IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2013

Project-Team PAROLE

Analysis, perception and recognition of speech

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

RESEARCH CENTER
Nancy - Grand Est

THEME
Language, Speech and Audio

Table of contents

1. Members **1**
2. Overall Objectives **2**
3. Research Program **3**
3.1. Introduction 3
3.2. Speech analysis and synthesis 3
3.2.1. Oral comprehension 3
3.2.1.1. Computer-assisted learning of prosody 4
3.2.1.2. Esophageal voices 4
3.2.2. Acoustic-to-articulatory inversion 4
3.2.3. Strategies of labial coarticulation 4
3.2.4. Speech synthesis 5
3.2.4.1. Text-to-speech synthesis 5
3.2.4.2. Acoustic-visual speech synthesis 5
3.3. Automatic speech recognition 6
3.3.1. Acoustic features and models 6
3.3.2. Robustness and invariance 6
3.3.3. Segmentation 7
3.3.4. Speech/text alignment 7
3.4. Speech to Speech Translation and Language Modeling 7
3.4.1. Word translation 8
3.4.2. Phrase translation 8
3.4.3. Language model 8
3.4.4. Decoding 8
4. Application Domains **8**
5. Software and Platforms **9**
5.1. WinSnoori 9
5.2. JSnoori 9
5.3. Xarticulators 9
5.4. SUBWEB 10
5.5. ANTS 10
5.6. CoALT 10
5.7. TTS SoJA 10
5.8. JCorpusRecorder 11
5.9. VisArtico 11
5.10. FASST 11
6. New Results **11**
6.1. Speech analysis and synthesis 11
6.1.1. Acoustic-to-articulatory inversion 12
6.1.1.1. Construction of articulatory models 12
6.1.1.2. Articulatory copy synthesis 12
6.1.2. Using articulography for speech animation 13
6.1.3. Acoustic analyses of non-native speech 13
6.1.4. Speech synthesis 13
6.1.5. Phonemic discrimination evaluation in language acquisition and in dyslexia and dysphasia 13
6.1.6. Enhancement of esophageal voice 14
6.1.6.1. Pitch detection 14
6.1.6.2. Real-time pitch detection for application to pathological voices 14
6.1.6.3. Voice conversion techniques applied to pathological voice repair 14

6.1.6.4. Signal reconstruction from short-time Fourier transform magnitude spectra 15
6.1.7. Audio source separation 15
6.2. Automatic speech recognition 15
6.2.1. Detailed acoustic modeling 15
6.2.2. Noise-robust speech recognition 16
6.2.3. Linguistic modeling 16
6.2.3.1. Random indexing 16
6.2.3.2. Continuous language models 16
6.2.3.3. Linguistic units for embedded systems 17
6.2.3.4. OOV proper name retrieval 17
6.2.4. Speech transcription 17
6.2.4.1. Combining recognizers 17
6.2.4.2. Spontaneous speech 18
6.2.4.3. Towards a structured output 18
6.2.5. Speech/text alignment 18
6.2.5.1. Alignment with non-native speech 18
6.2.5.2. Alignment with spontaneous speech 19
6.3. Machine translation and language modeling 19
6.3.1. Language modeling 19
6.3.1.1. Vocabulary selection 19
6.3.1.2. Music language modeling 19
6.3.2. Quality estimation of machine translation 19
6.3.3. Comparable corpora and multilingual sentiment analysis 20
6.3.4. Machine translation of arabic dialect 20
7. Bilateral Contracts and Grants with Industry **20**
8. Partnerships and Cooperations **20**
8.1. National initiatives 20
8.1.1. Equipex ORTOLANG 20
8.1.2. ANR ARTIS 21
8.1.3. ANR ViSAC 21
8.1.4. ANR ORFEO 22
8.1.5. ANR-DFG IFCASL 22
8.1.6. ANR ContNomina 23
8.1.7. FUI RAPSODIE 23
8.1.8. ADT FASST 23
8.1.9. ADT VisArtico 23
8.2. European initiatives 24
8.2.1.1. Interreg Allegro 24
8.2.1.2. Eureka - Eurostars i3DMusic 24
8.3. International initiatives 24
8.4. International research visitors 25
9. Dissemination **25**
9.1. Scientific animation 25
9.2. Teaching - supervision - juries 26
9.2.1. Teaching 26
9.2.2. Supervision 27
9.2.3. Juries 27
9.2.4. Participation to external committees 28
9.2.5. Participation to local committees 29
9.3. Popularization 29
10. Bibliography **29**

Project-Team PAROLE

Keywords: Natural Language, Speech, Recognition, Statistical Methods, Perception, Signal Processing

Creation of the Project-Team: 2001 May 01*, updated into Team:* 2014 January 01.

1. Members

Research Scientists

Yves Laprie [Team leader, CNRS, Senior Researcher, HdR]
Anne Bonneau [CNRS, Researcher]
Dominique Fohr [CNRS, Researcher]
Denis Jouvet [Inria, Senior Researcher, HdR]
Emmanuel Vincent [Inria, Researcher, from Jan 2013, HdR]

Faculty Members

Vincent Colotte [Univ. of Lorraine, Associate Professor]
Joseph Di Martino [Univ. of Lorraine, Associate Professor]
Irina Illina [Univ. of Lorraine, Associate Professor, HdR]
David Langlois [Univ. of Lorraine, Associate Professor]
Odile Mella [Univ. of Lorraine, Associate Professor]
Slim Ouni [Univ. of Lorraine, Associate Professor, HdR]
Agnès Piquard-Kipffer [Univ. Lorraine, Associate Professor]
Kamel Smaïli [Univ. of Lorraine, Professor, HdR]

Engineers

Ilef Ben Farhat [Inria, from Nov 2013]
Jérémy Miranda [Inria, granted by OSEO Innovation, from Sep 2013]
Yann Salaun [Inria]

PhD Students

Julie Busset [CNRS, granted by ANR, until Apr 2013]
Arseniy Gorin [Inria, CORDI-S]
Xabier Jaureguiberry [Institut Telecom-Institut Carnot, from Feb 2013]
Nathan Souviraà-Labastie [Institut Telecom-Région Bretagne, from Nov 2013]
Alex Mesnil [MenRT-ENS Lyon, Univ. of Lorraine, until Aug 2013]
Utpala Musti [Inria, CORDI-S, until Feb 2013]
Cyrine Nasri [ATER, Univ. of Lorraine]
Luiza Orosanu [Inria, CORDI-C, granted by OSEO Innovation]
Motaz Saad [Univ of Gaza]
Dung Tran [Inria, CORDI-S]

Post-Doctoral Fellows

Benjamin Elie [Inria, from Oct 2013]
Camille Fauth [CNRS, from Mar 2013]

Visiting Scientists

Mouhcine Chami [INPT, Maroco, from Jun 2013 until Jun 2013]
Karima Meftouh [Annaba University, until Oct 2013]

Administrative Assistants

Sylvie Musilli [Univ. of Lorraine, from Jan 2013]
Helene Zganic [Inria]

Others

Bruno Andriamiarina Miharimanana [Univ. of Lorraine, intern , from Feb 2013 until Jul 2013]
Émilien Casano [Univ. of Lorraine, intern , from Apr 2013 until Jun 2013]
Renaud Clement [Univ. of Lorraine, intern, from Mar 2013 until Jul 2013]
Thierno Dia [Inria, intern, from Mar 2013 until Jul 2013]
Marie Dormont [Inria, intern, from Oct 2013 until Dec 2013]
Baldwin Dumortier [Inria, intern , from Apr 2013 until Sep 2013]
Hugo Mathieu [Univ. of Lorraine, intern , from Apr 2013 until Jun 2013]
Jérémy Miranda [Univ. of Lorraine, intern, from Apr until Jun Sep 2013]
Imane Nkairi [Univ. of Lorraine, intern , from Feb 2013 until Jul 2013]
Romain Thomas [Univ. of Lorraine, intern, from Apr 2013 until Jun 2013]
Christophe Varray [ENS Cachan, intern, from Mar 2013 until Jul 2013]
Othman Zennaki [Univ. of Lorraine, intern, from Feb 2013 until Jun 2013]

2. Overall Objectives

2.1. Overall Objectives

PAROLE is a joint project to Inria, CNRS, University of Lorraine through the LORIA laboratory (UMR 7503). The purpose of our project is to automatically process speech signals to understand their meaning, and to analyze and enhance their acoustic structure. It inscribes within the view of offering efficient vocal technologies and necessitates works in analysis, perception and automatic recognition (ASR) of speech.

Our activities are structured in three topics:

- **Speech analysis and synthesis.** Our works are concerned with automatic extraction and perception of acoustic and visual cues, acoustic-to-articulatory inversion and speech synthesis. These themes give rise to a number of ongoing and future applications especially in the domain of foreign language learning.
- **Enriched automatic speech recognition.** Our works are concerned with stochastic models (HMM [1] and Bayesian networks), semi-supervised and smoothed training of these stochastic models, adaptation of a recognition system to important variability sources, and with enriching the output of speech recognition with higher-level information such as syntactic structure and punctuation marks. These topics give also rise to a number of ongoing and future applications: automatic transcription, speech/text alignment, audio indexing, keyword spotting, foreign language learning, dialog systems, vocal services...
- **Speech to speech translation and language modeling.** This axis concerns statistical machine translation. The objective is to translate speech from a source language to any target language. The main activity of the group which is in charge of this axis is to propose an alternative method to the classical five IBM's models. This activity should conduct to several applications: e-mail speech to text, translation of movie subtitles.

Our pluridisciplinary scientific culture combines works in phonetics, pattern recognition and artificial intelligence. This pluridisciplinarity turns out to be a decisive asset to address new research topics, particularly language learning that simultaneously require competences in automatic speech recognition and phonetics.

[1] Hidden Markov Models

3. Research Program

3.1. Introduction

Research in speech processing gave rise to two kinds of approaches:

- research that aims at explaining how speech is produced and perceived, and that therefore includes physiological aspects (vocal tract control), physical (speech acoustics), psychoacoustics (peripheral auditory system), and cognitive aspects (building sentences),
- research aiming at modeling the observation of speech phenomena (spectral analysis, stochastic acoustic or linguistic models).

The former research topic is motivated by the high specificity of speech among other acoustical signals: the speech production system is easily accessible and measurable (at least at first approach); acoustical equations are reasonably difficult from a mathematical point of view (with simplifications that are moderately restrictive); sentences built by speakers are governed by vocabulary and grammar of the considered language. This led acousticians to develop research aiming at generating artificial speech signals of good quality, and phoneticians to develop research aiming at finding out the origin of speech sound variability and at explaining how articulators are utilized, how sounds of a language are structured and how they influence each other in continuous speech. Lastly, that led linguists to study how sentences are built. Clearly, this approach gives rise to a number of exchanges between theory and experimentation and it turns out that all these aspects of speech cannot be mastered easily at the same time.

Results available on speech production and perception do not enable using an analysis by synthesis approach for automatic speech recognition. Automatic speech recognition thus gives rise to a second approach that consists in modeling observations of speech production and perception. Efforts focused onto the design of numerical models (first simple vectors of spectral shapes and now stochastic or neural models) of word or phoneme acoustical realizations, and onto the development of statistical language models.

These two approaches are complementary; the latter borrows theoretical results on speech from the former, which, in its turn, borrows some numerical methods. Spectral analysis methods are undoubtedly the domain where exchanges are most marked. The simultaneous existence of these two approaches is one of the particularities of speech research conducted in Nancy and we intend to enhance exchanges between them. These exchanges will probably grow in number because of new applications like: **(i)** computer aided foreign language learning which requires both reliable automatic speech recognition and fine acoustic and articulatory speech analysis, **(ii)** automatic recognition of spontaneous speech which requires robustness against noise and speaker variability.

3.2. Speech analysis and synthesis

Our research activities focus on acoustical and perceptual cues of speech sounds, speech modifications and acoustic-to-articulatory inversion. Our main applications concern the improvement of the oral component of language learning, speech synthesis and esophageal voices.

3.2.1. Oral comprehension

We developed tools to improve speech perception and production, and made perceptual experiments to prove their efficiency in language learning. These tools are also of interest for hearing impaired people, as well as for normally hearing people in noisy environments and also for children who learn to read (children who have language disabilities without cognitive deficit or hearing impairment and "normal" children).

3.2.1.1. Computer-assisted learning of prosody

We are studying automatic detection and correction of prosodic deviations made by a learner of a foreign language. This work implies three different tasks: (a) the detection of the prosodic entities of the learner's realization (lexical accent, intonative patterns), (b) the evaluation of the deviations, by comparison with a model, and (c) their corrections, both verbal and acoustic. This last kind of feedback is directly done on the learner's realization: the deviant prosodic cues are replaced by the prosodic cues of the reference. The identification and correction tasks use speech analysis and modification tools developed in our team. Perceptual experiments have shown the interest of speech modifications, both for L2 learners and people with hearing deficiencies [30].

Within the framework of the project Intonale, we also investigate the impact of a language intonational characteristics on the perception and production of the intonation of a foreign language.

3.2.1.2. Esophageal voices

It is possible for laryngectomees to learn a substitution voice: the esophageal voice. This voice is far from being natural. It is characterized by a weak intensity, a background noise that bothers listening, and a low pitch frequency. A device that would convert an esophageal voice to a natural voice would be very useful for laryngectomees because it would be possible for them to communicate more easily. Such natural voice restitution techniques would ideally be implemented in a portable device.

3.2.2. Acoustic-to-articulatory inversion

Acoustic-to-articulatory inversion aims at recovering the articulatory dynamics from speech signal that may be supplemented by images of the speaker face. Potential applications concern low bit rate speech coding, automatic speech recognition, speech production disorders assessment, articulatory investigations of phonetics, talking heads and articulatory feedback for language acquisition or learning.

One of the major difficulties of inversion is that an infinity of vocal tract shapes can give rise to the same speech spectrum. Acoustic-to-articulatory inversion methods are categorized into two families:

- methods that optimize a function generally combining speaker's articulatory effort and acoustical distance between natural and synthesized speech. They exploit constraints allowing the number of possible vocal tract shapes to be reduced.
- table look-up methods resting on an articulatory codebook of articulatory shapes indexed by their acoustical parameters (generally formant frequencies). After possible shapes have been recovered at each time, an optimization procedure is used to find an inverse solution in the form of an optimal articulatory path.

As our contribution only concerns inversion, we accepted widely used articulatory synthesis methods. We therefore chose Maeda's articulatory model, the acoustical-electrical analogy to compute the speech spectrum and the spatio-temporal method proposed by Maeda to generate the speech signal. As regards inversion, we chose Maeda's model to constrain vocal tract shapes because this model guarantees that synergy and compensation articulatory phenomena are still possible, and consequently, that articulatory deformations close to those of a human speaker may be recovered. The most important challenges in this domain are the inversion of any class of speech sounds and to perform inversion from standard spectral data, Mel Frequency Cepstral Coefficients (MFCC) for instance. Indeed at present, only vowels and sequences of vowels can be inverted, and only some attempts concern fricatives sounds. Moreover, most of the inversion techniques use formant frequencies as input data although formants cannot be extracted from speech easily and reliably.

3.2.3. Strategies of labial coarticulation

The investigation of labial coarticulations strategies is a crucial objective with the view of developing a talking head which would be understandable by lip readers, especially deaf persons.

In the long term, our goal is to determine a method of prediction of labial coarticulation adaptable to a virtual speaker. Predicting labial coarticulation is a difficult problem that gave rise to many studies and models. To predict the anticipatory coarticulation gestures (see [87] for an overall presentation of labial coarticulation), three main models have been proposed: the look-ahead model, the time-locked model and the hybrid model.

These models were often compared on their performance in the case of the prediction of anticipation protrusion in VCV or VCCV sequences where the first vowel is unrounded, the consonant(s) is neutral with respect to labial articulation and the last vowel is rounded. There is no general agreement about the efficiency of these models. More recent models have been developed. The one of Abry and Lallouache [78] advocates for the theory of expansion movements: the movement tends to be anticipated when no phonological constraint is imposed on labiality. Cohen and Massaro [83] proposed dominance functions that require a substantial numerical training.

Most of these models derive from the observations of a limited number of speakers. We are thus developing a more explicative model, i.e., essentially a phonetically based approach that tries to understand how speakers manage to control labial parameters from the sequence of phonemes to be articulated.

3.2.4. Speech synthesis

Data-driven speech synthesis is widely adopted to develop Text-to-Speech (TTS) synthesis systems. Basically, it consists of concatenating pieces of signal (units) selected from a pre-recorded sentence corpus. Our ongoing work on acoustic TTS was recently extended to study acoustic-visual speech synthesis (bimodal units).

3.2.4.1. Text-to-speech synthesis

Data-driven text-to-speech synthesis is usually composed of three steps to transform a text in speech signal. The first step is Natural Language Processing (NLP) which tags and analyzes the input text to obtain a set of features (phoneme sequence, word grammar categories, syllables...). It ends with a prosodic model which transforms these features into acoustic or symbolic features (F0, intensity, tones...). The second step uses a Viterbi algorithm to select units from a corpus recorded beforehand, which have the closest features to the prosodic features expected. The last step amounts to concatenate these units.

Such systems usually generate a speech signal with a high intelligibility and a naturalness far better than that achieved by old systems. However, building such a system is not an easy task [82] and the global quality mainly relies on the quality of the corpus and prosodic model. The prosodic model generally provides a good standard prosody, but, the generated speech can suffer from a lack of variability. Especially during the synthesis of extended passages, repetition of similar prosodic patterns can lead to a monotonous effect. Therefore, to avoid this problem due to the projection of linguistic features onto symbolic or acoustic dimensions (during NLP), we [84] proposed to perform the unit selection directly from linguistic features without incorporating any prosodic information. To compensate the lack of prosodic prediction, the selection needs to be performed with numerous linguistic features. The selection is no longer restrained by a prosodic model but only driven by weighted features. The consequence is that the quality of synthesis may drop in crucial instants. Our works deal to overcome this new problem while keeping advantage of the lack of prosodic model.

These works have an impact on the construction of corpus and on the NLP engine which needs to provide as much information as possible to the selection step. For instance, we introduced a chunker (shallow parser) to give us information on a potential rhythmic structure. Moreover, to perform the selection, we developed an algorithm to automatically weight the linguistic features given by the NLP. Our method relies on acoustic clustering and entropy information [84]. The originality of our approach leads us to design a more flexible unit selection step, constrained but not restrained.

3.2.4.2. Acoustic-visual speech synthesis

Audiovisual speech synthesis can be achieved using 3D features of the human face supervised by a model of speech articulation and face animation. Coarticulation is approximated by numerical models that describe the synergy of the different articulators. Acoustic signal is usually synthetic or natural speech synchronized with the animation of the face. Some of the audiovisual speech systems are inspired by recent development in speech synthesis based on samples and concatenative techniques. The main idea is to concatenate segments of recorded speech data to produce new segments. Data can be video or motion capture. The main drawback of these methods is that they focus on one field, either acoustic or visual. But (acoustic) speech is actually generated by moving articulators, which modify the speaker's face. Thus, it is natural to find out that acoustic and face movements are correlated. A key point is therefore to guarantee the internal consistency of the

acoustic-visual signal so that the redundancy of these two signals acknowledged as a determining perceptive factor, can really be exploited by listeners. It is thus important to deal with the two signals (acoustic and visual) simultaneously and to keep this link during the whole process. This is why we make the distinction between audiovisual speech synthesis (where acoustic is simply synchronized with animation) and acoustic-visual speech where speech is considered as a bimodal signal (acoustic and visual) as considered in our work. Our long-term goal is to contribute to the fields of acoustic speech synthesis and audiovisual speech synthesis by building a bimodal corpus and developing an acoustic-visual speech synthesis system using bimodal unit concatenation.

3.3. Automatic speech recognition

Automatic speech recognition aims at reproducing the cognitive ability of humans to recognize and understand oral speech. Our team has been working on automatic speech recognition for decades. We began with knowledge-based recognition systems and progressively made our research works evolve towards stochastic approaches, both for acoustic and language models. Regarding acoustic models, we have especially investigated HMM (Hidden Markov Models), STM (Stochastic Trajectory Models), multi-band approach and BN (Bayesian Networks). Regarding language models, our main interest has concerned Ngram approaches (word classes, trigger, impossible Ngram, etc).

The main challenge of automatic speech recognition is its robustness to multiple sources of variability [89]. Among them, we have been focusing on acoustic environment, inter- and intra-speaker variability, different speaking styles (prepared speech, spontaneous, etc.) and non-native pronunciations.

Another specifity of automatic speech recognition is the necessity to combine efficiently all the research works (in acoustic modeling, language modeling, speaker adaptation, etc.) into a core platform in order to evaluate them, and to go beyond pure textual transcriptions by enriching them with punctuation, syntax, etc., in order to make them exploitable by both humans and machines.

3.3.1. Acoustic features and models

The raw acoustic signal needs to be parameterized to extract the speech information it contains and to reduce its dimensionality. Most of our research and recognition technologies make use of the classical Mel Feature Cepstral Coefficients, which have proven since many years to be amongst the most efficient front-end for speech recognition. However, we have also explored alternative parameterizations to support some of our recent research progresses. For example, prosodic features such as intonation curves and vocal energy give important cues to recognize dialog acts, and more generally to compute information that relates to supra-phonemic (linguistic, dialog, ...) characteristics of speech. Prosodic features are developed jointly for both the Speech Analysis and Speech Recognition topics. We also developed a new robust front-end, which is based on wavelet-decomposition of the speech signal.

Concerning acoustic models, stochastic models are now the most popular approach for automatic speech recognition. Our research on speech recognition also largely exploits Hidden Markov Models (HMM). In fact, HMMs are mainly used to model the acoustic units to be recognized (usually triphones) in all of our recognition engines (ESPERE, ANTS...). Besides, we have investigated Bayesian Networks (BN) to explicitly represent random variables and their independence relationships to improve noise robustness.

3.3.2. Robustness and invariance

Part of our research activities about ASR aims at improving the robustness of recognizers to the different sources of variability that affect the speech signal and damage the recognition. Indeed, the issue of the lack of robustness of state-of-the-art ASR systems is certainly the most problematic one that still prevents the wide deployment of speech recognizers nowadays. In the past, we developed a large range of techniques to address this difficult topic, including robust acoustic models (such as stochastic trajectory and multi-band models) and model adaptation techniques (such as missing data theory). The following state-of-the-art approaches thus form our baseline set of technologies: MLLR (Maximum Likelihood Linear Regression), MAP (Maximum A Posteriori), PMC (Parallel Model Combination), CMN (Cepstral Mean Normalization), SAT (Speaker

Adaptive Training), HLDA (Heteroscedastic Linear Discriminant Analysis), Spectral Subtraction and Jacobian Adaptation.

These technologies constitute the foundations of our recent developments in this area, such as non-native speaker adaptation, out-of-vocabulary words detection and adaptation to pronunciation variations. Handling speech variabilities may also benefit from exploiting additional external or contextual sources of information to more tightly guide the speech decoding process. This is typically the role of the language model, which shall in this context be augmented with higher-level knowledge, such as syntactic or semantic cues. Yet, automatically extracting such advanced features is very challenging, especially on imperfect transcribed speech.

The performance of automatic speech recognition systems drastically drops when confronted with non-native speech. If we want to build an ASR system that takes into account non-native speech, we need to modify the system because, usually, ASR systems are trained on standard phone pronunciations and designed to recognize only native speech. In this way, three method categories can be applied: acoustic model transformation, pronunciation modeling and language modeling. Our contribution concerns the first two methods.

3.3.3. Segmentation

Audio indexing and automatic broadcast news transcription need the segmentation of the audio signal. The segmentation task consists in two steps: firstly, homogeneous segments are extracted and classified into speech, noise or music, secondly, speakers turns are detected in the extracted speech segments.

Speech/music segmentation requires to extract discriminant acoustic parameters. Our contribution concerns the MFCC and wavelet parameters. Another point is to find a good classifier. Various classifiers are commonly used: k-Nearest-Neighbors, Hidden Markov Models, Gaussian Mixture Models, Artificial Neural Networks.

As to detect speaker turns, the main approach consists of splitting the audio signal into segments that are assumed to contain only one speaker and then a hierarchical clustering scheme is performed for merging segments belonging to the same speaker.

3.3.4. Speech/text alignment

Speech/text alignment consists in finding time boundaries of words or phones in the audio signal knowing the orthographic transcription. The main applications of speech/text alignment are training of acoustic models, segmentation of audio corpus for building units for speech synthesis or segmentation of the sentence uttered by a learner of a foreign language. Moreover, speech/text alignment is a useful tool for linguistic researchers.

Speech/text alignment requires two steps. The first step generates the potential pronunciations of the sentence dealing with multiple pronunciations of proper nouns, liaisons, phone deletions, and assimilations. For that, the phonetizer is based on a phonetic lexicon, and either phonological rules or an automatic classifier as a decision tree. The second step finds the best pronunciation corresponding to the audio signal using acoustic HMM models and an alignment algorithm. The speech team has been working on this domain for a long time.

3.4. Speech to Speech Translation and Language Modeling

Speech-to-Speech Translation aims at translating a source speech signal into a target speech signal. A sequential way to address this problem is to first translate a text to another one. And after, we can connect a speech recognition system at the input and a text to speech synthesis system at the output. Several ways to address this issue exist. The concept used in our group is to let the computer learning from a parallel text all the associations between source and target units. A unit could be a word or a phrase. In the early 1990s [80] proposes five statistical translation models which became inescapable in our community. The basic idea of the model 1 is to consider that any word of the target language could be a potential translation of any source word. The problem is then to estimate the distribution probability of a target word given a source one. The translation problem is similar to the speech recognition one. Indeed, we have to seek the best foreign sentence given a source one. This one is obtained by decoding a lattice translation in which a language and translation models are used. Several issues have to be supported in machine translation as described below.

3.4.1. Word translation

The first translation systems identify one-to-one associations between words of target and source languages. This is still necessary in the present machine translation systems. In our group we develop a new concept to learn the translation table. This approach is based on computing all the inter-lingual triggers inside a parallel corpus. This leads to a pertinent translation table [95]. Obviously, this is not sufficient in order to make a realistic translation because, with this approach, one word is always translated into one word. In fact, it is possible to express the same idea in two languages by using different numbers of words. Thus, a more general one-to-one alignment has to be achieved.

3.4.2. Phrase translation

The human translation is a very complex process which is not only word-based. A number of research groups developed phrase-based systems which are different from the baseline IBM's model in training. These methods deal with linguistic units which consists in more than one word. The model supporting phrase-based machine translation uses reordering concept and additional feature functions. In order to retrieve phrases, several approaches have been proposed in the litterature. Most of them require word-based alignments. For example, Och and al. [96] collected all phrase pairs that were consistent with the word alignment provided by Brown's models.
We developed a phrase based algorithm which is based on finding first an adequate list of phrases. Then, we find out the best corresponding translations by using our concept of inter-lingual triggers. A list of the best translations of sequences is then selected by using simulated annealing algorithm.

3.4.3. Language model

A language model has an important role in a statistical machine translation. It ensures that the translated words constitute a valid linguistic sentence. Most of the community uses n-grams models, that is what we do also.

3.4.4. Decoding

The translation issue is treated as an optimization problem. Translating a sentence from English into a foreign language involves finding the best Foreign target sentence f^* which maximizes the probability of f given the English source sentence e. The Bayes rule allows to formulate the probability $P(f|e)$ as follows:

$$f^* = \arg\max_f P(f|e) = \arg\max_f P(e|f)P(f)$$

The international community uses either PHARAOH [93] or MOSES [92] based on a beam search algorithm. In our group we started decoding by PHARAOH but we moved recently to MOSES.

4. Application Domains

4.1. Application Domains

Our research is applied in a variety of fields from ASR to paramedical domains. Speech analysis methods will contribute to the development of new technologies for language learning (for hearing-impaired persons and for the teaching of foreign languages) as well as for hearing aids. In the past, we developed a set of teaching tools based on speech analysis and recognition algorithms of the group (cf. the ISAEUS [88] project of the EU that ended in 2000). We are continuing this effort towards the diffusion of a course on Internet.

Speech is likely to play an increasing role in man-machine communication. Actually, speech is a natural mean of communication, particularly for non-specialist persons. In a multimodal environment, the association of speech and designation gestures on touch screens can, for instance, simplify the interpretation of spatial reference expressions. Besides, the use of speech is mandatory in many situations where a keyboard is not available: mobile and on-board applications (for instance in the framework of the HIWIRE European project for the use of speech recognition in a cockpit plane), interactive vocal servers, telephone and domestic applications, etc. Most of these applications will necessitate to integrate the type of speech understanding process that our group is presently studying. Furthermore, speech to speech translation concerns all multilingual applications (vocal services, audio indexing of international documents). The automatic indexing of audio and video documents is a very active field that will have an increasing importance in our group in the forthcoming years, with applications such as economic intelligence, keyword spotting and automatic categorization of mails.

5. Software and Platforms

5.1. WinSnoori

WinSnoori is a speech analysis software that we have been developing for 15 years. It is intended to facilitate the work of the scientist in automatic speech recognition, phonetics or speech signal processing. Basic functions of WinSnoori enable several types of spectrograms to be calculated and the fine edition of speech signals (cut, paste, and a number of filters) as the spectrogram allows the acoustical consequences of all the modifications to be evaluated. Beside this set of basic functions, there are various functionalities to annotate phonetically or orthographically speech files, to extract fundamental frequency, to pilot the Klatt synthesizer and to utilize PSOLA resynthesis.

The current version of WinSnoori is available on http://www.winsnoori.fr.

5.2. JSnoori

JSnoori is written in Java and uses signal processing algorithms developed within WinSnoori software with the double objective of being a platform independent signal visualization and manipulation tool, and also for designing exercises for learning the prosody of a foreign language. JSnoori thus focused 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). Since phonetic segmentations and annotations play a central role in the derivation of diagnosis concerning the realization of prosody by learners, several tools have been incorporated to segment and annotate speech. In particular, a complete phonetic keyboard is available, several kinds of annotation can be used (phonemes, syllables and words) and forced alignment can exploit variants to cope with non native accents. In addition, JSnoori offers real time F0 calculation which can be useful from a pedagogical point of view.

5.3. Xarticulators

Xarticulators software is intended to delineate contours of speech articulators in X-ray images, to construct articulatory models and to synthesize speech from X-ray films. This software provide 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 together with the sound.

It also enables the construction of adaptable linear articulatory models from the X-ray images.

This year we particularly worked on the possibility of synthesizing speech from X-ray images. We thus substantially improved algorithms used to compute the centerline of the vocal tract in order to segment the vocal tract into elementary tubes approximating the propagation of a one-dimensional wave. We also developed time patterns used to synthesize sequences of voiceless consonants and vowels (VCV). In addition we also added the possibility of processing digitized manual delineation results made on sheet of papers in the seventies.

5.4. SUBWEB

We published in 2007 a method which allows to align sub-titles comparable corpora [94]. In 2009, we proposed an alignment web tool based on the developed algorithm. It allows to: upload a source and a target files, obtain an alignment at a sub-title level with a verbose option, and a graphical representation of the course of the algorithm. This work has been supported by CPER/TALC/SUBWEB [2].

5.5. ANTS

The aim of the Automatic News Transcription System (ANTS) is to transcribe radio or TV shows. ANTS is composed of several stages. The first processing steps aim at splitting the audio stream into homogeneous segments of a manageable size and at identifying the segment characteristics in order to allow the use of specific algorithms or models according to the nature of the segment. This includes broad-band/narrow-band speech segmentation, speech/music classification, speaker segmentation and clustering, detection of silences/breathing segments and generally speaker gender classification.

Each segment is then decoded using a large vocabulary continuous speech recognition engine, either the Julius engine or the Sphinx engine. The Julius engine operates in two passes: in the first pass, a frame-synchronous beam search algorithm is applied on a tree-structured lexicon assigned with bigram language model probabilities. The output of this pass is a word-lattice. In the second pass, a stack decoding algorithm using a trigram language model gives the N-best recognition sentences. The Sphinx engine processes the speech input segment in a single forward pass using a trigram language model.

Further processing passes are usually run in order to apply unsupervised adaptation processes on the feature computations (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 both systems can be efficiently combined for improved decoding performance.

The latest version which relies on a perl script exploits the multiple CPUs available on a computer to reduce the processing time, and runs on both a stand alone linux machine and on the cluster.

5.6. CoALT

CoALT (Comparing Automatic Labeling Tools) compares two automatic labelers or two speech-text alignment tools, ranks them and displays statistics about their differences. The main feature of our software is that a user can define its own criteria for evaluating and comparing two speech- text alignment tools. With CoALT, a user can give more importance to either phoneme labels or phoneme boundaries because the CoALT elastic comparison algorithm takes into account time boundaries. Moreover, by providing a set of phonetic rules, a user can define the allowed discrepancies between the automatic labeling result and the hand-labeling one.

5.7. TTS SoJA

TTS SoJA (Speech synthesis platform in Java) is a software for text-to-speech synthesis. The aim of this software is to provide a toolkit to test some steps of natural language processing and to provide a whole system of TTS based on non uniform unit selection algorithm. The software performs all steps from the text to the speech signal. Moreover, it provides a set of tools to elaborate 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 developed in Java. Some modules are in C. The platform is designed to make easy the addition of new modules. The software runs under Windows and Linux (tested on Mandriva, Ubuntu). It can be launch with a graphical user interface or directly integrated in a Java code or by following the client-server paradigm.

[2] http://wikitalc.loria.fr/dokuwiki/doku.php?id=operations:subweb

The software license should easily allow associations of impaired people to use the software. A demo web site has been built: http://soja-tts.loria.fr

5.8. JCorpusRecorder

JCorpusRecorder is a software for the recording of audio corpora. It provides a 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 with textual information given in input (nationality, speaker language, gender...). An easy syntactic tagging allows 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 now developed in Java (since 2013). It is currently used for the recording of sentences in several projects (including IFCASL).

5.9. VisArtico

VisArtico is intended to visualize articulatory data acquired using an articulograph [97]. It is intended for researchers that need to visualize data acquired from the articulograph with no excessive processing. It is well adapted to the data acquired using the AG500 and AG501 (developed by Carstens Medizinelektronik GmbH), and the articulograph NDI Wave, developed by Northern Digital Inc.

The software allows displaying the positions of the sensors that are simultaneously animated with the speech signal. It is possible to display the tongue contour and the lips contour. The software helps to find the midsagittal plane of the speaker and find the palate contour. In addition, VisArtico allows labeling phonetically the articulatory data.

All this information is very useful to researchers working in the field of speech production, as phoneticians for instance. VisArtico provides several possible views: (1) temporal view, (2) 3D spatial view and (3) 2D midsagittal view. In the temporal view, it is possible to display different articulatory trajectories in addition to the acoustic signal and eventually labels. The midsagittal view can display the tongue contour, the jaw, the lips and the palate.

VisArtico provides several tools to help to improve the quality of interpreting the data. It is a cross-platform software as it is developed in JAVA and does not need any additional external library or framework. It was tested and worked on Windows, Mac OS, and Linux. It should work on any system having JAVA installed. VisArtico is freely distributed via a dedicated website http://visartico.loria.fr.

5.10. FASST

The Flexible Audio Source Separation Toolbox (FASST) is a toolbox for audio source separation (http://bass-db.gforge.inria.fr/fasst/). It aims to become the reference software for research and applications of audio source separation. 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. Besides, it forms the basis of most of our current research in audio source separation, some of which may be incorporated into future versions of the software.

6. New Results

6.1. Speech analysis and synthesis

Participants: Anne Bonneau, Vincent Colotte, Dominique Fohr, Yves Laprie, Joseph Di Martino, Slim Ouni, Agnès Piquard-Kipffer, Emmanuel Vincent, Utpala Musti.

Signal processing, phonetics, health, perception, articulatory models, speech production, learning language, hearing help, speech analysis, acoustic cues, speech synthesis

6.1.1. Acoustic-to-articulatory inversion

The acoustic-to-articulatory inversion from cepstral data has been evaluated on the X-ray database, i.e. X-ray films recorded with the original speech signal. A codebook is used to represent the forward articulatory to acoustic mapping and we designed a loose matching algorithm using spectral peaks to access it. This algorithm, based on dynamic programming, allows some peaks in either synthetic spectra (stored in the codebook) or natural spectra (to be inverted) to be omitted. Quadratic programming is used to improve the acoustic proximity near each good candidate found during codebook exploration. The inversion [40], [10] has been tested on speech signals corresponding to the X-ray films. It achieves a very good geometric precision of 1.5 mm over the whole tongue shape unlike similar works which limit the error evaluation at 3 or 4 points corresponding to sensors located at the front of the tongue.

6.1.1.1. Construction of articulatory models

Articulatory models are intended to approximate the vocal tract geometry with a small number of parameters controlling linear deformation modes. Most of the models have been designed on images of vowels and thus offer a good coverage for vowels but are unable to provide a good approximation for consonants, especially in the region of the constriction. The first problem is related to the nature of contours used to derive linear components. When dealing with vowels there is no contact between the tongue and other fixed articulators (palate, teeth). Factor analysis used to determine linear modes of deformation of the tongue only takes into account the influence of the tongue muscles. This is no longer the case with consonants, since a contact is realized between the tongue and the palate, alveolar ridge or teeth for stops /k, g, t, d/ and the sonorant /l/ in French. The deformation factors thus incorporate the “clipping” effect of the palate. Following the idea of using virtual articulatory targets that lie beyond the positions that can be reached, here the palate, we edited delineated tongue contours presenting a contact with the palate. We chose a conservative solution which consists of keeping the tongue contour up to the contact point and extending it while guaranteeing a “natural shape”. These new contours do not cross the palate for more than 10 mm. As such, this modification alone is not sufficient, because the number of images corresponding to consonants is small even if the corpus used in this work is phonetically balanced. We thus duplicated a number of consonant X-ray images in order to increase the weight of deformation factors corresponding to the tongue tip which is essential for some consonants, /l/ for instance. This approach provides a very good fitting with original tongue contours, i.e. 0.83 mm in average with 6 components over the whole tongue contour and only 0.56 mm in the region of the main place of articulation, which is important with a view of synthesizing speech.

6.1.1.2. Articulatory copy synthesis

Acoustic features and articulatory gestures have always been studied separately. Articulatory synthesis could offer a nice solution to study both domains simultaneously provided that relevant information can be fed into the acoustic simulation. The first step consisted of connecting the 2D geometry given by mediosagittal images of the vocal tract with the acoustic simulation. Last year we thus developed an algorithm to compute the centerline of the vocal tract, i.e. a line which is approximately perpendicular to the wave front. The centerline is then used to segment the vocal tract into elementary tubes whose acoustic equivalents are fed into the acoustic simulation. A new version of the centerline algorithm [53] has been developed in order to approximate the propagation of a plane wave more correctly.

The work on the development of time patterns used to pilot the acoustic simulation has been continued by improving the choice of relevant X-ray images and the temporal transitions from one image to the following. This procedure has been applied successfully to copy sentences and VCV for four X-ray films of the DOCVACIM database[52]. More difficult transitions, i.e. those corresponding to consonant clusters, will be investigated this year.

In addition to the control of the acoustic simulation we started an informal cooperation with the IADI laboratory www.iadi-nancy.fr in order to record better static images of the vocal tract, and cineMRI, i.e. films, for a number of sentences.

6.1.2. Using articulography for speech animation

We are continuously working on the acquisition and analysis of the articulatory data using electromagnetic articulography (EMA).This year, we have conducted research to use EMA as motion capture data and we showed that it is possible to use it for audiovisual speech animation. In fact, as EMA captures the position and orientation of a number of markers, attached to the articulators, during speech, it performs the same function for speech that conventional motion capture does for full-body movements acquired with optical modalities, a long-time staple technique of the animation industry. We have processed EMA data from a motion-capture perspective and applied to the visualization of an existing multimodal corpus of articulatory data, creating a kinematic 3D model of the tongue and teeth by adapting a conventional motion capture based animation paradigm. Such an animated model can then be easily integrated into multimedia applications as a digital asset, allowing the analysis of speech production in an intuitive and accessible manner. In this work [61], we have addressed the processing of the EMA data, its co-registration with 3D data from vocal tract magnetic resonance imaging (MRI) and dental scans, and the modeling workflow. We will continue our effort in the future to improve this technique.

6.1.3. Acoustic analyses of non-native speech

Within the framework of the project IFCASL, we designed a corpus for the study of French and German, with both languages pronounced by French and German speakers, so as to put into light L1/L2 interferences. The corpus was constructed to control for several segmental and suprasegmental phenomena. German and French, for instance, show different kinds of voicing patterns. Whereas in French, the voicing opposition of stops is realized as voiced versus unvoiced, in German, the same difference is realized mostly as unaspirated versus aspirated. Furthermore, differences between the two language groups are expected with respect to the production of nasal vowels (absent in German), the realization of /h/ (not present in French, but in German). On the suprasegmental level, word stress and focus intonation are central to our investigation. Speakers produce both native and non-native speech, which allows for a parallel investigation of both languages.

We have conducted a pilot study on the realization of obstruents in word-final position -a typical example of L1-L2 interference on the segmental level-, which are subject to devoicing in German, but not in French. First results showed that German learners (beginners) had difficulties to voice French obstruents in this context, and, when listening to French realizations, tend to add a final schwa to achieve the expected realization.

6.1.4. Speech synthesis

We recall that within the framework of the ViSAC project we have developed bimodal acoustic-visual synthesis technique that concurrently generates the acoustic speech signal and a 3D animation of the speaker's outer face. This is done by concatenating bimodal diphone units that consist of both acoustic and visual information. In the visual domain, we mainly focus on the dynamics of the face rather than on rendering. The proposed technique overcomes the problems of asynchrony and incoherence inherent in classic approaches to audiovisual synthesis. The different synthesis steps are similar to typical concatenative speech synthesis but are generalized to the acoustic-visual domain. This year we have performed an extensive evaluation of the synthesis system using perceptual and subjective evaluations. The overall outcome of the evaluation indicates that the proposed bimodal acoustic-visual synthesis technique provides intelligible speech in both acoustic and visual channels [22]. For testing purposes we have also added a simple tongue model that is controlled by the generated phonemes. The purpose is to improve the quality of the audiovisual speech intelligibility.

Morover, we perform feature selection and weight tuning for a given unit-selection corpus to make the ranking given by the target cost function consistent with the ordering given by an objective dissimilarity measure. To find an objective metric highly correlated to perception we analyzed correlation between objective and subjective evaluation results. It shows interesting patterns which might help in designing better tuning metrics and objective evaluation techniques [55].

6.1.5. Phonemic discrimination evaluation in language acquisition and in dyslexia and dysphasia

We keep working on a project concerning identification of early predictors of reading, reading acquisition and language difficulties, more precisely in the field of specific developmental disabilities : dyslexia and dysphasia. A fair proportion of those children show a weakness in phonological skills, particularly in phonemic discrimination. However, the precise nature and the origin of the phonological deficits remain unspecified. In the field of dyslexia and normal acquisition of reading, our first goal was to contribute to identify early indicators of the future reading level of children. We based our work on the longitudinal study - with 85 French children - of [90], [91] which indicates that phonemic discrimination at the beginning of kindergarten is strongly linked to success and specific failure in reading acquisition. We study now the link between oral discrimination both with oral comprehension and written comprehension. Our analyses are based on the follow up of a hundred children for 4 years from kindergarten to end of grade 2 (from age 4 to age 8) [98].

6.1.6. Enhancement of esophageal voice

6.1.6.1. Pitch detection

Over the last two years, we have proposed two new real time pitch detection algorithms (PDAs) based on the circular autocorrelation of the glottal excitation, weighted by temporal functions, derived from the CATE [85] original algorithm (Circular Autocorrelation of the Temporal Excitation), proposed initially by J. Di Martino and Y. Laprie. In fact, this latter algorithm is not constructively real time because it uses a post-processing technique for the Voiced/Unvoiced (V/UV) decision. The first algorithm we developed is the eCATE algorithm (enhanced CATE) that uses a simple V/UV decision less robust than the one proposed later in the eCATE+ algorithm. We propose a recent modified version called the eCATE++ algorithm which focuses especially on the detection of the F0, the tracking of the pitch and the voicing decision in real time. The objective of the eCATE++ algorithm consists in providing low classification errors in order to obtain a perfect alignment with the pitch contours extracted from the Bagshaw or Keele databases by using robust voicing decision techniques. This algorithm has been published in Signal, Image and Video Processing, [14].

6.1.6.2. Real-time pitch detection for application to pathological voices

The work first rested on the CATE algorithm developed by Joseph Di Martino and Yves Laprie, in Nancy, 1999.The CATE (Circular Autocorrelation of the Temporal Excitation) algorithm is based on the computation of the autocorrelation of the temporal excitation signal which is extracted from the speech log-spectrum. We tested the performance of the parameters using Bagshaw database, which is constituted of fifty sentences, pronounced by a male and a female speaker. The reference signal is recorded simultaneously with a microphone and a laryngograph in an acoustically isolated room. These data are used for the calculation of the contour of the pitch reference. When the new optimal parameters from the CATE algorithm were calculated, we carried out statistical tests with the C functions provided by Paul BAGSHAW. The results obtained were very satisfactory and a first publication relative to this work was accepted and presented at the ISIVC 2010 conference [79]. At the same time, we improved the voiced / unvoiced decision by using a clever majority vote algorithm electing the actual F0 index candidate. Recently Fadoua Bahja developed a new algorithm based on wavelet transforms applied to the cepstrum excitation. The preliminary results obtained were satisfactory and a complete description of this latter study is under a submission process in an international journal.

6.1.6.3. Voice conversion techniques applied to pathological voice repair

Voice conversion is a technique that modifies a source speaker's speech to be perceived as if a target speaker had spoken it. One of the most commonly used techniques is the conversion by GMM (Gaussian Mixture Model). This model, proposed by Stylianou, allows for efficient statistical modeling of the acoustic space of a speaker. Let "x" be a sequence of vectors characterizing a spectral sentence pronounced by the source speaker and "y" be a sequence of vectors describing the same sentence pronounced by the target speaker. The goal is to estimate a function F that can transform each source vector as nearest as possible of the corresponding target vector. In the literature, two methods using GMM models have been developed: In the first method (stylianou,98), the GMM parameters are determined by minimizing a mean squared distance between the transformed vectors and target vectors. In the second method (kain,98), source and target vectors are combined in a single vector "z". Then, the joint distribution parameters of source and target speakers is estimated using the EM optimization technique. Contrary to these two well known techniques, the transform function F, in our

laboratory, is statistically computed directly from the data: no needs of EM or LSM techniques are necessary. On the other hand, F is refined by an iterative process. The consequence of this strategy is that the estimation of F is robust and is obtained in a reasonable lapse of time. Recently, we realized that one of the most important problems in speech conversion is the prediction of the excitation. In order to solve this problem we developed a new strategy based on the prediction of the cepstrum excitation pulses. Another very important problem in voice conversion concerns the prediction of the phase spectra. This study is under progress in the framework of an Inria ADT which began in September 2013.

6.1.6.4. Signal reconstruction from short-time Fourier transform magnitude spectra

Joseph Di Martino and Laurent Pierron developed in 2010 an algorithm for real-time signal reconstruction from short-time Fourier magnitude spectra [86]. Such an algorithm has been designed in order to enable voice conversion techniques we are developing in Nancy for pathological voice repair. Recently Mouhcine Chami, an assistant-professor of the INPT institute at Rabat (Morocco) proposed a hardware implementation of this algorithm using FPGAs. This implementation has been published in the SIIE 2012 conference [81]. Maryem Immassi, a PhD student of Mouhcine Chami, is comparing this algorithm with the state of the art RTISI-LA algorithm in the framework of a hardware implementation.

6.1.7. Audio source separation

Audio source separation is the task of extracting one or more target source signals from a given mixture signal. It 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. We studied the impact of sparsity penalties over the mixing filters [38] and we defined probabilistic priors [20] and deterministic subspace constraints [45] over the spatial covariance matrices. We also wrote a review paper about guided audio source separation for *IEEE Signal Processing Magazine* [28].

This paper highlighted that many guided separation techniques now exist that are closer than ever to successful industrial applications, as exemplified by the ongoing industrial collaborations of the team. In order to exploit our know-how for these real-world applications, we investigated issues such as the impact of audio coding [59], artifact reduction [21], real-time implementation [62], and latency [70]. Two patents have been filed [77], [76]. We also started a new research track on the fusion of multiple source separation techniques [46].

Finally, we pursued our long-lasting efforts on the evaluation of audio source separation by collecting the first-ever publicly available dataset of multichannel real-world noise recordings [71] and by conducting an experimental comparison of the two main families of techniques used for source separation [63].

6.2. Automatic speech recognition

Participants: Dominique Fohr, Jean-Paul Haton, Irina Illina, Denis Jouvet, Odile Mella, Emmanuel Vincent, Arseniy Gorin, Luiza Orosanu, Dung Tran.

stochastic models, acoustic models, language models, automatic speech recognition, speech transcription, training, robustness

6.2.1. 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.

The use of larger speech training corpora allows us increasing the size of the acoustic models (more parameters through more Gaussians components per density, and more shared densities) and this leads to improved performance. However, in such approaches, Gaussian components are estimated independently for each density. Thus, after having investigated last year the usage of multiple modeling approaches for better constraining the acoustic decoding space, recent studies have focused on enriching the acoustic models themselves in view of handling trajectory and speaker consistency in decoding.

This year a new modeling approach was developed that takes benefit 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, some Gaussian components are estimated on the data of each class. 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. The method allows us to better parameterize GMM-HMM without increasing significantly the number of model parameters. The experiments on French radio broadcast news data demonstrate the improvement of the accuracy with such parameterization compared to the models with similar, or even larger number of parameters [43].

Current experiments deal with stranded HMM. The objective of such an approach is to introduce in the GMM-HMM modeling some extra parameters to take into account the transition between the Gaussian components when moving from one frame to the next.

6.2.2. Noise-robust speech recognition

In many real-world conditions, the speech signal is overlapped with noise, including environmental sounds, music, or undesired extra speech. Source separation may then be used as a pre-processing stage to enhance the desired speech signal [64]. In practice, the enhanced signal always includes some distortions compared to the original clean signal. It is important to quantify which parts of the enhanced signal are reliable in order not to propagate these distortions to the subsequent feature extraction and decoding stages. A number of heuristic statistical uncertainty estimators and propagators have been proposed to this aim. We started some work aiming to improve the accuracy of these estimators and propagators. We also showed how to exploit uncertainty in order to train unbiased acoustic models directly from noisy data [24].

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. This challenge aims to recognize small or medium-vocabulary speech mixed with noise recorded in a real family home over the course of several weeks. We analyzed the outcomes of the first edition [16] which led to a special issue of *Computer Speech and Language* [15] and we organized a second edition in 2013 [66] which illustrated the progress made in two years over small-vocabulary speech and the remaining challenges towards robust recognition of medium-vocabulary speech [65].

6.2.3. Linguistic modeling

Usually the lexicon used by a speech recognition system refers to word entries, where each entry in the pronunciation lexicon specifies a possible pronunciation of a word, and the associated language model specifies the probability of a word knowing preceding words. However, whatever the size of the lexicon is, the size is always finite, and the speech recognition system cannot recognize properly words that are not present in the lexicon. In such cases, the unknown word is typically replaced by a sequence of short words which is acoustically similar to the unknown speech portion.

6.2.3.1. Random indexing

This year we studied the introduction of semantic information through the Random Indexing paradigm (RI) in statistical language models used in speech recognition. Random Indexing is a scalable alternative to LSA (Latent Semantic Analysis) for analyzing relationships between a set of documents and the terms they contain. We determined the best methods and parameters by minimizing the perplexity of a realistic corpus of 290000 words. We investigated 4 methods for training RI matrices, 4 weighting functions, several matrix sizes and how balancing the 4-gram and RI language model. We only obtained a relative gain of 3% [42].

6.2.3.2. Continuous language models

Language modeling plays an important role in automatic speech recognition because it constrains the decoder to search the most likely sequences of words according to a given language and a given task. A limitation of N-grams models is that they represent the words in a discrete space. It would be interesting to represent words in a continuous space where semantically close words would be projected in the same region of space. This projection can be achieved by recurrent neural networks. Moreover they are able to learn long-term

dependencies with the recurrent layer that can store a record of the past. During his master internship, Othman Zennaki integrated this new language model in our speech recognition system ANTS.

6.2.3.3. Linguistic units for embedded systems

In the framework of the RAPSODIE project, speech recognition is to be used to help communication with hard of hearing people. Because of requirements on memory and CPU (almost real time processing), various modeling approaches have been investigated with respect to linguistic units. The first approach has focused on analyzing the achieved phonetic decoding performance of various linguistic units (phonemes, syllables, words). Best phonetic decoding performance is achieved using word units and associated tri-gram language model, but at the expense of large CPU and memory requirements. Using directly phoneme units leads to the smallest models and requires little CPU, however, this also leads to the worst performance. The proposed approach relying on syllable units provides results which are rather close to the word based approach, but requires much less CPU [58], [57].

Further experiments are now focusing on combining word and syllable units, in view of having frequent words covered by the word units, and using syllables for decoding unknown words.

6.2.3.4. OOV proper name retrieval

Proper name recognition 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.

In the framework of the ContNomina project, we focus on increasing the vocabulary coverage of a speech transcription system by automatically retrieving proper names from contemporary diachronic text documents. We proposed methods that dynamically augment the automatic speech recognition system vocabulary, using lexical and temporal features in diachronic documents. We also studied different metrics for proper name selection in order to limit the vocabulary augmentation and therefore the impact on the ASR performances. Recognition results show a significant reduction of the word error rate using augmented vocabulary [56].

6.2.4. Speech transcription

The first complete version of the speech transcription system ANTS (see section 5.5) has been initially developed in the framework of the Technolangue project ESTER, and since then, the system has been regularly enriched through the integration of research results. The latest version can handle either HTK-based acoustic models through the Julius decoder, or Sphinx-based acoustic models with the CMU Sphinx decoders. In the last version, a Perl script encapsulates all the calls to the various tools used for diarization, model adaptation and speech recognition, and takes benefit of the multiple CPU available on the computer for parallelizing the different tasks as much as possible.

6.2.4.1. Combining recognizers

Last year in the context of the ETAPE speech transcription evaluation campaign, the Sphinx-based and Julius-based decoders have been further improved, and it was observed that combining the recognition outputs of several Sphinx-based and Julius-based decoder lead to a significant word error rate reduction compared to the best individual system.

More controlled experiments have then been performed to understand what was the main reason of the large performance improvement observed when combining Julius-based and Sphinx-based transcription system results. The Sphinx decoder processes the speech data in a forward pass, whereas the Julius decoder ends its decoding process by a backward pass. The Sphinx training and decoding scripts have been modified to process the speech material in a reverse time order; and various systems were developed by using different sets of acoustic features and different sets of acoustic units. It was then observed that combining several Sphinx-forward and several Sphinx-reverse decoders lead to much better results than combining the same amount of only Sphinx-forward decoders or only Sphinx-reverse decoders; and the achieved word error rate was consistent with the one obtained by combining the Sphinx-based (forward) and Julius-based (backward) decoders [49]. Hence, the improvement is mainly due to the fact that forward-based and backward-based processing are combined. Because heuristics are applied during decoding to limit the acoustic space that is explored, some hypotheses might be wrongly pruned when processing the data one way, and may be kept in

the active beam search when processing the other way. This is corroborated by the analysis of the word graph which show a large dissimilarity in the distribution of the number of words starting and ending in each frame [48].

Experiments have also shown that when the forward and backward decoders yield the same word hypothesis, this word is likely to be a correct answer. Recent experiments are investigating how far such behavior could help for unsupervised learning of acoustic models.

6.2.4.2. Spontaneous speech

During his master intership, Bruno Andriamiarina focuses on the new challenges brought by this spontaneity of the speech, making it difficult to be transcribed by the existing automatic speech recognition systems. He studied how to improve global performance of automatic speech recognition systems when dealing with spontaneous speech by adapting language model and pronunciation dictionary to this particular type of speech. He also studied the detection of disfluent speech portions (produced by spontaneous speech) in speech signal using a Gaussian Mixture Model (GMM)-based classifier trained on prosodic features covering the main prosodic characteristics (duration, fundamental frequency and energy).

6.2.4.3. Towards a structured output

The automatic detection of the prosodic structure of speech utterances has been investigated. The algorithm relies on a hierarchical representation of the prosodic organization of the speech utterances, and detects prosodic boundaries whether they are followed or not by pause. The detection of the prosodic boundaries and of the prosodic structures is based on an approach that integrates little linguistic knowledge and mainly uses the amplitude of the F0 slopes and the inversion of the F0 slopes as well as phone durations. The approach was applied on a corpus of radio French broadcast news and also on radio and TV shows which are more spontaneous speech data. The automatic prosodic segmentation results were then compared to a manual prosodic segmentation made by an expert phonetician [37].

Further work has focused on analyzing the links between manually set punctuation marks and this automatically detected prosodic structure, in view of using the prosodic structure for helping an automatic punctuation process.

6.2.5. Speech/text alignment

6.2.5.1. Alignment with non-native speech

Non-native speech alignment with text is one critical step in computer assisted foreign language learning. 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 - too short or too long -). 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.

In the case of French speakers learning English, we conducted a detailed analysis that has showed the benefit of taking into account non-native variants, and lead to determining the classes of phonemes whose temporal boundaries are the most accurate and which should be favored in the design of exercises for language learning[18].

In the framework of the IFCASL project, we proposed to use a two-step approach for automatic phone segmentation. The first step consists in determining the phone sequence that best explains the learner's utterance. This is achieved by force aligning the learner's speech utterance with a model representing the various possible pronunciation variants of the current sentence (both native and non-native variants need to be considered). In this step detailed acoustic Hidden Markov Models (HMMs) are used, with a rather large number of Gaussian components per mixture density. This kind of detailed acoustic models is the one that provides the best performance in automatic speech recognition. The second step consists in determining the phone boundaries. This is also achieved through a forced alignment process, but this time, the sequence of phones is known (as determined in the first step), and phone acoustic models with only a few Gaussians components per mixture density are used because it has been shown that they provide better temporal precision

than detailed acoustic models. For the training of the models used for both forced alignment steps, the speech of native and non-native speakers could be used, either directly or by MLLR (Maximum Likelihood Linear Regression) adaptation.

6.2.5.2. Alignment with spontaneous speech

In the framework of the ANR ORFEO, 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. As regards overlapping speech, the orthographic transcription of the audio files only provides a rather imprecise time information of the overlapping speech segment. As a first approach, 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 with the same time boundaries.

6.3. Machine translation and language modeling

Participants: Kamel Smaïli, David Langlois, Denis Jouvet, Emmanuel Vincent, Motaz Saad, Cyrine Nasri.

machine translation, statistical models

6.3.1. Language modeling

6.3.1.1. Vocabulary selection

In the framework of the ETAPE evaluation campaign a new machine learning based process was developed to select the most relevant lexicon to be used for the transcription of the speech data (radio and TV shows). The approach relies on a neural network trained to distinguish between words that are relevant for the task and those that are not. After training, the neural network (NN) is applied to each possible word (text tokens extracted from a very large text corpus). Then the words that have the largest NN output score are selected for creating the speech recognition lexicon. Such an approach can handle counts of occurrences of the words in various data subsets, as well as other complementary information, and thus offer more perspectives than the traditional unigram-based selection procedures [50].

6.3.1.2. 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 log-linear interpolation of multiple conditional distributions. We applied it to the joint modeling of "horizontal" (sequential) and "vertical" (simultaneous) dependencies between notes for polyphonic pitch estimation [26] and to the joint modeling of melody, key and chords for automatic melody harmonization [25]. 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 [74].

6.3.2. Quality estimation of machine translation

In the scope of Confidence Measures, we participated to the World Machine Translation evaluation campaign for the second year (WMT2013 http://www.statmt.org/wmt13/quality-estimation-task.html). More precisely, we proposed a Quality Estimation system to the Quality Estimation shared task. The goal was to predict the quality of translations generated by an automatic system. Each translated sentence is given a score between 0 and 1. The score is obtained by using several numerical or boolean features calculated according to the source and target sentences. We performed a linear regression of the feature space against scores in the range [0 ;1], to this end, we use a Support Vector Machine with 66 features. In this new participation, we proposed to increase the size of the training corpus. For that, we decided to use the post-edited and reference corpora in the training step after assigning a score to each sentence of these corpora. Then, we tune these scores on a development corpus. This leads to an improvement of 10.5% on the development corpus, in terms of Mean Average Error (average difference between reference and predicted scores), but achieves only a slight improvement on the test corpus. This work has been published in [51].

6.3.3. Comparable corpora and multilingual sentiment analysis

In the PhD Thesis of Motaz Saad, we work on collecting comparable corpora. For that purpose we presented a method which extracts and aligns comparable corpora at the article level from Wikipedia encyclopedia based on interlanguage links. To evaluate the closeness of corpora we proposed several comparability measures. Our evaluations show that the proposed comparability measures are able to capture the comparability degree of any comparable corpora [60]. We go further on the comparability of multilingual corpora by studying their comparability in terms of sentiment. The final objective is to propose a multilingual press review concerning a given topic. This review should use several multilingual resources (electronic newspapers), and should class resources according to the including sentiments (fear, joy...about the subject), polarity (against or not to the subject)...This conducts to study opinions across different languages by comparing the underlying messages written by different people having different opinions. We propose "Sentiment based Comparability Measures" to compare opinions in multilingual comparable articles without translating source/target into the same language [27].

6.3.4. Machine translation of arabic dialect

The translation of Arabic dialect constitutes a real challenge since it is an under-resourced language. In fact, Modern Standard Arabic is as any other evoluated language, it means it could be processed by the available tools but unfortunately in Arabic countries people speak an Arabic language which is inspired from the standard one but is different. Our objective is then to propose a speech to speech system converting modern standard Arabic to Algerian dialect. After collecting corpus, we decided to propose a method allowing to diacritize dialects in order to be able in the following to develop an acoustic model. For that, we considered the issue of diacritization as a machine translation issue, and we have developed a statistical machine translation which learns to transform an undiacritized corpus into a diacritized one [44].

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

Our policy in terms of technological and industrial partnership consists in favoring contracts that quite precisely fit our scientific objectives.

A three-day consulting contract was conducted with Technicolor (Rennes) in December 2013.

E. Vincent is involved through his former team (PANAMA) in an 18-month bilateral research contract with Canon Research Centre France (Rennes) which ended in July 2013 and in a 30-month bilateral research contract with the SME Studio MAIA (Boulogne-Billancourt).

8. Partnerships and Cooperations

8.1. National initiatives

8.1.1. Equipex ORTOLANG

Project acronym: ORTOLANG [3]

Project title: Open Resources and TOols for LANGuage

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

Coordinator: ATILF (Nancy)

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

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

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 our language bringing us up to the best international level;
- facilitate the use and transfer of resources and tools set up within public laboratories towards industrial partners, in particular towards SME which cannot often 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, such as text-speech alignment, speech visualization, syntactic parsing and annotation, ...

8.1.2. ANR ARTIS

Project acronym: ARTIS

Project title: Inversion articulatoire de la parole audiovisuelle pour la parole augmentée

Duration: January 2009 - June 2013

Coordinator: Yves Laprie (LORIA)

Other partners: Gipsa-Lab, LTCI, IRIT

Abstract: The main objective of ARTIS is to recover the temporal evolution of the vocal tract shape from the acoustic signal.

This contract started in January 2009 in collaboration with LTCI (Paris), Gipsa-Lab (Grenoble) and IRIT (Toulouse). Its main purpose is the acoustic-to-articulatory inversion of speech signals. Unlike the European project ASPI the approach followed in our group will focus on the use of standard spectra input data, i.e. cepstral vectors. The objective of the project is to develop a demonstrator enabling inversion of speech signals in the domain of second language learning.

This year the work has focused on the development of the inversion from cepstral data as input. We particularly worked on the comparison of cepstral vectors calculated on natural speech and those obtained via the articulatory to acoustic mapping. Bilinear frequency warping was combined with affine adaptation of cepstral coefficients. These two adaptation strategies enable a very good recovery of vocal tract shapes from natural speech. The second topic studied is the access to the codebook. Two pruning strategies, a simple one using the spectral peak corresponding to F2 and a more elaborated one exploiting lax dynamic programming applied on spectral peaks enable a very efficient access to the articulatory codebook used for inversion.

This year, the project focused on the articulatory synthesis in order to generate better sequences of consonant/vowel/consonant by developing time patterns coordinating source and vocal tract dynamics.

8.1.3. ANR ViSAC

Project acronym: VISAC

Project title: Acoustic-Visual Speech Synthesis by Bimodal Unit Concatenation

Duration: January 2009 - June 2013

Coordinator: Slim Ouni

Other partners: Magrit EPI (Inria)

Abstract: The main VISAC objective is to realize the bimodal (audio plus visual) synthesis of speech.

This contract started in January 2009 in collaboration with Magrit Inria team. The purpose of this project is to develop synthesis techniques where speech is considered as a bimodal signal with its acoustic and visual components that are considered simultaneously. This is done by concatenating bimodal diphone units, that is, units that comprise both acoustic and visual information. The latter is acquired using a stereovision technique. The proposed method addresses the problems of asynchrony and incoherence inherent in classic approaches to audiovisual synthesis. Unit selection is based on classic target and join costs from acoustic-only synthesis, which are augmented with a visual join cost. This final year of the project, we have performed an extensive evaluation of the synthesis system using perceptual and subjective evaluations. The overall outcome of the evaluation indicates that the proposed bimodal acoustic-visual synthesis technique provides intelligible speech in both acoustic and visual channels [22].

8.1.4. ANR ORFEO

Project acronym: ORFEO [4]

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 ORFEO objective is the constitution of a Corpus for the Study of Contemporary French.

In this project, we have provided 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 alignment.

8.1.5. 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 will be recorded by German speakers learning French, recoding 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.

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

8.1.6. ANR ContNomina

Project acronym: ContNomina

Project title: Exploitation of context for proper names recognition in the 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 will address:

- 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.

8.1.7. FUI RAPSODIE

Project acronym: RAPSODIE [5]

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.

In this project, the parole team 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.

8.1.8. ADT FASST

The Action de Développement Technologique Inria (ADT) FASST (2012–2014) is conducted by PAROLE in collaboration with the teams PANAMA and TEXMEX of Inria Rennes. It aims to reimplemented into efficient C++ code the Flexible Audio Source Separation Toolbox (FASST) originally developed in Matlab by A. Ozerov, E. Vincent and F. Bimbot in the METISS team of Inria Rennes. This will enable the application of FASST on larger data sets, and its use by a larger audience. The new C++ version will be released early 2014. The second year of the project will be devoted to the integration of FASST with speech recognition software in order to perform noise robust speech recognition.

8.1.9. ADT VisArtico

The technological Development Action (ADT) Inria Visartico just started this November (11/2013 - 10/2015). The purpose of this project is to develop and improve VisArtico, an articulatory vizualisation software. In addition to improve the basic functionalities, several articulatory analysis and processing will be integrated. We will also work on the integration of multimodal data.

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

8.2. European initiatives

8.2.1. Collaborations in European Programs, except FP7

8.2.1.1. Interreg Allegro

Program: Interreg

Project acronym: Allegro

Project title: Adaptive Language LEarning technology for the Greater Region

Duration: 01/01/2009 to 31/12/2012

Coordinator: Saarland University

Other partners: Supélec Metz and DFK Kaiserslautern

Abstract: Allegro is an Interreg project (in cooperation with the Department of COmputational LInguistics and Phonetics of the Saarland University and Supélec Metz) which started in April 2010. It is intended to develop software for foreign language learning. Our contribution consists of developing tools to help learners to master the prosody of a foreign language, i.e. the prosody of English by French learners, and then prosody of French by German learners. We started by recording (with the project Intonale) and segmenting of a corpus made up of English sentences uttered by French speakers and we analyzed specific problems encountered by French speakers when speaking English. The corrections were implemented in Jsnoori. The final review was held on May 15 in Saarbrücken.

8.2.1.2. Eureka - Eurostars i3DMusic

Besides the above contracts of which PAROLE is officially part, E. Vincent is 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: 01/10/2010 to 31/03/2014

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 will be 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. Declared Inria 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.

8.4. International research visitors

8.4.1. Visits of international scientists

- Mouhcin, Chami, INPT, Maroco, June,
- Karima Meftouh, Annaba University, until October,
- Amar Djeradi, USTHB, July, Algeria

9. Dissemination

9.1. Scientific animation

- Members of the team frequently review articles and papers for the following journals and conferences: Computer Speech and Language, Traitement du Signal, ICASSP, INTERSPEECH, AVSP, ASRU, SLAM, CHiME, JEP.
- Member of editorial boards:
  - Computer Speech and Language, special issue on Speech Separation and Recognition in Multisource Environments (E. Vincent) [15]
  - IEEE Transactions on Audio, Speech, and Language Processing (E. Vincent)
  - Eurasip Journal on Audio speech and Music Processing (Y. Laprie)
- Member of scientific committee of conference:
  - INTERSPEECH'2013 (area chair: D. Jouvet)
  - INTERSPEECH'2013 (K. Smaïli)
  - TALN 2013 (K. Smaïli)
  - IEEE Technical Committee on Audio and Acoustic Signal Processing (E. Vincent)
  - Steering Committee of the LVA conference series (E. Vincent)
- Special session:
  - Special Session at Interspeech 2013 "Articulatory data acquisition and processing" (Co-organizer: S. Ouni)
- General chair:
  - 12th International Conference on Auditory-Visual Speech Processing (AVSP2013), August 2013 (S. Ouni)
  - 2nd International Workshop on Machine Listening in Multisource Environments, Vancouver, June 2013 (E. Vincent)
  - Journée de la Fédération Charles Hermite 'Incertitudes: approches et défis', Nancy, December 2013 (E. Vincent)
  - 4th Joint Workshop on Hands-Free Speech Communication and Microphone Arrays, Nancy, May 2014 (E. Vincent)
- Organizer of the 2nd CHiME Speech Separation and Recognition Challenge (E. Vincent).
- Members of the team have been invited as lecturers:
  - Emmanuel Vincent, "Introduction to sound scene analysis - Source localization", "Source separation" and "Source classification", Journées Thématiques IRISA 'Localisation, séparation et suivi de sources audio', Rennes, February 2013 [32], [35], [34]
  - Emmanuel Vincent, "Leveraging online sound exposure and big data", 013 Hearing Aid Developers Forum, Oldenburg, June 2013 [33]

  - Emmanuel Vincent, “Uncertainties in speech and audio processing”, Journée de la Fédération Charles Hermite ‘Incertitudes: approches et défis’, Nancy, December 2013 [73]
  - David Langlois, “Uncertainty in Machine Translation”, Journée de la Fédération Charles Hermite ‘Incertitudes: approches et défis’, Nancy, December 2013
  - Yves Laprie and Slim Ouni, “Articulatory Data Acquisition and Processing”. Colloque Corpus et Outils en Linguistique Langue et Parole : Statuts, Usages et Mésusages, Jul 2013, Strasbourg, France.[31]
  - Anne Bonneau, “Troubles du traitement temporel du langage oral : vers des outils en recherche clinique”, 23ème Congrès de la Société Française de Neurologie Pédiatrique [30]
- Demonstration of VisArtico software at Interspeech 2013 (S. Ouni)

9.2. Teaching - supervision - juries

9.2.1. Teaching

List below is not exhaustive.

PhD

A. Piquard-Kipffer, Language acquisition and language pathology, 25h, Sorbonne Paris Cité University -EHESP, France

Master

K. Smaïli, Master research for ERASMUS MUNDUS LCT, Statistical language modeling for speech recognition and machine translation, Course done in English

K. Smaïli, Master for students from Luxembourg, Business Intelligence

D. Langlois, courses on "vigilance juridique" for teachers at primary school in the scope of using electronic resources (25HTED), ESPE of Academy Nancy-Metz, University of Lorraine, France

D. Langlois and V. Colotte, "Introduction to Speech Analysis and Recognition" (35HTED), University of Lorraine, France

O. Mella, Computer Networking (60HETD), M1, University of Lorraine, France

O. Mella, Project and internship supervision in Networking Engineering (18HETD), M2, University of Lorraine, France

A. Piquard-Kipffer, Language acquisition and language pathology, 200h, M1, M2, University of Lorraine, Blaise Pascal University, Paris 6 University, France

School of engineers

V. Colotte, XML (35HETD), Telecom Nancy, France

Licence

D. Langlois, Algorithmic, C Language (60HETD), Complexity (15HTED), Miage of Nancy, University of Lorraine, France

V. Colotte, C2i - Certificat Informatique et Internet (50HETD), System (80HETP), University of Lorraine, France

A. Piquard-Kipffer, Psycholinguistics, 30h, L3, University of Lorraine, France

O. Mella Networks and Network Programming(30HETD), L3, University of Lorraine, France

O. Mella From Chip to Internet (Introduction about Information Technology) (60 HETD), L1, University of Lorraine, France

I. Illina, Java Programming, 110h, L1, University of Lorraine, France

I. Illina, Graphical Interface, 30h, L1, University of Lorraine, France

I. Illina, Computer Operating System, 50h, L1, University of Lorraine, France

Adults

O. Mella, informatic courses for secondary school teachers (Informatique et Sciences du Numérique courses) (35HTED), ESPE of Academy Nancy-Metz, University of Lorraine, France

D. Langlois, informatic courses for secondary school teachers (Informatique et Sciences du Numérique courses) (100HTED), ESPE of Academy Nancy-Metz, University of Lorraine, France

Other

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

9.2.2. Supervision

HdR : Chiraz Latiri, ”Extraction de connaissances à partir de Textes : Méthodes et Applications", University of Lorraine, 24 June 2013.

HdR : Slim Ouni, ”Multimodal Speech: from articulatory speech to audiovisual speech, University of Lorraine, 29 November 2013.

PhD : Uptala Musti, “Acoustic-Visual Speech Synthesis by Bimodal Unit Selection”, University of Lorraine, 21 Feb 2013, Yves Laprie et Slim Ouni.

PhD: Gabriel Sargent, “Estimation de la structure des morceaux de musique par analyse multi-critères et contrainte de régularité”, University Rennes 1, Feb 2013, Emmanuel Vincent

PhD : Imen Jemaa, Inversion acoustique articulatoire à partir de coefficients cepstraux, University of Lorraine - ENIT (Tunis), 19 February 2013, Yves Laprie et Kaïs Ouni.

PhD : Julie Busset, Inversion acoustique articulatoire à partir de coefficients cepstraux, University of Lorraine, 25 March 2013, Yves Laprie.

PhD : Fadoua Bahja, Détection du fondamental de la parole en temps réel : application aux voix pathologiques, Université Mohammed V-Agdal Faculté des Sciences Rabat Marocco, 11 July 2013, Elhassan Ibn Elhadj and Joseph Di Martino.

PhD: Alexis Benichoux, “Fonctions de coût pour l’estimation des filtres acoustiques dans les mélanges réverbérants”, University Rennes 1, Oct 2013, Emmanuel Vincent

PhD in progress : Cyrine Nasri, New methods for machine translation, October 2011, K. Smaïli, C. Latiri.

PhD in progress : Othman Lachhab, Automatic speech recognition applied to pathological voices, Institut National des Postes et Télécommunications (INPT) Rabat Morocco, from November 2010, Elhassan Ibn Elhaj and Joseph Di Martino.

PhD in progress : Arseniy Gorin, Handling trajectories and speaker consistency in automatic speech recognition, October 2011, D. Jouvet.

PhD in progress : Motaz Saad, Cross-lingual concept mining, November 2011, K. Smaïli and D. Langlois.

PhD in progress : Dung Tran, Uncertainty handling for noise-robust automatic speech recogniton, December 2012, E. Vincent and D. Jouvet.

PhD in progress : Luiza Orosanu, Speech recognition for communication help for deaf or hard of hearing people, December 2012, D. 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, from Jan 2013, Emmanuel Vincent

PhD in progress: Xabier Jaureguiberry, “Fusion et optimisation de modèles pour la séparation de sources audio”, Télécom ParisTech, from Feb 2013, Emmanuel Vincent.

9.2.3. Juries

Participation in PhD thesis Jury for Souhir Gahbiche (Université de Paris-Sud, September 2013), K. Smaïli, reviewer

Participation in PhD thesis Jury for Marion Potet (Université de Grenoble, April 2013), K. Smaïli, reviewer

Participation in HDR Jury for Chiraz Latiti (University of Lorraine), June 2013, K. Smaïli, co-supervisor.

Participation in PhD thesis Jury for Penny Karanasou (Université Paris-Sud, June 2013), D. Jouvet.

Participation in PhD thesis Jury for Mohammed Bouallegue (Université Avignon et Pays du Vaucluse, December 2013), D. Jouvet, reviewer.

Participation in PhD thesis Jury for Tiraogo Abdoulaye Yves Zango (Université de Rennes, February 2013, Y. Laprie, reviewer.

Participation in PhD thesis Jury for Utpala Musti (University of Lorraine), February 2013, V. Colotte, member.

Participation in PhD thesis Jury for Julian Andrés VALDES VARGAS (Université de Grenoble), June 2013, Y. Laprie, reviewer.

Participation in PhD thesis Jury for Sebastien Le Maguer (Université de Rennes - ENSAT Lannion), July 2013, V. Colotte, member.

Participation in PhD thesis Jury for Amel Benamrane (Université de Strasbourg), December 2013, Y. Laprie, reviewer.

Participation in PhD thesis Jury for Julie Busset (University of Lorraine), March 2013, Y. Laprie, supervisor.

Participation in PhD thesis Jury for Imen Jemaa (University of Lorraine - ENIT (Tunis)), February 2013, Y. Laprie, co-supervisor.

Participation in HDR Jury for Slim Ouni (University of Lorraine), February 2013, Y. Laprie, supervisor.

Participation in PhD thesis Jury for Benoit Fuentes (Télécom ParisTech, March 2013), E. Vincent, reviewer

Participation in PhD thesis Jury for Ricard Marxer (University Pompeu Fabra, Barcelona, September 2013), E. Vincent, reviewer

Participation in PhD thesis Jury for Sašo Muševič (University Pompeu Fabra, Barcelona, September 2013), E. Vincent, reviewer

Participation in PhD thesis Jury for François Rigaud (Télécom ParisTech, December 2013), E. Vincent, reviewer

Participation in PhD thesis Jury for Charles Fox (Télécom ParisTech, December 2013), E. Vincent, reviewer

Participation in PhD thesis Jury for Alban Portello (University of Toulouse, December 2013), E. Vincent, reviewer

Participation in PhD thesis Jury for Stanislaw Gorlow (University Bordeaux 1, December 2013), E. Vincent, reviewer

9.2.4. Participation to external committees

Member of a Selection Committee of Le Mans (Université of Maine), LIUM (Laboratoire d'Informatique de l'Université du Maine), D. Langlois

Member of a Selection Committee of University of Lorraine (Ecole des Mines), LORIA, Y. Laprie

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

Member of a Visiting AERES Committee (LIMSI), Y. Laprie

9.2.5. Participation to local committees

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

Member of the Commission de Développement Technologique, A. Bonneau

Member of the Commission développement durable, D. Fohr

Member of Commission de médiation, I. Illina

Member of bureau of Ecole doctorale, I. Illina

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

Elected member of the Conseil de secteur MIAE of University of Lorraine, V. Colotte

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

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

President of the local "Comité des Utilisateurs des Moyens Informatiques" (Inria NGE), D. Fohr

9.3. Popularization

Demonstration (of audio processing and source separation) to high school students at Inria Nancy in March 2013 (E. Vincent).

Demonstration to students of Mines Nancy at Inria Nancy in December 2013 (S. Ouni).

Participation in the Renaissance festival in Nancy in June 2013 (E. Vincent, D. Fohr, Y. Salaün, Y. Laprie, J. di Martino) Demonstrations of automatic speech recognition, speech analysis, source separation, articulatory modeling and voice conversion were presented.

Participation in the 11th Rencontres européennes CNRS Jeunes “Sciences et Citoyens” in Pont-à-Mousson in November 2013 (E. Vincent).

Participation in the Forum Sciences Cognitives, presentation to students in Cognitive Sciences at the University of Lorraine, in November 2013 (S. Ouni) .

10. Bibliography

Major publications by the team in recent years

[1] M. ABBAS, K. SMAÏLI, D. BERKANI. *Multi-category support vector machines for identifying Arabic topics*, in "Journal of Research in Computing Science", 2009, vol. 41

[2] A. BONNEAU, Y. LAPRIE. *Selective acoustic cues for French voiceless stop consonants*, in "The Journal of the Acoustical Society of America", 2008, vol. 123, pp. 4482-4497, http://hal.inria.fr/inria-00336049/en/

[3] C. CERISARA, S. DEMANGE, J.-P. HATON. *On noise masking for automatic missing data speech recognition: a survey and discussion*, in "Computer Speech and Language", 2007, vol. 21, n^{o} 3, pp. 443-457

[4] J.-P. HATON, C. CERISARA, D. FOHR, Y. LAPRIE, K. SMAÏLI. , *Reconnaissance Automatique de la Parole. Du signal à son interprétation*, Dunod, 2006, http://hal.inria.fr/inria-00105908/en/

[5] C. LATIRI, K. SMAÏLI, C. LAVECCHIA, D. LANGLOIS. *Mining monolingual and bilingual corpora*, in "Intelligent Data Analysis", November 2010, vol. 14, n^{o} 6, pp. 663-682, http://hal.inria.fr/inria-00545493/en

[6] C. LAVECCHIA, K. SMAÏLI, D. LANGLOIS, J.-P. HATON. *Using inter-lingual triggers for Machine translation*, in "Eighth conference INTERSPEECH 2007", Antwerp/Belgium, 08 2007, http://hal.inria.fr/inria-00155791/en/

[7] S. OUNI, Y. LAPRIE. *Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion*, in "Journal of the Acoustical Society of America (JASA)", 2005, vol. 118 (1), pp. 444–460, http://hal.archives-ouvertes.fr/hal-00008682/en/

[8] S. RAYBAUD, D. LANGLOIS, K. SMAÏLI. *"This sentence is wrong." Detecting errors in machine-translated sentences.*, in "Machine Translation", August 2011, vol. 25, n° 1, pp. p. 1–34 [*DOI :* 10.1007/s10590-011-9094-9], http://hal.inria.fr/hal-00606350/en

Publications of the year

Doctoral Dissertations and Habilitation Theses

[9] 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, http://hal.inria.fr/tel-00927147

[10] J. BUSSET. , *Inversion acoustique articulatoire à partir de coefficients cepstraux*, Université de Lorraine, March 2013, http://hal.inria.fr/tel-00838913

[11] I. JEMAA. , *Suivi de Formants par analyse en Multirésolution*, Université de Lorraine and Faculté des Sciences de Tunis and Faculté des Sciences de Tunis, February 2013, http://hal.inria.fr/tel-00836717

[12] U. MUSTI. , *Synthèse Acoustico-Visuelle de la Parole par Séléction d'Unités Bimodales*, Université de Lorraine, February 2013, http://hal.inria.fr/tel-00927121

[13] S. OUNI. , *Parole Multimodale : de la parole articulatoire à la parole audiovisuelle*, Université de Lorraine, November 2013, Habilitation à Diriger des Recherches, http://hal.inria.fr/tel-00927119

Articles in International Peer-Reviewed Journals

[14] 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], http://hal.inria.fr/hal-00831660

[15] J. BARKER, E. VINCENT. *Special Issue on Speech Separation and Recognition in Multisource Environments*, in "Computer Speech and Language", February 2013, vol. 27, n° 3, pp. 619-620 [*DOI :* 10.1016/J.CSL.2012.09.005], http://hal.inria.fr/hal-00743532

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

[17] 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], http://hal.inria.fr/hal-00934941

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

[19] S. DEMANGE, S. OUNI. *An episodic memory-based solution for the acoustic-to-articulatory inversion problem*, in "Journal of the Acoustical Society of America", May 2013, vol. 133, n° 5, pp. 2921-2930 [*DOI :* 10.1121/1.4798665], http://hal.inria.fr/hal-00834556

[20] N. DUONG, E. VINCENT, R. GRIBONVAL. *Spatial location priors for Gaussian model based reverberant audio source separation*, in "EURASIP Journal on Advances in Signal Processing", September 2013, 149 p. [*DOI :* 10.1186/1687-6180-2013-149], http://hal.inria.fr/hal-00865125

[21] J. LE ROUX, E. VINCENT. *Consistent Wiener filtering for audio source separation*, in "IEEE Signal Processing Letters", January 2013, vol. 20, n° 3, pp. 217-220 [*DOI :* 10.1109/LSP.2012.2225617], http://hal.inria.fr/hal-00742687

[22] S. OUNI, V. COLOTTE, U. MUSTI, A. TOUTIOS, B. WROBEL-DAUTCOURT, M.-O. BERGER, C. LAVECCHIA. *Acoustic-visual synthesis technique using bimodal unit-selection*, in "EURASIP Journal on Audio, Speech, and Music Processing", June 2013, n° 2013:16 [*DOI :* 10.1186/1687-4722-2013-16], http://hal.inria.fr/hal-00835854

[23] S. OUNI. *Tongue control and its implication in pronunciation training*, in "Computer Assisted Language Learning", January 2013 [*DOI :* 10.1080/09588221.2012.761637], http://hal.inria.fr/hal-00834554

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

[25] S. RACZYNSKI, S. FUKAYAMA, E. VINCENT. *Melody harmonisation with interpolated probabilistic models*, in "Journal of New Music Research", October 2013, vol. 42, n° 3, pp. 223-235 [*DOI :* 10.1080/09298215.2013.822000], http://hal.inria.fr/hal-00876128

[26] S. RACZYNSKI, E. VINCENT, S. SAGAYAMA. *Dynamic Bayesian networks for symbolic polyphonic pitch modeling*, in "IEEE Transactions on Audio, Speech and Language Processing", April 2013, vol. 21, n° 9, pp. 1830-1840, http://hal.inria.fr/hal-00803886

[27] M. SAAD, D. LANGLOIS, K. SMAILI. *Extracting Comparable Articles from Wikipedia and Measuring their Comparabilities*, in "Procedia - Social and Behavioral Sciences", 2013, vol. 95, pp. 40-47 [*DOI :* 10.1016/J.SBSPRO.2013.10.620], http://hal.inria.fr/hal-00907442

[28] E. VINCENT, N. BERTIN, R. GRIBONVAL, F. BIMBOT. *From blind to guided audio source separation*, in "IEEE Signal Processing Magazine", May 2014, http://hal.inria.fr/hal-00922378

Articles in Non Peer-Reviewed Journals

[29] A. PIQUARD-KIPFFER, L. SPRENGER-CHAROLLES. *Early predictors of future reading skills: A follow-up of French-speaking children from the beginning of kindergarten to the end of the second grade (age 5 to 8)*, in "L'Année psychologique - Topics in Cognitive Psychology", 2013, vol. 113, n° 4, pp. 491-521 [*DOI :* 10.4074/S0003503313014012], http://hal.inria.fr/hal-00925411

Invited Conferences

[30] A. BONNEAU. *Troubles du traitement temporel du langage oral : vers des outils en recherche clinique*, in "23eme SFNP congres de la société française de neurologie pédiatrique", Nancy, France, SFNP société française de neurologie pédiatrique, January 2013, http://hal.inria.fr/hal-00925640

[31] Y. LAPRIE, S. OUNI. *Articulatory Data Acquisition and Processing*, in "Colloque Corpus et Outils en Linguistique Langue et Parole : Statuts, Usages et Mésusages", Strasbourg, France, July 2013, http://hal.inria.fr/hal-00921142

[32] E. VINCENT. *Introduction to sound scene analysis - Source localization*, in "Journées Thématiques 'Localisation, séparation et suivi de sources audio'", Rennes, France, February 2013, http://hal.inria.fr/hal-00833535

[33] E. VINCENT. *Leveraging online sound exposure and big data*, in "2013 Hearing Aid Developers Forum", Oldenburg, Germany, June 2013, http://hal.inria.fr/hal-00833239

[34] E. VINCENT. *Source classification*, in "Journées Thématiques 'Localisation, séparation et suivi de sources audio'", Rennes, France, February 2013, http://hal.inria.fr/hal-00833537

[35] E. VINCENT. *Source separation*, in "Journées Thématiques 'Localisation, séparation et suivi de sources audio'", Rennes, France, February 2013, http://hal.inria.fr/hal-00833536

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

International Conferences with Proceedings

[37] K. BARTKOVA, D. JOUVET. *Automatic Detection of the Prosodic Structures of Speech Utterances*, in "SPECOM - 15th International Conference on Speech and Computer - 2013", Pilsen, Czech Republic, M. ŽELEZNÝ, I. HABERNAL, A. RONZHIN (editors), Lecture Notes in Artificial Intelligence, Springer Verlag, September 2013, vol. 8113, pp. 1-8, http://hal.inria.fr/hal-00834318

[38] A. BENICHOUX, E. VINCENT, R. GRIBONVAL. *A fundamental pitfall in blind deconvolution with sparse and shift-invariant priors*, in "ICASSP - 38th International Conference on Acoustics, Speech, and Signal Processing - 2013", Vancouver, Canada, May 2013, http://hal.inria.fr/hal-00800770

[39] 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. , http://hal.inria.fr/hal-00931859

[40] J. BUSSET, Y. LAPRIE. *Acoustic-to-articulatory inversion by analysis-by-synthesis using cepstral coefficients*, in "ICA - 21st International Congress on Acoustics - 2013", Montréal, Canada, June 2013, http://hal.inria.fr/hal-00836808

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

[42] D. FOHR, O. MELLA. *Combination of Random Indexing based Language Model and N-gram Language Model for Speech Recognition*, in "INTERSPEECH - 14th Annual Conference of the International Speech Communication Association - 2013", Lyon, France, August 2013, http://hal.inria.fr/hal-00833898

[43] A. GORIN, D. JOUVET. *Efficient constrained parametrization of GMM with class-based mixture weights for Automatic Speech Recognition*, in "LTC'13 - 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics", Poznań, Poland, December 2013, http://hal.inria.fr/hal-00923202

[44] S. HARRAT, K. MEFTOUH, M. ABBAS, K. SMAÏLI. *Diacritics Restoration for Arabic Dialects*, in "INTERSPEECH 2013", Lyon, France, ISCA, August 2013, http://hal.inria.fr/hal-00925815

[45] N. ITO, E. VINCENT, N. ONO, S. SAGAYAMA. *General algorithms for estimating spectrogram and transfer functions of target signal for blind suppression of diffuse noise*, in "2013 IEEE International Workshop on Machine Learning for Signal Processing", Southampton, United Kingdom, September 2013, http://hal.inria.fr/hal-00849791

[46] X. JAUREGUIBERRY, G. RICHARD, P. LEVEAU, R. HENNEQUIN, E. VINCENT. *Introducing a simple fusion framework for audio source separation*, in "2013 IEEE International Workshop on Machine Learning for Signal Processing", Southampton, United Kingdom, September 2013, 6 p. , http://hal.inria.fr/hal-00846834

[47] I. JEMAA, K. OUNI, Y. LAPRIE, S. OUNI, J.-P. HATON. *A new Automatic Formant Tracking approach based on scalogram maxima detection using complex wavelets*, in "CEIT - International Conference on Control, Engineering & Information Technology - 2013", Sousse, Tunisia, June 2013, http://hal.inria.fr/hal-00836854

[48] 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, http://hal.inria.fr/hal-00834296

[49] 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, http://hal.inria.fr/hal-00834282

[50] D. JOUVET, D. LANGLOIS. *A Machine Learning Based Approach for Vocabulary Selection for 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. 60-67, http://hal.inria.fr/hal-00834302

[51] D. LANGLOIS, K. SMAÏLI. *LORIA System for the WMT13 Quality Estimation Shared Task*, in "ACL 2013 Eighth Workshop on Statistical Machine Translation", Sofia, Bulgaria, 2013, http://hal.inria.fr/hal-00923623

[52] Y. LAPRIE, M. LOOSVELT, S. MAEDA, R. SOCK, F. HIRSCH. *Articulatory copy synthesis from cine X-ray films*, in "InterSpeech - 14th Annual Conference of the International Speech Communication Association - 2013", Lyon, France, August 2013, http://hal.inria.fr/hal-00836838

[53] S. MAEDA, Y. LAPRIE. *Vowel and prosodic factor dependent variations of vocal-tract length*, in "InterSpeech - 14th Annual Conference of the International Speech Communication Association - 2013", Lyon, France, August 2013, http://hal.inria.fr/hal-00836829

[54] J. MIRANDA, S. OUNI. *Mixing faces and voices: a study of the influence of faces and voices on audiovisual intelligibility.*, in "AVSP - 12th International Conference on Auditory-Visual Speech Processing - 2013", Annecy, France, August 2013, http://hal.inria.fr/hal-00835855

[55] U. MUSTI, V. COLOTTE, S. OUNI, C. LAVECCHIA, B. WROBEL, M.-O. BERGER. *Automatic Feature Selection for Acoustic-Visual Concatenative Speech Synthesis: Towards a Perceptual Objective Measure*, in "AVSP - Audio Visual Speech Processing", Annecy, France, September 2013, http://hal.inria.fr/hal-00925115

[56] I. NKAIRI, I. ILLINA, G. LINARÈS, D. FOHR. *Exploring temporal context in diachronic text documents for automatic OOV proper name retrieval*, in "Language & Technology Conference", Poznań, Poland, December 2013, pp. 540-544, http://hal.inria.fr/hal-00924696

[57] L. OROSANU, D. JOUVET. *Comparison and Analysis of Several Phonetic Decoding Approaches*, 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. 161-168, http://hal.inria.fr/hal-00834313

[58] L. OROSANU, D. JOUVET. *Comparison of approaches for an efficient phonetic decoding*, in "InterSpeech - 14th Annual Conference of the International Speech Communication Association - 2013", Lyon, France, August 2013, http://hal.inria.fr/hal-00834284

[59] M. PUIGT, E. VINCENT, Y. DEVILLE, A. GRIFFIN, A. MOUCHTARIS. *Effects of audio coding on ICA performance: an experimental study*, in "11th IEEE Int. Workshop on Electronics, Control, Measurement, Signals and their application to Mechatronics", Toulouse, France, June 2013, http://hal.inria.fr/hal-00817202

[60] M. SAAD, D. LANGLOIS, K. SMAÏLI. *Comparing Multilingual Comparable Articles Based On Opinions*, in "Proceedings of the 6th Workshop on Building and Using Comparable Corpora", Sofia, Bulgaria, Association for Computational Linguistics ACL, August 2013, pp. 105-111, http://hal.inria.fr/hal-00851959

[61] I. STEINER, K. RICHMOND, S. OUNI. *Speech animation using electromagnetic articulography as motion capture data*, in "AVSP - 12th International Conference on Auditory-Visual Speech Processing - 2013", Annecy, France, August 2013, pp. 55-60, http://hal.inria.fr/hal-00835856

[62] J. THIEMANN, E. VINCENT. *A fast EM algorithm for Gaussian model-based source separation*, in "EUSIPCO - 21st European Signal Processing Conference - 2013", Marrakech, Morocco, September 2013, http://hal.inria.fr/hal-00840366

[63] J. THIEMANN, E. VINCENT. *An experimental comparison of source separation and beamforming techniques for microphone array signal enhancement*, in "MLSP - 23rd IEEE International Workshop on Machine Learning for Signal Processing - 2013", Southampton, United Kingdom, September 2013, http://hal.inria.fr/hal-00850173

[64] D. TRAN, E. VINCENT, D. JOUVET, K. ADILOGLU. *Using full-rank spatial covariance models for noise-robust ASR*, in "CHiME - 2nd International Workshop on Machine Listening in Multisource Environments - 2013", Vancouver, Canada, June 2013, pp. 31-32, http://hal.inria.fr/hal-00801162

[65] E. VINCENT, J. BARKER, S. WATANABE, J. LE ROUX, F. NESTA, M. MATASSONI. *The Second 'CHiME' Speech Separation and Recognition Challenge: An overview of challenge systems and outcomes*, in "2013 IEEE Automatic Speech Recognition and Understanding Workshop", Olomouc, Czech Republic, December 2013, http://hal.inria.fr/hal-00862750

[66] E. VINCENT, J. BARKER, S. WATANABE, J. LE ROUX, F. NESTA, M. MATASSONI. *The second 'CHiME' Speech Separation and Recognition Challenge: Datasets, tasks and baselines*, in "ICASSP - 38th International

Conference on Acoustics, Speech, and Signal Processing - 2013", Vancouver, Canada, May 2013, pp. 126-130, http://hal.inria.fr/hal-00796625

National Conferences with Proceedings

[67] J. BUSSET, M. CADOT. *Fouille d'images animées : cinéradiographies d'un locuteur*, in "FOSTA 2013, atelier de EGC 2013", Toulouse, France, January 2013, pp. 1-12, http://hal.inria.fr/hal-00773448

Conferences without Proceedings

[68] M. BARKAT-DEFRADAS, C. FAUTH, F. HIRSCH, B. AMY DE LA BRETÈQUE, J. SAUVAGE, C. DODANE. *Rauque 'n' Roll : La raucité, entre symptôme pathologique & expression artistique*, in "5° Journées de Phonétique Clinique", Liège, Belgium, December 2013, http://hal.inria.fr/hal-00918332

[69] L. CATANESE, N. SOUVIRAÀ-LABASTIE, B. QU, S. CAMPION, G. GRAVIER, E. VINCENT, F. BIMBOT. *MODIS: an audio motif discovery software*, in "Show & Tell - Interspeech 2013", Lyon, France, August 2013, http://hal.inria.fr/hal-00931227

[70] L. S. R. SIMON, A. VIMOND, E. VINCENT. *Effects of audio latency in a disc jockey interface*, in "21st International Congress on Acoustics", Montreal, Canada, June 2013, http://hal.inria.fr/hal-00798322

[71] J. THIEMANN, N. ITO, E. VINCENT. *The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings*, in "21st International Congress on Acoustics", Montreal, Canada, Acoustical Society of America, June 2013, http://hal.inria.fr/hal-00796707

[72] E. VINCENT, J. BARKER, S. WATANABE, J. LE ROUX, F. NESTA, M. MATASSONI. *Overview of the 2nd 'CHiME' Speech Separation and Recognition Challenge*, in "CHiME - 2nd International Workshop on Machine Listening in Multisource Environments - 2013", Vancouver, Canada, June 2013, http://hal.inria.fr/hal-00833252

[73] E. VINCENT. *Incertitudes en traitement de la parole et de l'audio*, in "Journée scientifique de la Fédération Charles Hermite " Incertitudes: approches et défis "", Nancy, France, December 2013, http://hal.inria.fr/hal-00923070

Research Reports

[74] S. RACZYNSKI, E. VINCENT. , *Genre-based music language modelling with latent hierarchical Pitman-Yor process allocation*, Inria, March 2013, n^{o} RT-0434, http://hal.inria.fr/hal-00804567

[75] A. ROUSSEAU, A. DARNAUD, B. GOGLIN, C. ACHARIAN, C. LEININGER, C. GODIN, C. HOLIK, C. KIRCHNER, D. RIVES, E. DARQUIE, E. KERRIEN, F. NEYRET, F. MASSEGLIA, F. DUFOUR, G. BERRY, G. DOWEK, H. ROBAK, H. XYPAS, I. ILLINA, I. GNAEDIG, J. JONGWANE, J. EHREL, L. VIENNOT, L. GUION, L. CALDERAN, L. KOVACIC, M. COLLIN, M.-A. ENARD, M.-H. COMTE, M. QUINSON, M. OLIVI, M. GIRAUD, M. DORÉMUS, M. OGOUCHI, M. DROIN, N. LACAUX, N. ROUGIER, N. ROUSSEL, P. GUITTON, P. PETERLONGO, R.-M. CORNUS, S. VANDERMEERSCH, S. MAHEO, S. LEFEBVRE, S. BOLDO, T. VIÉVILLE, V. POIREL, A. CHABREUIL, A. FISCHER, C. FARGE, C. VADEL, I. ASTIC, J.-P. DUMONT, L. FÉJOZ, P. RAMBERT, P. PARADINAS, S. DE QUATREBARBES, S. LAURENT. , *Médiation Scientifique : une facette de nos métiers de la recherche*, March 2013, 34 p. , http://hal.inria.fr/hal-00804915

Patents and standards

[76] G. KERGOURLAY, J. CITÉRIN, E. NGUYEN, L. LE SCOLAN, J. THIEMANN, E. VINCENT, N. BERTIN, F. BIMBOT. , *Sound source separation method*, 2013, n° 1313218.8, http://hal.inria.fr/hal-00923802

[77] G. KERGOURLAY, J. THIEMANN, E. VINCENT, N. BERTIN, F. BIMBOT. , *Method and apparatus for sound source separation based on a binary activation model*, 2013, n° 1304774.1, http://hal.inria.fr/hal-00923803

References in notes

[78] C. ABRY, T. LALLOUACHE. *Le MEM: un modèle d'anticipation paramétrable par locuteur: Données sur l'arrondissement en français*, in "Bulletin de la communication parlée", 1995, vol. 3, n° 4, pp. 85–89

[79] F. BAHJA, J. DI MARTINO, E. H. IBN ELHAJ. *Real-Time Pitch Tracking using the eCate Algorithm*, in "5th International Symposium on I/V Communications over fixed and Mobile Networks - ISIVC 2010", Rabat, Maroc, 2010, pp. 1-4, ISBN : 978-1-4244-5996-4 [*DOI :* 10.1109/ISVC.2010.5656254], http://hal.inria.fr/inria-00545435

[80] P. F. BROWN. *A statistical Approach to MAchine Translation*, in "Computational Linguistics", 1990, vol. 16, pp. 79-85

[81] M. CHAMI, J. DI MARTINO, L. PIERRON, E. H. IBN ELHAJ. *Real-Time Signal Reconstruction from Short-Time Fourier Transform Magnitude Spectra Using FPGAs*, in "5th. International Conference on Information Systems and Economic Intelligence - SIIE 2012", Djerba, Tunisie, 2012, http://hal.inria.fr/hal-00761783

[82] R. CLARK, K. RICHMOND, S. KING. *Festival 2 - Build your own general purpose unit selection speech synhtesiser*, in "ISCA 5th Speech Synthesis Workshop", Pittsburgh, 2004, pp. 201–206

[83] M. COHEN, D. MASSARO. , *Modeling coarticulation in synthetic visual speech*, 1993

[84] V. COLOTTE, R. BEAUFORT. *Linguistic features weighting for a Text-To-Speech system without prosody model*, in "proceedings of EUROSPEECH/INTERSPEECH 2005", 2005, pp. 2549-2552, http://hal.ccsd.cnrs.fr/ccsd-00012561/en/

[85] J. DI MARTINO, Y. LAPRIE. *An Efficient F0 Determination Algorithm Based on the Implicit Calculation of the Autocorrelation of the Temporal Excitation Signal*, in "6th European Conference on Speech Communication & Technology - EUROSPEECH'99", Budapest, Hongrie, 1999, 4 p. , http://hal.inria.fr/inria-00098759

[86] J. DI MARTINO, L. PIERRON. , *Synthétiseur numérique audio amélioré*, June 2010, n° 10/02674, patent no. 10/02674, Oesovox, http://hal.inria.fr/inria-00546967

[87] E. FARNETANI. *Labial coarticulation*, in "In Coarticulation: Theory, data and techniques", Cambridge, W. J. HARDCASTLE, N. HEWLETT (editors), Cambridge university press, 1999, chap. 8

[88] M.-C. HATON. *The teaching wheel: an agent for site viewing and subsite building*, in "Int. Conf. Human-Computer Interaction", Heraklion, Greece, 2003

[89] J.-P. HATON, C. CERISARA, D. FOHR, Y. LAPRIE, K. SMAÏLI. , *Reconnaissance Automatique de la Parole Du signal à son interprétation*, UniverSciences (Paris) - ISSN 1635-625X, DUNOD, 2006, 392 p. , http://hal.inria.fr/inria-00105908/en/

[90] A. KIPFFER-PIQUARD. , *Prédiction de la réussite ou de l'échec spécifiques en lecture au cycle 2. Suivi d'une population "à risque" et d'une population contrôle de la moyenne section de maternelle à la deuxième année de scolarisation primaire*, ARNT - Lille, 2006, 277 p. , http://hal.inria.fr/inria-00185312/en/

[91] A. KIPFFER-PIQUARD. *Prédiction dès la maternelle de la réussite et de l'échec spécifique à l'apprentissage de la lecture en fin de cycle 2*, in "Les troubles du développement chez l'enfant", Amiens France, L'HARMATTAN, 2007, http://hal.inria.fr/inria-00184601/en/

[92] P. KOEHN, H. HOANG, A. BIRCH, C. CALLISON-BURCH, M. FEDERICO, N. BERTOLDI, B. COWAN, W. SHEN, C. MORAN, R. ZENS, C. DYER, O. BOJAR, A. CONSTANTIN, E. HERBST. *Moses: Open Source Toolkit for Statistical Machine Translation*, in "Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session", June 2007

[93] P. KOEHN. *Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models*, in "6th Conference Of The Association For Machine Translation In The Americas", Washington, DC, USA, 2004, pp. 115-224

[94] C. LAVECCHIA, K. SMAÏLI, D. LANGLOIS. *Building a bilingual dictionary from movie subtitles based on inter-lingual triggers*, in "Translating and the Computer", Londres Royaume-Uni, 2007, http://hal.inria.fr/inria-00184421/en/

[95] C. LAVECCHIA, K. SMAÏLI, D. LANGLOIS, J.-P. HATON. *Using inter-lingual triggers for Machine translation*, in "Eighth conference INTERSPEECH 2007", Antwerp/Belgium, 08 2007, http://hal.inria.fr/inria-00155791/en/

[96] F. J. OCH, H. NEY. *Improved statistical alignment models*, in "ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics", Morristown, NJ, USA, Association for Computational Linguistics, 2000, pp. 440–447

[97] 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, États-Unis, September 2012, http://hal.inria.fr/hal-00730733

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