Activity Report 2014

Team MULTISPEECH

Speech Modeling for Facilitating Oral-Based Communication

RESEARCH CENTER
Nancy - Grand Est

THEME
Language, Speech and Audio

Table of contents

Team MULTISPEECH

Keywords: Speech, Machine Learning, Statistical Methods, Perception, Recognition, Audio, Linguistics, Natural Language, Signal Processing, Modeling

The team was known as PAROLE during the first semester of 2014, and as MULTISPEECH during the second semester of 2014. As Kamel Smaïli and David Langlois have moved to the LORIA team SMART, that they have created, in January 2014, and as Antoine Liutkus was recruited in January 2014, the core team (senior researchers, researchers, and associate professors) was the same over the whole year 2014.

Hence, this activity report covers the research activity of the team over the whole 2014 year (i.e., refers to PAROLE for the first semester, and to MULTISPEECH for the second semester). The research program presented in this activity report is the one of the MULTISPEECH proposal, and the new results are presented according to the corresponding research directions.

Creation of the Team: 2014 July 01.

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 Lorraine, Associate Professor]

Engineers

Ilef Ben Farhat [Inria]
Julie Busset [CNRS]
Antoine Chemardin [CNRS, from Oct 2014]
Yann Salaün [Inria, until Nov 2014]
Aghilas Sini [CNRS, from Nov 2014]

PhD Students

Baldwin Dumortier [Inria, from Jun 2014]
Arseniy Gorin [Inria]
Xabier Jaureguiberry [Institut Telecom]
Luiza Orosanu [Inria]
Imran Sheikh [Univ. Lorraine]
Nathan Souviraà-Labastie [Univ. Rennes 1]
Dung Tran [Inria]

Post-Doctoral Fellows

Benjamin Elie [Inria]
Camille Fauth [CNRS, until Jul 2014]
Thibaut Fux [Inria, from Sep 2014]

Emad Girgis [Inria, from Nov 2014]

Visiting Scientists

Andrea Bandini [Univ. of Bologna, from Oct 2014]
Dayana Ribas [Advanced Technologies Application Center, La Habana, Cuba, from Sep 2014]

Administrative Assistants

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

Others

Guillaume Gris [Ecole Polytechnique, from Mar 2014 until Jul 2014]
Simon Meoni [Univ. Lorraine, from Mar 2014 until Jun 2014]
Aghilas Sini [Univ. Paul Sabatier, Toulouse, from Mar 2014 until Aug 2014]

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 three following aspects that are particularly considered, namely:

- ***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 will also be 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 will be investigated as a way of making speech recognition systems more robust to noise. Dealing with spontaneous speech and handling new proper names are two critical aspects that will be 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. Confidence measures associated with word recognition results aim at providing information on the reliability of the hypothesized words. Finally, with respect to phonetic segment boundaries and prosodic parameters 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

As mentioned previously, MULTISPEECH is structured along three research directions that are associated to the three challenges previously described: 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 jaw, lips, tongue, soft palate and larynx 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 audiovisual speech. On the one hand, we study the relationship from acoustic speech signal to vocal tract, in the context of acoustic-to-articulatory inversion, and from vocal tract to acoustic speech, in the context of articulatory synthesis. On the other hand, we work on expressive audiovisual speech synthesis, where both expressive acoustic speech and visual signals are generated from text. 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 sounds robustly. From the point of view of perception, these contrasts enable efficient processes of categorization in the peripheral and central human auditory system. 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 MRI data of the entire vocal tract together with denoised speech signals, and evaluating several approaches of acoustic simulation. This will enable the acoustic mismatch to be better controlled and the determination of the potential precision of inversion to be evaluated in particular.

Up to now, acoustic-to-articulatory inversion has been addressed as an instantaneous problem, articulatory gestures being recovered by concatenating local solutions via the determination of trajectories minimizing some articulatory cost. 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.

This area of research relies on the equipment available in the laboratory to acquire articulatory data: articulograph Carstens AG501, head-neck antenna to acquire MRI of the vocal tract at Nancy Hospital, and multimodal acquisition system. Very few sites in France benefit from such a combination of acquisition devices.

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. Our research impacts both audiovisual and acoustic-only synthesis fields.

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 (3D markers on the face, data acquired by MAGRIT team) and on an extension of our non-uniform 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 where the tongue deformations come from EMA data analysis; and a lip-model to tackle the main recurrent problem of the lack of some lip markers in the 3D data. We will also improve the TTS engine to increase the accuracy of the unit selection simultaneously into the acoustic and visual domains (learning weights, feature selection...).

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 given prosodic entities such as emphasis, in order to highlight some important key words. Expressivity could be introduced before the unit selection step but also by developing algorithms intended to modify the parameters of prosody (in the acoustic domain, and in the visual domain as well). 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 this issue we will also investigate Hidden Markov Model (HMM) based synthesis. The flexibility of the HMM-based approach enables the adjustment of the modeling parameters according to the available data and an easy adaption of the system to various conditions. This point will rely upon our experience in HMM modeling.

To acquire the facial data, we consider using marker-less motion capture system using a kinect-like system with a face tracking software. The software presents a user-friendly interface to track and visualize the motion in real time. Audio is also acquired synchronously with facial data. The advantage of this new system is to acquire rapidly the movements of the face with an acceptable quality. This system is used as an alternative relatively low-cost system to the VICON system.

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 which are 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 (i.e., at the segmental level and at the supra-segmental level). Moreover, as categorization is a perceptual process, studies on the emergence of categories must also rely on perceptual experiments.

Studies on native sounds have been an important research area of the team for years, leading to the notion of "selective" acoustic cues and the development of acoustic detectors. This know-how will be exploited in the study of non-native sounds. 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). Results aim 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, 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 complex 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, which is a critical aspect that is studied in the project. 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 in order to overcome the limited size of the current static lexicons used, especially with respect to proper names. Statistical approaches are also useful for generating speech signals. Along this direction, MULTISPEECH mainly considers voice transformation techniques, with their application to pathological voices, and statistical speech synthesis applied to expressive multimodal speech synthesis.

3.3.1. Acoustic modeling

Acoustic modeling is a key issue for automatic speech recognition. Despite progress made for many years, acoustic modeling is still far from perfect, and current speech recognition applications rely on strong constraints (limited vocabulary, speaker adaptation, restricted syntax...) to achieve acceptable performance. As the acoustic models represent the acoustic realization of the sounds, they have to account for many variability sources, such as speaker characteristics, microphones, noises, etc. Extension of the HMM formalism based on the Dynamic Bayesian Networks (DBN) formalism are investigated further for handling such variability sources; as well as other approaches to dynamically constrain the search space according to known or estimated characteristics of the utterance being processed. Deep Neural Networks (DNN) based approaches will also be investigated as means of making speech recognition systems more accurate and robust. Speaker dependent modeling and speaker adaptation will also be investigated in relation with HMM-based speech synthesis and statistical voice conversion.

State-of-the-art speech recognition systems are still very sensitive to the quality of speech signals they have to deal with; their performance degrades rapidly when they deal 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 Bayesian source separation techniques using multiple microphones and/or models of non-speech events. Some of the challenges include building a non-parametric model of the sources in the time-frequency-channel domain, linking the parameters of this model to the cepstral representation used in speech processing, modeling the temporal

structure of environmental noise, and exploiting large audio data sets to automatically discover new models. Beyond the definition of such complex models, the difficulty is to design scalable estimation algorithms robust to overfitting, that will be integrated in the FASST [6] framework that was recently developed.

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 the processing of spontaneous speech. Collaborations are ongoing with the SMarT team on linguistic modeling aspects.

Proper names are relevant keys in information indexing, but are a real problem in transcribing many diachronic spoken documents (such as radio or TV shows) 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 possibly collected over the web. Random Indexing (RI) and Latent Dirichlet Allocation (LDA) are two possible approaches to be used for this purpose. Also, to overcome the limitations of current n-gram based language models, we investigate language models defined on a continuous space in order to achieve a better generalization on unseen data, and to model long-term dependencies. This is achieved through neural network based approaches. We also want to introduce into these new models additional relevant information such as linguistic features, semantic relation, topic or user-dependent information.

Spontaneous speech utterances are often ill-formed and frequently contain disfluencies (hesitations, repetitions...) that degrade speech recognition performance. This is partly due to the fact that disfluencies are not properly represented in linguistic models estimated from clean text data (coming from newspapers for example); hence a particular effort will be set for improving the modeling of these events.

Attention will also be set on pronunciation lexicons in particular 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. Automatic grapheme-to-phoneme state-of-the-art approaches, based for example on Joint Multigram Models (JMM) or Conditional Random Fields (CRF) will be further investigated and combined.

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. 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).

In addition to the statistical aspects of the voice conversion approaches, signal processing is critical for good quality speech output. Information on the fundamental frequency is chaotic in the case of esophageal speech or non-existent in the case of the whispered voice. So after applying voice conversion techniques for enhancing pathological voices, the excitation spectrum must be predicted or corrected. That is the challenge that is addressed in the project. Also, in the context of acoustic feedback in foreign language learning, voice modification approaches (either statistical or not) 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. Speaker-dependent HMM modeling constitute the basis of such an approach. 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. The reasons are twofold: first, parameters (F0, spectrum, duration...) are modeled independently and the models, even when taking into account dynamics,

do not manage to generate parameters with a good precision. Second, the HMM generates sequences of feature vectors from which the actual speech signals are reconstructed, and this impacts on its quality. 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. One secondary objective will be to unify the HMM-based and the concatenation-based approaches.

3.4. Uncertainty estimation and exploitation in speech processing

After the explicit modeling presented and the statistical modeling that were previously described, we focus here 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). The first case (speech recognition) is connected with the computation of prosodic features for structuring speech recognition output, whereas the second case (forced alignment) is important in the context of non-native speech segmentation for automatic feedbacks in language learning.

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 and/or the uncertainty on 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 (in collaboration with the SYNALP team). 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 or utterance modality, or in the determination of the structure of the utterance. In a first approach, uncertainty estimation will rely on the comparison, and possibly the combination, of several estimators (several segmentation processes, several pitch algorithms).

4. Application Domains

4.1. Introduction

Approaches and models developed in the MULTISPEECH project are intended to be used for facilitating oral-based 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, represent 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. Also, the system should provide indications on what should be corrected, a guidance which is considered as necessary by specialists in the oral aspects of language learning. Consequently, 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. For example, with respect to prosody, the diagnosis provided through both a text and a visual display, comes from an evaluation of the melodic curve and of the phoneme durations of the learner's realization; and the perceptual feedback consists in a replacement of the learner's prosodic cues by those of the reference; i.e., the signal of the learner's utterance is modified in order to reflect the prosodic cues (duration and F0) of the reference in order to make the learner aware of the expected prosodic cues. The diagnosis step strongly relies on the studies on categorization of sounds and prosody in the mother tongue and in the second language, and also depends on the influence between them. Furthermore, reliable diagnosis on individual utterances 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 also 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 improving their autonomy. The following applications are considered in the project.

The first one 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/or syllables, which takes also into account the reliability of the recognized items.

The second 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.

The 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

The 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.

The 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.

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 application 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. New Software and Platforms

5.1. Introduction

This software section is organized along three main axes: tools for automatic speech processing, then visualization tools used to display different aspects of speech data and which possibly feature other functionalities; and finally tools and platforms for acquiring articulatory data.

5.2. Speech processing tools

Participants: Denis Jouvet, Dominique Fohr, Odile Mella, Irina Illina, Emmanuel Vincent, Antoine Liutkus, Vincent Colotte, Yann Salaün, Antoine Chemardin.

These automatic speech processing tools deal with audio data transcription (ANTS), audio sources separation (FASST), speech-text alignment (LASTAS) and text-to-speech synthesis (SoJA).

5.2.1. ANTS (Automatic News Transcription System)

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 of a manageable size, and then each segment is decoded using the most adequate acoustic model with a large vocabulary continuous speech recognition engine (Julius or Sphinx). Further processing passes are run in order to apply unsupervised adaptation processes on the features (VTLN: Vocal Tract Length Normalization) and/or on the model parameters (MLLR: Maximum Likelihood Linear Regression), or to use Speaker Adaptive Training (SAT) based models. Moreover decoding results of several systems can be efficiently combined for improved decoding performance. The latest version takes advantage of the multiple CPUs available on a computer, and runs on both standalone linux machines and on clusters.

5.2.2. FASST (Flexible Audio Source Separation Toolbox)

FASST [1] is a toolbox for audio source separation distributed under the Q Public License. Version 2 in C++ has been developed in the context of the ADT FASST (conducted by MULTISPEECH in collaboration with the PANAMA and TEXMEX teams from Inria Rennes - cf. 8.1.6) and released in January 2014. Its unique feature is the possibility for users to specify easily a suitable algorithm for their use case thanks to the general modeling and estimation framework proposed in [6]. It forms the basis of most of our current research in audio source separation, some results of which will be incorporated into future versions of the software.

5.2.3. KAM (Kernel Additive Modelling)

The Kernel Additive Modelling framework for source separation [13], [42] has been proposed this year by Liutkus et al. as a new and effective approach to source separation. In 2014, two different implementations of KAM have been registered with the APP: a Matlab version matKAM and a python version pyKAM. The former is under a aGPL license, while the latter is under a proprietary license. The rationale for this choice is that the Matlab version is to be mainly disseminated for research purpose to the colleagues in the field, that mainly use Matlab, while the python version is more liable to lead to industrial transferts.

5.2.4. LASTAS (Loria Automatic Speech-Text Alignment Software)

LASTAS is a software for aligning a speech signal with its corresponding orthographic transcription. 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 together the boundaries at the phone level and at the word level.

This year, this software has been included in a web application for speech-text automatic alignement, named ASTALI, which will soon be available [2].

[1] http://bass-db.gforge.inria.fr/fasst/
[2] http://astali.loria.fr

5.2.5. CoALT (Comparing Automatic Labeling Tools)

CoALT is a software for comparing the results of several automatic labeling processes through user defined criteria [70].

5.2.6. SoJA (Speech synthesis platform in Java)

SOJA [3] is a software for Text-To-Speech synthesis (TTS) which relies on a non uniform unit selection algorithm. It performs all steps from text to speech signal output. Moreover, a set of associated tools is available for elaborating a corpus for a TTS system (transcription, alignment...). Currently, the corpus contains 1800 sentences (about 3 hours of speech) recorded by a female speaker. Most of the modules are in Java; some are in C. The software runs under Windows and Linux. It can be launched with a graphical user interface or directly integrated in a Java code or by following the client-server paradigm. We will consider extending and making SoJA more modular and able to handle both acoustic and visual features, in order to use it for both acoustic-only synthesis and audiovisual synthesis. In the future, the text-to-speech synthesis platform will get extended to take into account expressivity features.

5.3. Speech visualization tools

Participants: Yves Laprie, Slim Ouni, Julie Busset, Aghilas Sini, Ilef Ben Farhat.

This set of tools aims at visualizing various aspects of speech data: speech audio signal (SNOORI), Electro-Magnetographic Articulography (EMA) data (VisArtico) and speech articulators from X-ray images (Xarticulators).

5.3.1. SNOORI: speech analysis and visualization software

JSnoori is written in Java and uses signal processing algorithms developed within the WinSnoori [4] 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.

We added the possibility of developing scripts for JSnoori by using Jython which allows Java classes of JSnoori to be used from Python. This required some refactoring of JSnoori classes in order to make them more independent from the JSnoori context.

5.3.2. VisArtico: Visualization of EMA Articulatory data

VisArtico [5] 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 [72]. Several researchers showed interest in this application. In fact, VisArtico is very useful for the speech science community, and it makes the use of articulatory data more accessible. The software is a cross-platform application (i.e., running under Windows, Linux and Mac OS).

[3] http://soja-tts.loria.fr
[4] http://www.loria.fr/~laprie/WinSnoori/
[5] http://visartico.loria.fr/

Within the framework of an Inria ADT project (cf. 8.1.7), we are implementing several improvements to the software. It is possible to use VisArtico to import and export several articulatory data formats. In addition, it possible to insert images (MRI or X-Ray, for instance) to compare the EMA data with data obtained through other acquisition techniques. Finally, it is possible to generate a movie for any articulatory-acoustic sequence. These improvements (and others) extend the capabilities of VisArtico and make it more useful and widely used. The software will also provide a demonstration module that will produce articulatory synthesis from EMA data or text. It animates the vocal tract, using articulatory data and generates the corresponding acoustic signal. VisArtico is freely available for research.

5.3.3. Xarticulators: delineation of speech articulators in medical images

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

We added new articulatory model construction features intended to approximate the tongue shape more correctly when the tongue contacts the palate during the stop closure of /k/ and /t/ and we added more complete modeling of the epiglottis and the larynx region. Future developments will focus on the development of time patterns to synthesize any speech sound and on the coupling between vocal folds and vocal tract.

5.4. Data acquisition

Participants: Vincent Colotte, Slim Ouni, Yves Laprie.

The nature of our research makes us highly concerned by acquisition and processing of speech data. Besides acquisition of speech audio signals, we are concerned with the acquisition of articulatory data, mainly ElectroMagnetographic Articulography (EMA) data using an articulograph and Magnetic Resonance Imaging (MRI) data. EMA captures articulatory movements in three dimensions (3D) with a high temporal resolution by tracking tiny sensors attached to speech articulators such as the tongue, teeth, and lips. MRI is a non-invasive, hazard-free medical imaging technique allowing for high-resolution scans of the vocal tract.

5.4.1. JCorpusRecorder

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 of the sentence to pronounce. This software is suitable for recording sentences with information to guide the speaker. The sentences can be presented randomly. The software is developed in Java. It is currently used for the recording of sentences in several projects.

5.4.2. EMA acquisition platform

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 (sterilization, calibration, etc.). 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) to very high frequency (1250Hz). In addition, we have developed a powerful tool, VisArtico, to visualize articulatory data acquired using an articulograph.

5.4.3. MRI acquisition platform

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, we acquired static MRI data for two speakers (approximately 90 blocked articulations corresponding to vowels and consonants followed by a vowel) and we carried out preliminary experiments intended to acquire dynamic data.

6. New Results

6.1. Highlights of the Year

The version 2 of our source separation toolbox FASST [65] has been downloaded more than 300 times since its release in January 2014.

6.2. Explicit modeling of speech production and perception

Participants: Yves Laprie, Slim Ouni, Vincent Colotte, Anne Bonneau, Agnès Piquard-Kipffer, Martine Cadot [Univ. Lorraine], Antoine Liutkus, Emmanuel Vincent, Odile Mella, Benjamin Elie, Camille Fauth, Julie Busset, Andrea Bandini, Guillaume Gris, Simon Meoni.

6.2.1. Articulatory modeling

6.2.1.1. Acquisition of articulatory data

Acquisition of articulatory data plays a central role in the construction of articulatory models and investigation of articulatory gestures. In cooperation with the IADI laboratory (Nancy hospital) we thus conducted a series of preliminary experiments intended to acquire cine-MRI data. Images of the film are reconstructed thanks to the cine-GRICS algorithm developed at IADI [56].

The second research track concerns ultrasound (US) imaging which presents the interest of offering a good temporal resolution without any health hazard and at a reasonable price. However, it cannot be used alone because there is no reference coordinate system and no spatial calibration. We thus used a multimodal acquisition system developed by the Magrit team, which uses electromagnetography sensors to locate the US probe, and the method used to calibrate the US modality. We experimented this system to investigate the most appropriate acquisition protocol for Magnetic Resonance Imaging [37].

We also use an articulograph to acquire articulatory data. Within the framework of the EQUIPEX ORTOLANG, we acquired this year an AG501, a 24-channel articulograph. This system is the most advanced electromagnetography acquisition system. It has been used for two articulatory studies: (1) investigating the effects of posture and noise on speech production [48] and (2) studying the pauses in spontaneous speech from an articulatory point of view. We also conducted an exploratory study on retrieving the 3D shape of the palate from electromagnetography tracings (the work of Simon Meoni, a master student in Cognitive Sciences).

6.2.1.2. Acoustic-to-articulatory inversion

Our previous works about acoustic-to-articulatory inversion relied on the exploration of a vast articulatory codebook covering the whole articulatory space that could be reached by a speaker. This solution presents the main drawback of requiring the construction a codebook for each speaker. We thus developed a multimodal approach to estimate the area function and the length of the vocal tract of oral vowels. The method is based on an iterative technique consisting in deforming an initial area function so that the output acoustic vector matches a specified target. The chosen acoustic vector is the formant frequency pattern. In order to regularize the ill-defined problem, several constraints are added to the algorithm. First, the lip termination area is estimated via a facial capture software. Then, the area function is constrained so that it does not get too far from a neutral position, and so that it does not change too quickly from a temporal frame to the next, when dealing with dynamic inversion. The method proves to be efficient for approximating the area function and the length of the vocal tract for oral French vowels, both in static and dynamic configurations.

6.2.1.3. Articulatory models

The development of articulatory models is a crucial aspect of articulatory synthesis since this determines the success of synthesis. The previous model was developed for X-ray images. This means that the laryngeal part of the model associates the larynx with the piriform sinuses event if these two structures are not in the same sagittal plane. The new model separates the two structures if needed. Additionally, the larynx and the epiglottis are controlled independently which corresponds to the anatomical truth. Previous attempts to modeling epiglottis used principal component analysis applied to the contours drawn on X-ray images. Unfortunately the width of the epiglottis varies from one image to the other and PCA thus learns a spurious "inflating" component. The new model uses the epiglottis centerline plus a constant width which prevents this error.

The second major improvement concerns the use of virtual targets in the construction of the articulatory model. Virtual targets are used to separate the contribution of the tongue contour from those of the palate. The objective it to render the articulation of consonants more correctly since they require a contact between the tongue and the palate at a very precise point [38].

These two improvements of the articulatory model were used in the articulatory copy synthesis experiments [11].

The construction of models was also tackled from a data mining point of view. A robust data mining approach was designed to automatically extract complex statistically significant connections between data (e.g. interactions between more than two variables). This work could be used for data other than X-ray images [54].

6.2.2. Expressive acoustic-visual synthesis

Right now, we are investigating the state-of-the-art of the field of expressive speech and how to acquire efficiently expressive speech corpus. As a first step, we are also investigating visual acquisition techniques to track facial expression. This is the work of the visiting PhD student Andrea Bandini (from University of Bologna). Another step toward expressive speech synthesis is to have an expressive face model. In this context, the expressivity is mainly based on the dynamics. In fact, when the human facial movements are natural and accurately replicated on the 3D model, we can reach a reasonable expressivity. In this context, we are conducting new research toward an expressive talking head. In this context, we acquired a high-resolution 3D model of a human speaker and we are developing methods to animate the model using motion capture data. This was the work the master student Guillaume Gris. We also investigated the advantage of generating visual speech from sequences of 2D Images, when the 3D data is lacking [43].

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

Categorization of sounds and prosody for non-native speech is the object of the ANR+DFG project IFCASL devoted to French and German languages. Within this project, we built a bilingual corpus and started a study about the realization of (final) voicing in both languages. We also gave a training course about non-native phonetic realizations for a Spring School devoted to *Individualized centered approaches to speech processing* [63].

6.2.3.1. Bilingual speech corpus of French and German language learners

We designed a corpus of native and non-native speech for the French-German language pair, with a special emphasis on phonetic and prosodic aspects. To our knowledge there is no suitable corpus, in terms of size and coverage, currently available for this target language pair [9].

We adopted a two step process to create the corpus. Firstly, a bilingual corpus including all sounds of each language and all speech phenomena of potential interest was recorded from a few speakers (14), and analyzed. Its analysis revealed/confirmed: 1) the existence of special strategies due to sentence reading and sentence listening conditions, 2) the importance of recording duration (the recording sessions should not last more than one hour to avoid subjects' fatigue), 3) the frequence and importance of some mispronunciations (voicing problems, erroneous presence (or absence) of /h/ for German (or French) non-native speakers, rhythm ...). Secondly, we specified and collected the final corpus [24], which is focused on the problems revealed by

the preliminary corpus. One hundred speakers (50 French and 50 German speakers), beginners and advanced speakers, recorded 60 sentences in their second language and 30 in their native language, which gave a total amount of about 6000 non-native and 3000 native sentence realizations. The sentences were read in two conditions depending upon whether or not the subjects listen to a reference before producing the sentence. A small text as well as sentences devoted to focus analysis completed the corpus. The data was segmented and labelled at word and phone levels by an automatic alignment algorithm elaborated by our team (cf. 6.4.3.2). The outputs were then manually checked at the levels of phones and words (phonetic transcription) and corrections were made if necessary. In order to check the homogeneity of the corrections made by the seven annotators, phone boundaries were compared with those achieved by a golden annotator on a few sentences using the CoALT tool.

6.2.3.2. Devoicing of final obstruents by German learners

We investigated a typical example of L1-L2 interference: the realization of voiced fricatives in final position, where the opposition between voiced and unvoiced consonants is neutralized in German (with a bias towards unvoiced consonants) but not in French. As a consequence, German speakers learning French as a second language often produce unvoiced fricatives in final position instead of the expected voiced consonants. We analysed the production of French voiced fricatives for 40 non-native (beginners and advanced speakers) and 8 native speakers. We measured the ratio of locally unvoiced frames in the consonantal segment and also the ratio of consonantal duration vs. the duration of the preceding vowel. Results showed that the realizations of French fricatives by German speakers varied with speakers, speakers' level and experimental condition (there were two conditions depending on whether or not the subjects listened to a reference before producing the sentence) [23]. As could be expected we observed a continuum between typically voiced and typically unvoiced realizations, and best level speakers tend to produce more typically French realizations. Our next study will concern the perceptual identification of learners' realizations and the link between perceptual answers and acoustic cues values.

6.3. Complex statistical modeling of speech

Participants: Emmanuel Vincent, Antoine Liutkus, Denis Jouvet, Dominique Fohr, Irina Illina, Joseph Di Martino, Emad Girgis, Arseniy Gorin, Nathan Souviraà-Labastie, Luiza Orosanu, Imran Sheikh, Xabier Jaureguiberry, Baldwin Dumortier.

6.3.1. Acoustic modeling

6.3.1.1. Theory for audio source separation

Our work on audio source separation was marked by the release of version 2 of our toolbox FASST, which was demonstrated at ICASSP 2014 [65], and by the publication of a review paper about guided audio source separation for *IEEE Signal Processing Magazine* [16]. Audio source separation is an inverse problem, which requires the user to guide the separation process using prior models for the source signals and the mixing filters or for the source spectra and their spatial covariance matrices.

On the topic of the mixing parameters, we studied the impact of sparsity penalties over the mixing filters [8] and deterministic subspace constraints [10] over the spatial covariance matrices.

Modelling the spectra of the sources is a fundamental problem in source separation, that aims at catching their main features while requiring few parameters to estimate. We proposed a new framework called Kernel Additive Modelling (KAM). In contrast to Nonnegative Matrix Factorization approaches (NMF), KAM permits to model sources spectro-temporal evolutions only locally. It generalizes many methods from the state-of-the-art, including REPET (voice/music separation) and HPSS (harmonic/percussive separation) and is the first framework to settle them on principled statistical grounds. This year, we have thus been very active not only in diffusing REPET and its variants to a large audience, notably through the publication of a chapter book on the topic [58], but also by establishing many international collaborations on KAM, leading to the publication of one journal paper in IEEE TSP [13] and to two international conference papers [25], [42].

In parallel, we started a new research track on the fusion of multiple source separation techniques. In the specific case when the source spectra are modeled by NMF, the number of components of the NMF is known to have a noticeable influence on separation quality. Many methods have been proposed to select the best order for a given task. To go further, we proposed to use model averaging. As existing techniques do not allow an effective averaging, we introduced a generative model in which the number of components is a random variable and we proposed a modification to conventional variational Bayesian (VB) inference. Initial experiments showed promising results [33], [32].

6.3.1.2. Audio separation based on multiple observations

An interesting scenario for informed audio source separation is when the signals to separate can be observed through deformed references. We proposed a general approach for 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 [46], [66].

A related topic is the removal of interferences from live recordings. In this scenario, there are as many microphones as source signals, but each microphone captures not only its dedicated source, but also some interference from the other ones. We proposed a variant of KAM, called KAM for Interference Removal (KAMIR) that permits to address this scenario. The corresponding study has been achieved in collaboration with New York and Erlangen universities.

6.3.1.3. Separation and dereverberation

In order to complement source separation by dereveberation of the source signals, we devoted some work to the estimation of the reverberation time (RT60). In many situations, the room impulse response (RIR) is not available and the RT60 must be blindly estimated from a speech or music signal. Current methods often implicitly assume that reverberation dominates direct sound, which restricts their applicability to relatively small rooms or distant sound sources. We proposed a blind RT60 estimation method that is independent of the room size and the source distance and showed that the estimation error is significantly reduced even in the case when reverberation dominates [21].

6.3.1.4. Corpora for audio separation

Finally, we pursued our long-lasting efforts on the evaluation of audio source separation by providing more details about the DEMAND dataset, that is the first-ever publicly available dataset of multichannel real-world noise recordings [55]. Furthermore, we have continued our efforts on providing corpora for the evaluation of music source separation methods (notably for music/voice separation) and target at significantly extending the SiSEC corpus in 2015 to several hundreds complete recordings, to be used for the first time at SiSEC 2015.

6.3.1.5. Detailed acoustic modeling

Acoustic models aim at representing the acoustic features that are observed for the sounds of the language, as well as for non-speech events (silence, noise, ...). Currently context-dependent hidden Markov models (CD-HMM) constitute the state of the art for speech recognition. However, for text-speech alignment, simpler context-independent models are used as they provide better performance.

In conventional HMM-based approaches that rely on Gaussian mixture densities (GMM), the Gaussian components are estimated independently for each density. Thus, we have focused recent studies on enriching the acoustic models themselves in view of handling trajectory and speaker consistency in decoding. A new modeling approach was developed that takes advantage of the multiple modeling ideas and involves a sharing of parameters. The idea is to use the multiple modeling approach to partition the acoustic space according to classes (manual classes or automatic classification). Then, for each density, Gaussian components are estimated using the data associated to the classes. These class-based Gaussian components are then pooled to provide the set of Gaussian components of the density. Finally class dependent mixture weights are estimated for each density; such approach allows us to better parameterize GMM-HMM without increasing significantly the number of model parameters. Experiments on French radio broadcast news data demonstrated the improvement of the accuracy with such parameterization compared to models with a similar, or even a larger number of parameters. Another approach has been proposed that combines the structuring of the Gaussian components of the densities with respect to some data classes, with the stranded-based approach

which introduces probabilities for the transitions between the Gaussian components of the densities when moving from one frame to the next. A detailed analysis of stranded GMM was conducted on data containing different types of non-phonetic variability [29]. The combination of stranded GMM with class-structured densities was evaluated on an English connected digits task using adult and child data [27] and for phonetic decoding on a larger French telephone speech database [26]. This approach was later combined with feature normalization [28].

6.3.1.6. Robust acoustic modeling

In the framework of using speech recognition for helping communication with deaf or hard of hearing people, robustness of the acoustic modeling is investigated. Current studies relate to improving robustness with respect to speech signal level and environment noise through multicondition training and enhanced set of acoustic features.

6.3.1.7. Unsupervised acoustic model training

In previous experiments relating to the combination of speech decoder outputs for improving speech recognition performance [4], it was observed that when a forward-based and a backward-based decoder were providing a same word hypothesis, such common word hypothesis is correct in more than 90% of the cases [71]. Hence, we have investigated how such behavior can help for selecting data for unsupervised training of acoustic models. Best performance is achieved when selecting automatically transcribed data (speech segments) that have the same word hypotheses when processed by the Sphinx forward-based and the Julius backward-based transcription systems, and this selection process outperforms confidence measure based selection. Overall, selecting automatically transcribed speech segments that have the same word hypotheses for the two speech transcription systems, and adding this automatically transcribed and selected data to the manually transcribed data leads to significant word error rate reductions on the ESTER2 data (radio broadcast news) when compared to the baseline system trained only on manually transcribed speech data [34].

6.3.1.8. Score normalization

Existing techniques for robust ASR typically compensate distortion on the features or on the model parameters themselves. By contrast, a number of normalization techniques have been defined in the field of speaker verification that operate on the resulting log-likelihood scores. We provided a theoretical motivation for likelihood normalization due to the so-called "hubness" phenomenon and we evaluated the benefit of several normalization techniques on ASR accuracy for the 2nd CHiME Challenge task. We showed that symmetric normalization (S-norm) reduces the relative error rate by 43% alone and by 10% after feature and model compensation [53].

6.3.2. Linguistic modeling

6.3.2.1. Out-of-vocabulary proper name retrieval

Recognition of proper names 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. 8.1.4), 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 proper names and increase the ASR vocabulary size. We focus on exploiting the lexical context based on temporal information from diachronic documents. Our assumption is that time is an important feature for capturing name-to-context dependencies. We also studied different metrics for proper name selection in order to limit the vocabulary augmentation: a method based on Mutual Information and a new method based on cosine-similarity measure. Recognition results show a significant reduction of the proper name error rate using augmented vocabulary [30][31].

6.3.2.2. Hybrid language modeling

In the framework of using speech recognition for helping communication with deaf or hard of hearing people, the handling of out-of-vocabulary words is a critical aspect. Indeed, the size of the vocabulary is always limited (even if large or very large), and the system is not able to recognize words out of its lexicon. Such words would then be transcribed as sequences of short words which involve similar sounds as the unknown word. However the interpretation of such sequences of small word require a lot of efforts. Hence the idea of combining in a single model a set of words (the most frequent and/or most relevant for the application context) and a set of syllables. With such an approach,unknown words are usually recognized as sequences of syllables which are easier to interpret. By setting different thresholds on the confidence measures associated to the recognized words (or syllables), the most reliable word hypotheses can be identified, and they have correct recognition rates between 70% and 92% [44][45].

6.3.2.3. Music language modeling

Similarly to speech, music involves several levels of information, from the acoustic signal up to cognitive quantities such as composer style or key, through mid-level quantities such as a musical score or a sequence of chords. The dependencies between mid-level and lower- or higher-level information can be represented through acoustic models and language models, respectively. We pursued our pioneering work on music language modeling, with a particular focus on the modeling of long-term structure [20]. We also proposed a new Bayesian n-gram topic modeling and estimation technique, which we applied to genre-dependent modeling of chord sequences and to music genre classification [15].

6.3.3. Speech generation by statistical methods

6.3.3.1. Enhancing pathological voice by voice conversion techniques

Enhancing the pathological voice in order to make it more intelligible would allow persons having this kind of voice to communicate more easily with those around them. In our group we chose to improve the pathological voice by means of voice conversion techniques. Since we began this study, we have succeeded to predict the complete magnitude spectrum. In doing so, we free ourselves from the prediction of the fundamental frequency of speech (F0). Such an interesting result allows us to obtain converted speech of good audio quality. Now in order to obtain perfect conversion, we are trying, with Emad Girgis, a postdoctoral student who began his work in November 2014, to predict the phase spectrum. To achieve this goal, Emad intends to use Deep Neural Networks (DNN). We expect first results in the beginning of 2015.

6.3.3.2. Enhancing pathological voice by voice recognition techniques

Another possibility for enhancing the pathological voice is to recognize it. Othman Lachhab, a PhD student, is working on the recognition of the esophageal voice: using high order temporal derivatives combined with an Heteroscedastic Linear Discriminant Analysis (HLDA) he reached an interesting phone recognition rate of 63.59% [36]. Currently Othman, is trying to improve his results by using voice conversion techniques. Using these techniques pathological features are projected in a clean-natural speech feature space, and preliminary results exhibit an increase of 1.70% of the phone recognition rate.

6.3.3.3. F0 detection using wavelet transforms

Another possible interesting track for improving voice conversion techniques is to predict the fundamental frequency of speech. For doing so, it is necessary to have a good F0 detector. As part of her thesis, Fadoua Bahja developed many F0 detection algorithms [69] [1]. The latest, using wavelet transform for denoising the cepstrum signal, has been submitted for publication in an international journal.

6.4. Uncertainty estimation and exploitation in speech processing

Participants: Emmanuel Vincent, Dominique Fohr, Odile Mella, Denis Jouvet, Agnès Piquard-Kipffer, Dung Tran.

6.4.1. Uncertainty and acoustic modeling

In many real-world conditions, the speech signal is overlapped with noise, including environmental sounds, music, or undesired extra speech. Speech enhancement is useful but insufficient: some distortion remains in the enhanced signal which must be quantified in order not to be propagated to the subsequent feature extraction and decoding stages. The framework of uncertainty decoding assumes that this distortion has a Gaussian distribution and seeks to estimate its covariance matrix [5]. A number of uncertainty estimators and propagators have been proposed for this purpose, which typically operate on diagonal covariance matrices and are based on fixed mathematical approximations or heuristics. We obtained more accurate uncertainty estimates by propagating the full uncertainty covariance matrix and by fusing multiple uncertainty estimators [50], [51]. Overall, we obtained 18% relative 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 and propagator.

In order to motivate further work by the community, we created a new international evaluation campaign on that topic in 2011: the CHiME Speech Separation and Recognition Challenge [2]. After two successful editions in 2011 and 2013, we started working and collecting a new corpus towards the organization of a third edition to be announced in 2015.

6.4.2. 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. 8.1.5), our goal is to find the best way for displaying the speech transcription results. 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 (cf. 6.3.2.2); hence such information can be used to adjust the display of the transcription results.

We have adopted a two-step process. Firstly, we conducted a feasibility study with three hard-of-hearing persons including written display tests on print media and interviews. Secondly, we set up an experimental protocol with five hard-of-hearing persons. It included comprehension tests of 40 written sentences recorded by a French native speaker video projected onto a screen. We have also conducted parallel interviews. Their analysis revealed: (1) the interest of the participants in the project; (2) their difficulties to read International Phonetic Alphabet; (3) the importance of knowing the context of communication; (4) the need for aid in case of errors of the speech recognition system by emphasing the words that are supposed to be well recognized by the system. At this stage of the experimental period, the best display associates writing in a bold spelling the words that are supposed to be correctly recognized, and writing in a normal font using simplified French phonetics the words that are possibly wrongly recognized (according to their confidence measure). The next step will be to set up another experimental protocol in order to compare the current display in three conditions (written sentences vs written sentences with oral and lip reading vs lip reading only).

6.4.3. Uncertainty and phonetic segmentation

As described below, phonetic segmentation has been studied this year for spontaneous speech and non-native speech. Moreover, some portions (of about 30 secondes) of various speech documents have been manually annotated (checking and correction of an automatic segmentation). In the future this manually annotated data will be used to analyze the accuracy of the automatic segmentation, and also to elaborate measures that estimate the quality of the segmentation.

6.4.3.1. Alignment with spontaneous speech

Within the ANR ORFEO project (cf. 8.1.2), we addressed the problem of the alignment of spontaneous speech. The ORFEO audio files 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. For selecting the best acoustic models, we compared the alignment outputs obtained with the different acoustic models by using our tool CoALT and the manually annotated portions described above.

We also designed a new automatic grapheme-phoneme tool to generate the potential pronunciations of words and proper names. For what concerns overlapping speech, among the different orthographic transcripts corresponding to the overlapping area, we determined as the main transcript the one that best matches the audio signal, the others are kept in other tiers (in a Praat TextGrid file) with the same time boundaries.

6.4.3.2. Alignment with non-native speech

Non-native speech alignment with text is one critical step in computer assisted foreign language learning [3]. The alignment is necessary to analyze the learner's utterance, in view of providing some prosody feedback (as for example bad duration of some syllables). However, non-native speech alignment with text is much more complicated than native speech alignment. This is due to the pronunciation deviations observed on non-native speech, as for example the replacement of some target language phonemes by phonemes of the mother tongue, as well as errors in the pronunciations. Non-native speech alignment with text is currently studied in the ANR IFCASL project (see 8.1.3).

6.4.4. Uncertainty and prosody

A statistical analysis was conducted on a large annotated speech corpus to investigate the links between punctuation and automatically detected prosodic structures. The speech data comes froms radio broadcast news and TV shows, that were manually annotated during French speech transcription evaluation campaigns. These corpora contain more than 3 million words and almost 350,000 punctuation marks. The detection of the prosodic boundaries and of the prosodic structures is based on an automatic approach that integrates little linguistic knowledge and mainly uses the amplitude and the direction of the F0 slopes, as well as phone durations. A first analysis of the occurrences of the punctuation marks, with respect to various sub-corpora, has highlighted the variability among annotators. Then, a detailed analysis of the prosodic parameters with respect to the punctuation marks, whether followed or not by a pause, and of the links between the automatically detected prosodic structures and the manually annotated punctuation marks was conducted [18].

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

Besides the contracts listed below, for which MULTISPEECH is officially part of, E. Vincent was involved through his former team (PANAMA) in another 30-month bilateral research contract with Studio MAIA.

7.1.1. 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.

7.1.2. Venatech

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.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. Equipex ORTOLANG

Project acronym: ORTOLANG [6]

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 (Open Resources and TOols for LANGuage) 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 which will:

- enable a real mutualization of analysis research, of modeling and automatic treatment of the French language;
- 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;
- 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.

8.1.2. ANR ORFEO

Project acronym: ORFEO [7]

Project title: Outils et Ressources pour le Français Ecrit 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 have provided so far an automatic alignment at the word and phoneme levels for audio files from the corpus TCOF (Traitement de Corpus Oraux en Français). This corpus contains mainly spontaneous speech, recorded under various conditions with a large SNR range and a lot of overlapping speech. We tested different acoustic models and different adaptation methods for the forced speech-text alignment. Other corpora are currently being processed.

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

8.1.3. 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 has mainly focused on the design of a corpus of French sentences and text that has been recorded by German speakers learning French, recording a corpus of German sentences read by French speakers, and tools for annotating French and German corpora. Beforehand, two preliminary small corpora have been designed and recorded in order to bring to the fore the most interesting phonetic issues to be investigated in the project. In addition this preliminary work was used to test the recording devices so as to guarantee the same quality of recording in Saarbrücken and in Nancy, and to design and develop recording software.

In this project, we also provided an automatic alignment procedure at the word and phoneme levels for 4 corpora: French sentences uttered by French speakers, French sentences uttered by German speakers, German sentences uttered by French speakers, German sentences uttered by German speakers.

8.1.4. 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 (Loria)

Other partners: LIA, Synalp

Abstract: the project ContNomina 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;
- 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.

8.1.5. FUI RAPSODIE

Project acronym: RAPSODIE [8]

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 embed a speech recognition system, adapted to this task. Another application of the device will be environment vocal control for handicapped persons.

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

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.

8.1.6. ADT FASST

The Action de Développement Technologique Inria (ADT) FASST (2012–2014) was conducted by PAROLE in collaboration with the teams PANAMA and TEXMEX of Inria Rennes. It reimplemented into efficient C++ code the Flexible Audio Source Separation Toolbox (FASST) originally developed in Matlab by the METISS team of Inria Rennes. This enabled the application of FASST on larger data sets, and its use by a larger audience. The new C++ version was released in January 2014. Two modules were also developed for HTK and Kaldi in order to perform noise robust speech recognition by uncertainty decoding.

8.1.7. ADT VisArtico

The technological Development Action (ADT) Inria Visartico (2013–2015) aims at developing and improving VisArtico, an articulatory vizualisation software. In addition to improving the basic functionalities, several articulatory analysis and processing tools are being integrated. We will also work on the integration of multimodal data.

8.2. European Initiatives

8.2.1. Collaborations in European Programs, except FP7 & H2020

E. Vincent was responsible for his former team (PANAMA) of the following project.

Program: Eureka - Eurostars

Project acronym: i3DMusic

Project title: Real-time Interactive 3D Rendering of Musical Recordings

Duration: October 2010 to March 2014

Coordinator: Audionamix (FR)

Other partners: EPFL (CH), Sonic Emotion (CH)

Abstract: The i3DMusic project aims to enable real-time interactive respatialization of mono or stereo music content. This is achieved through the combination of source separation and 3D audio rendering techniques. PANAMA is responsible for the source separation work package, more precisely for designing scalable online source separation algorithms and estimating advanced spatial parameters from the available mixture.

8.3. International Initiatives

8.3.1. Inria International Partners

8.3.1.1. Informal International Partners

E. Vincent is involved as an associate member in the national Japanese JSPS Grant-in-Aid for Scientific Research project on distributed microphone arrays led by Nobutaka Ono from the National Institute of Informatics together with other partners from the University of Tsukuba and Tokyo Institute of Technology.

A. Liutkus is involved in a national project in Ireland, still at the proposal stage, on the topic of Audio Forensics, led by Derry Fitzgerald (Cork Institute of Technology). He is an associate researcher on some workpackages of this project, notably those focusing on the theory of audio source separation.

A. Liutkus is co-advisor for the Ph.D. of Donal O'Donovan (Cork Institute of Technology, Ireland), whose Ph.D. topic lies in the applications of the Kernel Additive Modelling framework to image processing.

8.3.2. Participation in other International Programs

A. Liutkus is an associate researcher in a national project in the USA, funded by the National Science Fundation (NSF) on the program "Cyber-Human Systems" (CHS) under the name "CHS:Small: Robust Interactive Audio Source Separation" and led by Bryan Pardo (Northwestern University, Chicago).

8.4. International Research Visitors

8.4.1. Visits of International Scientists

RIBAS Dayana

Date: Sep 2014 - Dec 2014

Institution: CENATAV Advanced Technologies Application Center, La Habana (Cuba)

BANDINI Andrea

Date: Oct 2014 - Mar 2015

Institution: University of Bologna, Bologna, Italy.

8.4.2. Visits to International Teams

8.4.2.1. Explorer program

VINCENT Emmanuel

Date: Jun 2014 - Aug 2014

Institution: Mitsubishi Electric Research Labs (USA)

LIUTKUS Antoine

Date: Oct 2014 - Dec 2014

Institution: BU (Turkey)

Description: This Explorer program had several objectives. First, it aims at studying several ambitious scientific problems, such as the analysis of multimodal and multirate data and also to extend Nonnegative Matrix Factorization to alpha-stable models, significantly generalizing the classical Gaussian model for audio signals. Second, this program is the occasion to build an international academic network involving researchers of the Bogazici University. It is planned to submit an ambitious proposal for a Marie-Curie International Training Network (ITN) in 2015.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

9.1.1.1. General chair, scientific chair

General chair, 4th Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA), Nancy, May 2014 (E. Vincent)

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

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

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

9.1.1.2. Organizing committee membership

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

9.1.2. Scientific events selection

9.1.2.1. Responsible of conference program committee

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

9.1.2.2. Conference program committee membership

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

RFIA'2014 - 19ème congrès national sur la Reconnaissance de Formes et l'Intelligence Artificielle (D. Jouvet)

9.1.2.3. Reviewer

JEP'2014 - 30ème édition des Journées d'Etudes sur la Parole (D. Jouvet, Y. Laprie)

RFIA'2014 - 19ème congrès national sur la Reconnaissance de Formes et l'Intelligence Artificielle (D. Jouvet)

ISMIR'2014 - International Symposium of Music Information Retrieval (A. Liutkus)

INTERSPEECH'2014 - 15th Annual Conference of the International Speech Communication Association (A. Bonneau, D. Jouvet, Y. Laprie, S. Ouni)

ICASSP'2015 - 40th International Conference on Speech, Acoustic and Signal Processing (A. Bonneau, D. Jouvet, S. Ouni)

9.1.3. Journal

9.1.3.1. Editorial board membership

IEEE Transactions on Audio, Speech, and Language Processing (E. Vincent)

Traitement du Signal (E. Vincent)

Speech Communications (D. Jouvet)

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

9.1.3.2. Reviewing activities

Computer Speech and Language (D. Jouvet)

Traitement Automatique des Langues (D. Jouvet)

Traitement du Signal (D. Jouvet)

IEEE Transactions on Audio, SPeech, and Language Processing (A. Liutkus)

Digital Signal Processing (A. Liutkus)

IEEE Letters on Signal Processing (A. Liutkus, Y. Laprie)

Signal Processing (A. Liutkus)

Speech Communication (S. Ouni)

Journal of the American Acoustical Society (Y. Laprie)

9.1.4. Miscellaneous

9.1.4.1. Tutorial

Tutorial on Informed Audio Source Separation at ICASSP 2014 in Florence, Italy (A. Ozerov and A. Liutkus and G. Richard). http://www.icassp2014.org/tutorials.html#3

9.1.4.2. Invited lecture

Les sons à domicile [62], Séminaire SAILOR "Imaginer des nouveaux lieux de vie" (E. Vincent)

Evaluation campaigns and reproducibility [68], Journée GdR ISIS "reproductibilité en traitement du signal et des images" (E. Vincent)

Acquisition and processing of X-ray and MRI data, in Spring school "Individualized-centered approaches to speech processing" in Dagstuhl, 7-9 April (Y. Laprie)

Construction of articulatory models and exploitation of articulatory data, in Spring school "Individualized-centered approaches to speech processing" in Dagstuhl, 7-9 April (Y. Laprie)

Language pathology, Séminaire "Dépistage des troubles des apprentissages" in EHESP, Rennes, 6-10 January (A. Piquard-Kipffer)

Phonological processes in beginning reading, Séminaire "Le bilan orthophonique" in ESENESR,Poitiers, 25 March (A. Piquard-Kipffer)

Relation of phonemic discrimination to learning to read, Séminaire "Les liens entre langage oral et langage écrit, la place de la discrimination phonémique" in Blaise Pascal University, Clermont-Ferrand, 6 May (A. Piquard-Kipffer)

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Master: Cadot Martine, Biostatistics, 60 hours, M1, University of Lorraine, France

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

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

Licence: Di Martino Joseph, Programming in C/C++, 45 hours, L3, 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

Master: Vincent Colotte, Introduction to Speech Analysis and Recognition, 18h, M1 , 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, 86h, L2-L3, University of Lorraine, France

Master: Odile Mella, Computer Networking, 45h, M1, University of Lorraine, France

Master : Odile Mella, Computer Networking, 14h, M2, University of Lorraine, France

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

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

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

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

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

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

Master: Piquard-Kipffer Agnès, Deaf people & reading, 9 hours,Departement Orthophonie, University of Lorraine, France

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

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

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

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

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

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

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

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, 48 hours, L1, University of Lorraine, France

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

Licence: Slim Ouni, Innovation in Computer Science, 38 hours, L3, University of Lorraine, France

Master: Slim Ouni, Game Design, 30 hours, M1, University of Lorraine, France

Master: Slim Ouni, Innovative Information Technology, 30 hours, M1 University of Lorraine, France

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

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

9.2.2. Supervision

PhD : Arseniy Gorin, "Acoustic model structuring for improving automatic speech recognition performance", University of Lorraine, 26 November 2014, Denis Jouvet.

PhD in progress: Othman Lachhab, "Pathological voice recognition using voice conversion techniques", November 2010, El Hassane Ibn Elhaj and Joseph Di Martino.

PhD in progress : Dung Tran, "Uncertainty handling for noise robust automatic speech recognition", December 2012, Emmanuel Vincent and Denis Jouvet.

PhD in progress : Luiza Orosanu, "Speech recognition for helping communication for deaf or hard of hearing people", December 2012, Denis Jouvet.

PhD in progress: Nathan Souviraà-Labastie, "Localisation et séparation de sources sonores pour la reconnaissance de la parole en environnement réel", University Rennes 1, January 2013, Frédéric Bimbot and Emmanuel Vincent.

PhD in progress: Xabier Jaureguiberry, "Fusion et optimisation de modèles pour la séparation de sources audio", Télécom ParisTech, February 2013, Gaël Richard and Emmanuel Vincent.

PhD in progress: Imran Sheikh, "Exploitation du contexte pour la reconnaissance de noms propres dans les documents diachroniques", January 2014, Irina Illina.

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.

9.2.3. Juries

Participation in PhD thesis Jury for Stefan Ziegler (University of Rennes 1, January 2014), D. Jouvet, reviewer.

Participation in PhD thesis Jury for Nicolas López (Télécom ParisTech, July 2014), E. Vincent, reviewer.

Participation in PhD thesis Jury for Thiago Fraga da Silva (University of Paris-South, September 2014), D. Jouvet, reviewer.

Participation in PhD thesis Jury for Antti Hurmalainen (Tampere University of Technology, Finland, October 2014), E. Vincent, reviewer.

Participation in PhD thesis Jury for Raphaël Laurent (University of Grenoble, October 2014), Y. Laprie, reviewer.

Participation in PhD thesis Jury for Lucie Steiblé (University of Strasbourg, December 2014), Y. Laprie, reviewer.

Participation in PhD thesis Jury for Fayssal Bouarourou (University of Strasbourg, December 2014), Y. Laprie.

9.2.4. Participation to external committees

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

Member of a Selection Committee of University of Lorraine (UFR Math Info), LORIA, Y. Laprie

Elected member of the Conseil du Pôle Scientifique AM2I of University of Lorraine, Y. Laprie

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

Member of the Conseil du Pôle Scientifique AM2I of University of Lorraine, A. Piquard-Kipffer

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

9.2.5. Participation to local committees

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

Member of the Comipers Enseignant, E. Vincent

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

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

member of the "Commission locale dévelopement durable", D.Fohr

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

Member of DFD staff, I. Illina

9.3. Popularization

Panelist at a public meeting of the SAILOR consortium on the topic of “Imagining new places to live”, University of Lorraine, April 2014 (E. Vincent).

Demonstration at Journée Sciences et Musique, IRISA Rennes, October 2014 (N. Souviraà-Labastie, E. Vincent).

Vulgarization of music signal processing at Inria to students of the Lycée Loritz, June 2014 (Y. Laprie, A. Liutkus)

10. 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

Publications of the year

Doctoral Dissertations and Habilitation Theses

[7] A. GORIN. *Acoustic Model Structuring for Improving Automatic Speech Recognition Performance*, University of Lorraine, November 2014, https://hal.inria.fr/tel-01102029

Articles in International Peer-Reviewed Journals

[8] A. BENICHOUX, L. S. R. SIMON, E. VINCENT, R. GRIBONVAL. *Convex regularizations for the simultaneous recording of room impulse responses*, in "IEEE Transactions on Signal Processing", January 2014 [*DOI :* 10.1109/TSP.2014.2303431], https://hal.inria.fr/hal-00934941

[9] C. FAUTH, A. BONNEAU, O. MELLA, V. COLOTTE, D. FOHR, D. JOUVET, Y. LAPRIE, J. TROUVAIN. *Constitution d'un Corpus de Français Langue Etrangère destiné aux Apprenants Allemands*, in "SHS Web of Conferences", July 2014, vol. 8, 14 p. [*DOI :* 10.1051/SHSCONF/20140801186], https://hal.inria.fr/hal-01080630

[10] 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", July 2014, https://hal.inria.fr/hal-01020255

[11] Y. LAPRIE, R. SOCK, B. VAXELAIRE, B. ELIE. *Comment faire parler les images aux rayons X du conduit vocal ?*, in "SHS Web of Conferences", July 2014, vol. 8, 14 p. [*DOI :* 10.1051/SHSCONF/20140801344], https://hal.inria.fr/hal-01059887

[12] N. LIU, A. LIUTKUS, J.-F. AUBRY, L. MARSAC, M. TANTER, L. DAUDET. *Random Calibration for Accelerating MR-ARFI Guided Ultrasonic Focusing in Transcranial Therapy*, in "Physics in Medicine and Biology", January 2015, vol. 60, n^{o} 3, 21 p. [*DOI :* 10.1088/0031-9155/60/3/1069], https://hal.inria.fr/hal-01104616

[13] A. LIUTKUS, D. FITZGERALD, Z. RAFII, B. PARDO, L. DAUDET. *Kernel Additive Models for Source Separation*, in "IEEE Transactions on Signal Processing", June 2014 [*DOI :* 10.1109/TSP.2014.2332434], https://hal.inria.fr/hal-01011044

[14] A. LIUTKUS, D. MARTINA, S. POPOFF, G. CHARDON, O. KATZ, G. LEROSEY, S. GIGAN, L. DAUDET, I. CARRON. *Imaging With Nature: Compressive Imaging Using a Multiply Scattering Medium*, in "Scientific Reports", July 2014, vol. 4 [*DOI :* 10.1038/SREP05552], https://hal.inria.fr/hal-01025647

[15] S. RACZYNSKI, E. VINCENT. *Genre-based music language modelling with latent hierarchical Pitman-Yor process allocation*, in "IEEE/ACM Transactions on Audio, Speech, and Language Processing", January 2014, vol. 22, n^{o} 3, pp. 672-681, https://hal.inria.fr/hal-00804567

[16] E. VINCENT, N. BERTIN, R. GRIBONVAL, F. BIMBOT. *From blind to guided audio source separation: How models and side information can improve the separation of sound*, in "IEEE Signal Processing Magazine", May 2014, vol. 31, n^{o} 3, pp. 107-115, https://hal.inria.fr/hal-00922378

Invited Conferences

[17] E. VINCENT, A. SINI, F. CHARPILLET. *Audio source localization by optimal control of a mobile robot*, in "40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", Brisbane, Australia, April 2015, https://hal.inria.fr/hal-01103949

International Conferences with Proceedings

[18] K. BARTKOVA, D. JOUVET. *Links between Manual Punctuation Marks and Automatically Detected Prosodic Structures*, in "Speech Prosody 2014", Dublin, Ireland, May 2014, https://hal.archives-ouvertes.fr/hal-00998031

[19] J. BELIAO, A. LIUTKUS. *OOPS: une approche orientée objet pour l'interrogation et l'analyse linguistique de l'interface prosodie/syntaxe/discours*, in "4e Congrès Mondial de Linguistique Française", Berlin, Germany, July 2014, vol. 8, pp. 2565-2581 [*DOI :* 10.1051/SHSCONF/20140801273], https://hal.archives-ouvertes.fr/hal-01053422

[20] F. BIMBOT, G. SARGENT, E. DERUTY, C. GUICHAOUA, E. VINCENT. *Semiotic Description of Music Structure: an Introduction to the Quaero/Metiss Structural Annotations*, in "AES 53rd International Conference on Semantic Audio", London, United Kingdom, January 2014, 12 p. , P1-1, https://hal.archives-ouvertes.fr/hal-00931859

[21] B. DUMORTIER, E. VINCENT. *Blind RT60 estimation robust across room sizes and source distances*, in "IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)", Firenze, Italy, May 2014, https://hal.inria.fr/hal-00941061

[22] B. ELIE, Y. LAPRIE. *Audiovisual to area and length functions inversion of human tract* , in "Eusipco 2014", Lisbonne, Portugal, September 2014, https://hal.inria.fr/hal-01096547

[23] C. FAUTH, A. BONNEAU. *L1-L2 interference: the case of devoicing of French voiced obstruents in final position by German learners - Pilot study*, in "International Workshop on Multilinguality in Speech Research: Data, Methods and Models", Dagstuhl, Germany, Bernd Möbius et Jürgen Trouvain, Université de la Sarre, Allemagne, April 2014, https://hal.inria.fr/hal-01095183

[24] C. FAUTH, A. BONNEAU, F. ZIMMERER, J. TROUVAIN, B. ANDREEVA, V. COLOTTE, D. FOHR, D. JOUVET, J. JÜGLER, Y. LAPRIE, O. MELLA, B. MÖBIUS. *Designing a Bilingual Speech Corpus for French and German Language Learners: a Two-Step Process*, in "LREC - 9th Language Resources and Evaluation Conference", Reykjavik, Iceland, The European Language Resources Association, May 2014, https://hal.inria.fr/hal-00979026

[25] D. FITZGERALD, A. LIUTKUS, Z. RAFII, B. PARDO, L. DAUDET. *Harmonic/Percussive Separation Using Kernel Additive Modelling*, in "IET Irish Signals & Systems Conference 2014", Limerick, Ireland, June 2014, https://hal.inria.fr/hal-01000001

[26] A. GORIN, D. JOUVET. *Component Structuring and Trajectory Modeling for Speech Recognition*, in "Interspeech", Singapoore, Singapore, September 2014, https://hal.inria.fr/hal-01063653

[27] A. GORIN, D. JOUVET. *Explicit trajectories and speaker class modeling for child and adult speech recognition*, in "XXXème édition des Journées d'Etudes sur la Parole", Le Mans, France, June 2014, https://hal.inria.fr/hal-01080343

[28] A. GORIN, D. JOUVET. *Structured GMM Based on Unsupervised Clustering for Recognizing Adult and Child Speech*, in "SLSP - 2nd International Conference on Statistical Language and Speech Processing", Grenoble, France, October 2014, pp. 108 - 119 [*DOI :* 10.1007/978-3-319-11397-5_8], https://hal.inria.fr/hal-01090472

[29] A. GORIN, D. JOUVET, E. VINCENT, D. TRAN. *Investigating Stranded GMM for Improving Automatic Speech Recognition*, in "4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA 2014)", Nancy, France, May 2014, https://hal.inria.fr/hal-01003054

[30] I. ILLINA, D. FOHR, G. LINARES. *Extension du vocabulaire d'un système de transcription avec de nouveaux noms propres en utilisant un corpus diachronique*, in "Journées d'Etude sur la parole", Le Mans, France, June 2014, https://hal.inria.fr/hal-01092214

[31] I. ILLINA, D. FOHR, G. LINARES. *Proper Name Retrieval from Diachronic Documents for Automatic Speech Transcription using Lexical and Temporal Context*, in "Workshop on Speech, Language and Audio in Multimedia", Penang, Malaysia, September 2014, https://hal.inria.fr/hal-01092224

[32] X. JAUREGUIBERRY, E. VINCENT, G. RICHARD. *Multiple-order non-negative matrix factorization for speech enhancement*, in "Interspeech", Singapore, June 2014, 4 p. , https://hal.archives-ouvertes.fr/hal-01023399

[33] X. JAUREGUIBERRY, E. VINCENT, G. RICHARD. *Variational Bayesian model averaging for audio source separation*, in "SSP (IEEE Workshop on Statistical Signal Processing)", Australia, June 2014, 4 p. , https://hal.archives-ouvertes.fr/hal-00986909

[34] D. JOUVET, D. FOHR. *About Combining Forward and Backward-Based Decoders for Selecting Data for Unsupervised Training of Acoustic Models*, in "INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association", Singapour, Singapore, September 2014, https://hal.inria.fr/hal-01090483

[35] S. KIRBIZ, A. OZEROV, A. LIUTKUS, L. GIRIN. *Perceptual coding-based informed source separation*, in "22nd European Signal Processing Conference (EUSIPCO-2014)", Lisbonne, Portugal, September 2014, https://hal.inria.fr/hal-01016314

[36] O. LACHHAB, J. DI MARTINO, E. H. IBN ELHAJ, A. HAMMOUCH. *Improving the recognition of pathological voice using the discriminant HLDA transformation*, in "3rd International IEEE Colloquium on Information Science and Technology", Tetuan-Chefchaouen, Morocco, October 2014, https://hal.inria.fr/hal-01093309

[37] Y. LAPRIE, M. ARON, M.-O. BERGER, B. WROBEL-DAUTCOURT. *Studying MRI acquisition protocols of sustained sounds with a multimodal acquisition system*, in "10th International Seminar on Speech Production (ISSP)", Köln, Germany, May 2014, https://hal.inria.fr/hal-01002121

[38] Y. LAPRIE, B. VAXELAIRE, M. CADOT. *Geometric articulatory model adapted to the production of consonants*, in "10th International Seminar on Speech Production (ISSP)", Köln, Germany, May 2014, https://hal.inria.fr/hal-01002125

[39] 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

[40] 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

[41] A. LIUTKUS, D. MARTINA, S. GIGAN, L. DAUDET. *Compressed sensing under strong noise. Application to imaging through multiply scattering media*, in "European Signal Processing Conference (EUSIPCO)", Lisbon, Portugal, September 2014, https://hal.inria.fr/hal-01074786

[42] A. LIUTKUS, Z. RAFII, B. PARDO, D. FITZGERALD, L. DAUDET. *Kernel Spectrogram models for source separation*, in "HSCMA", Nancy, France, May 2014, https://hal.inria.fr/hal-00959384

[43] U. MUSTI, S. OUNI, Z. ZIHENG. *3D Visual Speech Animation from Image Sequences*, in "Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)", Bangalore, India, ACM, December 2014, https://hal.archives-ouvertes.fr/hal-01086073

[44] L. OROSANU, D. JOUVET. *Combining words and syllables for speech transcription*, in "XXXème édition des Journées d'Etudes sur la Parole", Le Mans, France, June 2014, https://hal.inria.fr/hal-01080351

[45] L. OROSANU, D. JOUVET. *Hybrid language models for speech transcription*, in "INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association", Singapour, Singapore, September 2014, https://hal.inria.fr/hal-01090478

[46] N. SOUVIRAÀ-LABASTIE, A. OLIVERO, E. VINCENT, F. BIMBOT. *Audio source separation using multiple deformed references*, in "Eusipco", Lisboa, Portugal, September 2014, https://hal.inria.fr/hal-01017571

[47] 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)", Brisbane, Australia, April 2015, https://hal.archives-ouvertes.fr/hal-01108675

[48] I. STEINER, P. KNOPP, S. MUSCHE, A. SCHMIEDEL, A. BRAUN, S. OUNI. *Investigating the effects of posture and noise on speech production*, in "10th International Seminar on Speech Production (ISSP)", Cologne, Germany, Susanne Fuchs, Martine Grice, Anne Hermes, Leonardo Lancia, Doris Mücke, May 2014, https://hal.archives-ouvertes.fr/hal-01086066

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

[50] D. TRAN, E. VINCENT, D. JOUVET. *Extension of uncertainty propagation to dynamic MFCCs for noise robust ASR* , in "IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)", Florence, Italy, May 2014, https://hal.inria.fr/hal-00954654

[51] D. TRAN, E. VINCENT, D. JOUVET. *Fusion of Multiple Uncertainty Estimators and Propagators for Noise Robust ASR*, in "IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)", Florence, Italy, May 2014, https://hal.inria.fr/hal-00955185

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

[53] E. VINCENT, A. GKIOKAS, D. SCHNITZER, A. FLEXER. *An investigation of likelihood normalization for robust ASR*, in "Interspeech", Singapore, Singapore, September 2014, https://hal.inria.fr/hal-01006142

National Conferences with Proceedings

[54] M. CADOT, Y. LAPRIE. *Méthodologie 3-way d'extraction d'un modèle articulatoire de la parole à partir des données d'un locuteur*, in "Atelier Fouille de Données Complexes des 14èmes Journées Francophones "Extraction et Gestion des Connaissances"", Rennes, France, January 2014, pp. 1-12, https://hal.archives-ouvertes.fr/hal-00934436

[55] J. THIEMANN, E. VINCENT, S. VAN DE PAR. *Spatial properties of the DEMAND noise recordings*, in "40th Annual German Congress on Acoustics (DAGA 2014)", Oldenburg, Germany, March 2014, https://hal.inria.fr/hal-00985979

Conferences without Proceedings

[56] P.-A. VUISSOZ, F. ODILLE, Y. LAPRIE, E. VINCENT, G. HOSSU, J. FELBLINGER. *Speech Cine SSFP with optical microphone synchronization and motion compensated reconstruction*, in "ISMRM Workshop on Motion Correction in MRI", Tromso, Norway, July 2014, https://hal.inria.fr/hal-00994526

[57] 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*, in "2e Congrès de la SFRMBM", Grenoble, France, March 2015, https://hal.inria.fr/hal-01104230

Scientific Books (or Scientific Book chapters)

[58] Z. RAFII, A. LIUTKUS, B. PARDO. *REPET for Background/Foreground Separation in Audio*, in "Blind Source Separation", G. NAIK, W. WANG (editors), Springer Berlin Heidelberg, 2014, pp. 395-411 [*DOI :* 10.1007/978-3-642-55016-4_14], https://hal.inria.fr/hal-01025563

Research Reports

[59] R. BADEAU, A. LIUTKUS. *Proof of Wiener-like linear regression of isotropic complex symmetric alpha-stable random variables*, September 2014, https://hal.archives-ouvertes.fr/hal-01069612

[60] J. LE ROUX, E. VINCENT. *A categorization of robust speech processing datasets*, September 2014, n° Mitsubishi Electric Research Labs TR2014-116, https://hal.inria.fr/hal-01063805

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

Scientific Popularization

[62] E. VINCENT. *Les sons à domicile*, April 2014, Séminaire SAILOR "Imaginer des nouveaux lieux de vie", Séminaire SAILOR "Imaginer des nouveaux lieux de vie", https://hal.inria.fr/hal-00977674

Other Publications

[63] A. BONNEAU. *Phonetic variation in non-native speech*, April 2014, Spring School : "Individual-centered Approaches to Speech Processing", https://hal.inria.fr/hal-01095804

[64] A. PIQUARD-KIPFFER. *Critères d'évaluation d'un album numérique pour des enfants en difficulté de langage*, December 2014, pp. 287-309, In M. Frisch (Eds) Le réseau Idéki : objets de recherche, d'éducation et de formation émergents, problématisés, mis en tension, réélaborés. Préface de Joël Lebeaume. Paris : L'harmattan, Collection I.D, 287-309, https://hal.inria.fr/hal-01097278

[65] Y. SALAÜN, E. VINCENT, N. BERTIN, N. SOUVIRAÀ-LABASTIE, X. JAUREGUIBERRY, D. T. TRAN, F. BIMBOT. *The Flexible Audio Source Separation Toolbox Version 2.0*, May 2014, ICASSP, https://hal.inria.fr/hal-00957412

[66] N. SOUVIRAÀ-LABASTIE, A. OLIVERO, E. VINCENT, F. BIMBOT. *Multi-channel audio source separation using multiple deformed references*, November 2014, https://hal.inria.fr/hal-01070298

[67] D. T. TRAN, E. VINCENT, D. JOUVET. *Nonparametric uncertainty estimation and propagation for noise robust ASR*, January 2015, https://hal.inria.fr/hal-01114329

[68] E. VINCENT. *Evaluation campaigns and reproducibility*, January 2014, Journée GdR ISIS "reproductibilité en traitement du signal et des images", https://hal.inria.fr/hal-00927741

References in notes

[69] F. BAHJA. *Détection du fondamental de la parole en temps réel : application aux voix pathologiques*, Université Mohammed V-Agdal UFR Informatique et Télécommunications Laboratoire LRIT Unité associée au CNRST, URAC 29, Faculté des sciences, June 2013, https://tel.archives-ouvertes.fr/tel-00927147

[70] D. FOHR, O. MELLA. *CoALT: A Software for Comparing Automatic Labelling Tools*, in "Language Resources and Evaluation LREC 2012", Istanbul, Turkey, May 2012, pp. 325-328, https://hal.archives-ouvertes.fr/hal-00761781

[71] D. JOUVET, D. FOHR. *Analysis and Combination of Forward and Backward based Decoders for Improved Speech Transcription*, in "TSD - 16th International Conference on Text, Speech and Dialogue - 2013", Pilsen, Czech Republic, I. HABERNAL, V. MATOUŠEK (editors), Lecture Notes in Artificial Intelligence, Springer Verlag, September 2013, vol. 8082, pp. 84-91, https://hal.inria.fr/hal-00834296

[72] S. OUNI, L. MANGEONJEAN, I. STEINER. *VisArtico: a visualization tool for articulatory data*, in "13th Annual Conference of the International Speech Communication Association - InterSpeech 2012", Portland, OR, United States, September 2012, https://hal.inria.fr/hal-00730733