. Enjeux et rapports à l'écrit. Etude préliminaire auprès d'adultes déficients auditifs*, in "3ème Colloque International IDEKI 2015. Didactiques, Métiers de l'humain et Intelligence collective. Construction de savoirs et de dispositifs", COLMAR, France, IDEKI, December 2015, https://hal.inria.fr/hal-01243142

Scientific Books (or Scientific Book chapters)

[71] *LNCS 9237 - Proceedings of the 12th International Conference on Latent Variable Analysis and Signal Separation*, Springer, Liberec, Czech Republic, August 2015, https://hal.inria.fr/hal-01183618

Research Reports

[72] V. COLOTTE, C. EMILIEN. *JCorpusRecorder*, Université de Lorraine, December 2015, https://hal.inria.fr/hal-01244553

[73] A. LIUTKUS. *Scale-Space Peak Picking*, Inria Nancy - Grand Est (Villers-lès-Nancy, France), January 2015, https://hal.inria.fr/hal-01103123

[74] M. MOUSSALLAM, A. LIUTKUS, L. DAUDET. *Listening to features*, Institut Langevin, ESPCI - CNRS - Paris Diderot University - UPMC, January 2015, 24 p. , https://hal.inria.fr/hal-01118307

[75] A. A. NUGRAHA, A. LIUTKUS, E. VINCENT. *Multichannel audio source separation with deep neural networks*, Inria, June 2015, n^{o} RR-8740, https://hal.inria.fr/hal-01163369

[76] L. S. R. SIMON, E. VINCENT. *Combining blockwise and multi-coefficient stepwise approches in a general framework for online audio source separation*, Inria, August 2015, n^{o} RR-8766, 18 p. , https://hal.inria.fr/hal-01186948

Scientific Popularization

[77] A. PIQUARD-KIPFFER. *Faire voir une histoire : Louis et son incroyable chien Noisette in Les cahiers pédagogiques, Dossier Lire et écrire avec la littérature numérique coordonné par Yaël Boublil et Jacques Crinon. 7p*, in "Les Cahiers Pédagogiques", 2016, https://hal.inria.fr/hal-01191878

Other Publications

[78] P. CARROLL, J. TROUVAIN, F. ZIMMERER, Y. LAPRIE, O. MELLA, D. FOHR. *Improvements for a German Vowel Trainer CAPT Tool*, November 2015, Individualized Feedback for Computer-Assisted Spoken Language Learning, Poster, https://hal.archives-ouvertes.fr/hal-01243043

[79] 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 vocal folds with glottal chink*, September 2015, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01199792

[80] D. FITZGERALD, A. LIUTKUS, R. BADEAU. *Projection-based demixing of spatial audio*, January 2016, working paper or preprint, https://hal.inria.fr/hal-01260588

[81] X. JAUREGUIBERRY, E. VINCENT, G. RICHARD. *Fusion methods for speech enhancement and audio source separation*, October 2015, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01120685

[82] S. OUNI, G. GRIS. , ACM (editor) *Dynamic realistic lip animation using a limited number of control points*, Proceeding SIGGRAPH '15 ACM SIGGRAPH 2015 Posters, ACM, August 2015, 1 p. , SIGGRAPH 2015, Poster [*DOI :* 10.1145/2787626.2787628], https://hal.inria.fr/hal-01188997

[83] 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. Etude préliminaire auprès d'adultes déficients auditifs*, March 2016, 15 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

[84] E. VINCENT, E. A. P. HABETS. *Advanced spatial speech and audio processing*, August 2015, LVA/ICA 2015 Summer School on Latent Variable Analysis and Signal Separation, https://hal.inria.fr/hal-01183505

[85] E. VINCENT. *Is audio signal processing still useful in the era of machine learning?*, October 2015, 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), https://hal.inria.fr/hal-01183506

[86] P.-A. VUISSOZ, F. ODILLE, Y. LAPRIE, E. VINCENT, J. FELBLINGER. *Sound synchronization and motion compensated reconstruction for speech Cine MRI*, May 2015, ISMRM 2015 Annual Meeting, Poster, https://hal.inria.fr/hal-01183504

[87] P.-A. VUISSOZ, F. ODILLE, E. VINCENT, J. FELBLINGER, Y. LAPRIE. *Synchronisation vocale et mouvement compensé en reconstruction pour une ciné IRM de la parole*, March 2015, 2e Congrès de la SFRMBM (Société Française de Résonance Magnétique en Biologie et Médecine), Poster, https://hal.inria.fr/hal-01104230