INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team Parole

Analysis, perception and recognition of speech

Nancy - Grand Est

Theme : Audio, Speech, and Language Processing

Activity Report

2010

Table of contents

is joint project to INRIA, CNRS, Henri Poincaré University and Nancy 2 University through LORIA laboratory (UMR 7503). For more details, we invite the reader to consult the team web site at http://parole.loria.fr/.

1. Team

Research Scientists

Yves Laprie [Team Leader, Research Director CNRS, HdR]
Anne Bonneau [Research scientist CNRS]
Christophe Cerisara [Research scientist CNRS, HdR]
Dominique Fohr [Research scientist CNRS]
Denis Jouvet [Research Director INRIA, HdR]

Faculty Members

Vincent Colotte [Assistant Professor, Henri Poincaré University]
Joseph di Martino [Assistant Professor, Henri Poincaré University]
Jean-Paul Haton [Professor emerit, Henri Poincaré University, Institut Universitaire de France]
Marie-Christine Haton [Professor emerit, Henri Poincaré University, HdR]
Irina Illina [Assistant Professor, I.U.T. Charlemagne, Nancy 2 University, HdR]
David Langlois [Assistant Professor, IUFM, Henri Poincaré University]
Agnès Kipffer-Piquard [Assistant Professor, IUFM, Henri Poincaré University]
Odile Mella [Assistant Professor, Henri Poincaré University]
Slim Ouni [Assistant Professor, I.U.T. Charlemagne, Nancy 2 University]
Kamel Smaïli [Professor, Nancy 2 University, HdR]
Sébastien Demange [ATER until August 2010, INPL]

Technical Staff

Fabian Monnay [ADT Handicom]
Benjamin Ujvari-Cseh [ADT Handicom, September to December 2010]

PhD Students

Christian Gillot [MENRT grant, thesis to be defended in 2012]
Caroline Lavecchia [ATER I.U.T. Charlemagne, Nancy 2 University, thesis defended in June 2010]
Sylvain Raybaud [MENRT grant, thesis to be defended in 2011]
Ammar Werghi [COADVISE-FP7 program since October 2009]
Fadoua Bahja [COADVISE-FP7 program since May 2009]
Julie Busset [CNRS since 1st September 2009]
Utpala Musti [INRIA Cordi grant since 1st October 2009]
Imen Jemaa [Joint supervision with ENIT Tunis, since April 2009]
Nadia Amar [Joint supervision with ENIT Tunis, thesis to be defended in 2011]

Post-Doctoral Fellows

Eren Akdemir [Since October 15th, 2010]
Larbi Mesbahi [Since December 1st, 2010, Interreg Allegro project]
Asterios Toutios [University Nancy 2, ANR Visac]
Frederick Stouten [until August 2010]
Frédéric Tantini
Sébastien Demange [from September 2010, Eurostar Emospeech project]

Administrative Assistant

Hélène Zganic [INRIA]

2. Overall Objectives

2.1. Overall Objectives

PAROLE is a joint project to INRIA, CNRS, Henri Poincaré University and Nancy 2 University 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 variabilities, 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 Langage 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.

Our policy in terms of industrial partnership consists in favoring contracts that quite precisely fit our scientific objectives. We are involved in an ANR project about audiovisual speech synthesis, another about acoustic-to-articulatory inversion of speech (ARTIS), another about the processing of articulatory data (DOCVACIM) and in a national evaluation campaign of automatic speech recognition systems (ESTER). We also coordinated until January 2009 the 6th PCRD project ASPI about acoustic-to-articulatory inversion of speech, and the Rapsodis ARC until october 2009. Additionally, we are also participating to a number of regional projects.

3. Scientific Foundations

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).

[1] Hidden Markov 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.

Within the framework of a new project (see 7.2.2), 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. Phonemic discrimination in language acquisition and language disabilities

We have started the development of 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. Reading acquisition in alphabetic systems is described as depending on the efficiency of phonological skills which link oral and written language. Phonemic awareness seems to be strongly linked to success or specific failure in reading acquisition. A fair proportion of dyslexic and dysphasic 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 [42], [43] which indicates that phonemic discrimination at the beginning of kindergarten (at age 5) can predict some 25% of the variance in reading level at the end of Grade 2 (at age 8). This longitudinal study showed that there was a difference of numbers of errors between a "control group" and a group "at risk" for dyslexia when presented with pairs of pseudowords which differ only by a single phonemic feature. Our goal was to specify if there was a difference of type of errors between these two groups of children. Identifying reading and reading related-skills in dyslexic teenagers was our second goal. We used EVALEC, the computerized tool developed by [55].

In the field of dysphasia, our goal was to contribute to identify the nature of the phonemic discrimination difficulties with dysphasic children. Do the profiles of dysphasic children differ from those who are simply retarded speakers. Is there a difference in number of errors or of type of errors ?

3.2.1.3. 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.

Works on acoustic-to-articulatory inversion widely rely on an analysis by synthesis approach that covers three essential aspects:

Solving acoustic equations. In order to solve the acoustic equations adapted to the vocal tract, one assumes that the sound wave is a plane wave in the vocal tract and that it can be unbend. There are two families of solving methods:

(i) frequency methods through the acoustical-electrical analogy,

(ii) spatio-temporal methods, through the direct solving of finite difference equations derived from Webster equations.

Measuring the vocal tract. This represents an important obstacle because there does not exist any reliable method enabling a precise measurement in time and dimension. MRI (Magnetic Resonance Imaging) enables 3D measurements but is not sufficiently fast and X-rays only allow a sagittal slice of the vocal tract to be captured while involving not acceptable health hazards.

Articulatory modeling. Articulatory models aim at describing all the possible vocal tract shapes with a small number of parameters, while preserving deformations observed on a real vocal tract. Present articulatory models often derive from data analysis of cineradiography moving pictures. One of the most widely used is the one built by Maeda [50].

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, MFCC (Mel Frequency Cepstral Coefficients) 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 coarticulation 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 [39] 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 (Vowel Consonant Vowel) 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 [34] advocates for the theory of expansion movements: the movement tends to be anticipated when no phonological constraint is imposed on labiality. Cohen and Massaro [37] 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 [36] 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 [38] 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 absence of any 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, a chunker (shallow parser) can be introduced to give information about a potential rhythmic structure. Moreover, to perform the selection, an algorithm can be used to automatically weight the linguistic features given by the NLP. Our method relies on acoustic clustering and entropy information [38]. 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 ngrams).

The main challenge of automatic speech recognition is its robustness to multiple sources of speech variability [41]. Among them, we have been focusing on acoustic environment, inter- and intra-speaker variability, different speaking styles (prepared speech, spontaneous speech...) and non-native pronunciations.

Another specificity of automatic speech recognition is the necessity to combine efficiently all the research works (in acoustic modeling, langage 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 or syntax 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 word 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 (ASR) 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 alignement 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 Parole 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 adress 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 adress 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 [35] 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 [49]. 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 alignement 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 consist 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. [51] 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 algorithms.

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 [45] or MOSES [44] 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 [40] 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

5.1. WinSnoori

Participant: Yves Laprie.

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 main improvement concerns automatic formant tracking which is now available with other tools for copy synthesis. It is now possible to determine parameters for the formant synthesizer of Klatt quite automatically. The first step is formant tracking, then the determination of F0 parameters and finally the adjustment of formant amplitudes for the parallel branch of the Klatt synthesizer enable a synthetic speech signal to be generated. The automatic formant tracking that has been implemented is an improved version of the concurrent curve formant tracking [47]. One key point of this tracking algorithm is the construction of initial rough estimates of formant trajectories. The previous algorithm used a mobile average applied onto LPC roots. The window is sufficiently large (200 ms) to remove fast varying variations due to the detection of spurious roots. The counterpart of this long duration is that the mobile average prevents formants fairly far from the mobile average to be kept. This is particularly sensitive in the case of F2 which presents low frequency values for back vowels. A simple algorithm to detect back vowels from the overall spectral shape and particularly energy levels has been added in order to keep extreme values of F2 which are relevant.

Together with other improvements reported during the last years, formant tracking enables copy synthesis. The current version of WinSnoori is available on http://www.winsnoori.fr.

5.2. LABCORP

Participants: David Langlois, Kamel Smaïli.

In the past, we developed a labelling tool which allows syntactic ambiguities to be solved. The syntactic class of each word is assigned depending on its effective context. This tool is based on a large dictionary (230000 lemmas) extracted from BDLEX and a set of 230 classes determined by hand. This tool has a labelling error of about 1 %.

Such a tool is dedicated to tag a text with predefined sets of *Parts of Speech*. A tagger needs a time-consuming manual pre-tagging to bootstrap the training parameters. It is then difficult to test numerous tag sets as needed for our research activities. However, this stage could be skipped [46]. That's why we developed another tagger based on a unsupervised tagging algorithm. This method has been used to estimate the parameters of a new tagger using the classes of the former one. The new tagger is now integrated into the TTS platform developed in the team (see 5.11).

5.3. Automatic lexical clustering

Participants: David Langlois, Kamel Smaïli.

In order to adapt language models in ASR applications, we use a toolkit we developed in the past. This tool automatically creates word classes. This toolkit exploits the simulated annealing algorithm. Creating these classes requires a vocabulary (set of words) and a training corpus. The resulting set of classes minimizes the perplexity of the corresponding language model. Several options are available: the user can fix the resulting number of classes, the initial classification, the value of the final perplexity, etc.

5.4. SUBWEB

Participants: David Langlois, Kamel Smaïli.

We published in 2007 a method which allows to align sub-titles comparable corpora [48]. 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. ESPERE

Participant: Dominique Fohr.

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

ESPERE (Engine for SPEech REcognition) is an HMM-based toolbox for speech recognition which is composed of three processing stages: an acoustic front-end, a training module and a recognition engine. The acoustic front-end is based on MFCC parameters: the user can customize the parameters of the filterbank and the analyzing window.

The training module uses Baum-Welch re-estimation algorithm with continuous densities. The user can define the topology of the HMM models. The modeled units can be words, phones or triphones and can be trained using either an isolated training or an embedded training.

The recognition engine implements a one-pass time-synchronization algorithm using the lexicon of the application and a grammar. The structure of the lexicon allows the user to give several pronunciations per word. The grammar may be word-pair or bigram.

ESPERE contains more than 20000 C++ lines and runs on PC-Linux or PC-Windows.

5.6. SELORIA

Participant: Odile Mella.

SELORIA is a toolbox for speaker diarization.

The system contains the following steps:

- Speaker change detection: to find points in the audio stream which are candidates for speaker change points, a distance is computed between two Gaussian modeling data of two adjacent given-length windows. By sliding both windows on the whole audio stream, a distance curve is obtained. A peak in this curve is thus considered as a speaker change point.
- Segment recombination: too many speaker turn points detected during the previous step result in a lot of false alarms. A segment recombination using BIC is needed to recombine adjacent segments uttered by the same speaker.
- Speaker clustering: in this step, speech segments of the same speaker are clustered. Top-down clustering techniques or bottom-up hierarchical clustering techniques using BIC can be used.
- Viterbi re-segmentation: the previous clustering step provides enough data for every speaker to estimate multi-gaussian speaker models. These models are used by a Viterbi algorithm to refine the boundaries between speakers.
- Second speaker clustering step (called cluster recombination). This step uses Universal Background Models (UBM) and the Normalized Cross Likelihood Ratio (NCLR) measure.

This toolbox is derived from mClust designed by LIUM.

5.7. ANTS

Participants: Dominique Fohr, Irina Illina, Odile Mella.

The aim of the Automatic News Transcription System (ANTS) is to transcribe radio broadcast news. ANTS is composed of five stages: broad-band/narrow-band speech segmentation, speech/music classification, speaker segmentation and clustering, detection of silences/breathing segments and large vocabulary speech recognition. The three first stages split the audio stream into homogeneous segments with a manageable size and allow the use of specific algorithms or models according to the nature of the segment.

Speech recognition is based on the Julius engine and 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.

A real time version of ANTS has been developped. The transcription is done in real time on a quad-core PC.

5.8. J-Safran

Participant: Christophe Cerisara.

J-Safran is the “Java Syntaxico-semantic French Analyser”. Its development has started in June 2009 from the collaboration between Parole and Talaris in the context of the RAPSODIS project. It is an open-source dependency parser that is dedicated to oral speech. J-Safran has participated to the Passage evaluation campaign in 2009/2010; J-Safran ranked around the average amongst all other participating systems, which is satisfactory given its very recent creation. Recent progresses concern the inclusion of partial Semantic Role Labeling inference on top of syntactic parsing.

It is distributed under the Cecill-C licence, and can be downloaded at http://www.loria.fr/~cerisara/jsafran/index.html

5.9. JTRANS

Participant: Christophe Cerisara.

JTrans is an open-source software for semi-automatic alignement of speech and textual corpus. It is written 100% in JAVA and exploits libraries developed since several years in our team. Two algorithms are available for automatic alignment: a block-Viterbi and standard forced-alignement Viterbi. The latter is used when manual anchors are defined, while the former is used for long audio files that do not fit in memory. It is designed to be intuitive and easy to use, with a focus on GUI design. The rationale behind JTrans is to let the user control and check on-the-fly the automatic alignment algorithms. It is bundled for now with a French phonetic lexicon and French models.

Recent improvements include automatic speech transcription in addition to text/speech alignement. JTrans is developed in the context of the CPER MISN TALC project, in collaboration between the Parole and Talaris INRIA teams, and CNRS researchers from the ATILF laboratory. It is distributed under the Cecill-C licence, and can be downloaded at http://jtrans.gforge.inria.fr

5.10. STARAP

Participants: Dominique Fohr, Odile Mella.

STARAP (Sous-Titrage Aidé par la Reconnaissance Automatique de la Parole) is a toolkit to help the making of sub-titles for TV shows. This toolkit performs:

- Parameterization of speech data;
- Clustering of parameterized data;
- Gaussian Mixture Models (GMM) training;
- Viterbi recognition.

This toolkit was realised in the framework of the STORECO contract and the formats of the input and output files are compatible with HTK toolkit.

5.11. TTS SoJA

Participant: Vincent Colotte.

TTS SoJA (Speech synthesis platform in Java) is a text-to-speech synthesis software system. 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 algorithms. The software performs all steps from 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 launched with a graphical user interface or directly integrated in a Java code or by following the client-server paradigm.

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

6. New Results

6.1. Speech Analysis and Synthesis

Participants: Anne Bonneau, Vincent Colotte, Dominique Fohr, Yves Laprie, Joseph di Martino, Slim Ouni, Asterios Toutios, Nadia Amar, Imen Jemaa, Sébastien Demange, Ammar Werghi, Fadoua Bahja, Agnès Piquard-Kipffer, Utpala Musti, Fabian Monnay.

6.1.1. Acoustic-to-articulatory inversion

6.1.1.1. Building new articulatory models

The possibility of generating the same sounds as those uttered by the speaker (or at least vocal tract transfer functions not too far from those observed) via the articulatory model and the acoustic simulation constitutes the underlying hypothesis of an analysis by synthesis method of acoustic-to-articulatory inversion. The articulatory model, and consequently its construction, thus plays a crucial role in inversion. We exploited four X-ray films for this purpose. The first step consisted in developing a software tool, called Xarticulators, offering a large spectrum of automatic and manual tools to outline contours. Then, we chose to use curvilinear coordinates to approximate tongue contours because this allows any tongue shape (rounded and moved very back in the vocal tract or even retroflex) to be described correctly. In order to increase the phonetic coverage of the films, MRI images of the same speaker have been registered and incorporated in the analysis. The tongue shape was analyzed via PCA after the subtraction of the mandible movement. The linear model obtained was complemented by a collision algorithm managing interactions with the epiglottis considered as a passive articulator.

The resulting model was evaluated on a database of contours outlined from X-ray films used to construct the articulatory model designed by Maeda. The deviation is less than 1mm and all tongue shapes, including those of consonants are well approximated.

6.1.1.2. Adaptation of cepstral coefficients for inversion

The inversion of speech requires spectra of natural speech to be compared with spectra synthesized via the articulatory synthesizer. This comparison cannot be carried out directly because the source is not taken into account in the synthetic spectra. Furthermore, the mismatch between the speaker's vocal tract and the articulatory model, even after adaptation, introduces some deviation between the two spectra. The objective is to enable a direct comparison of cepstral vectors for both types of speech. Traditionally, a lifter is applied to weaken first coefficients linked to the spectra tilt and last coefficients linked to the source. However, this only provides a partial solution. Since we have at our disposal X-ray films and corresponding audio signals for one speaker we designed a true adaptation procedure.

First, we outlined contours of speech articulators either automatically or by hand in order to derive the vocal tract shape at any time point in the film. This enabled the synthesis of spectra, and thus cepstral vectors, along time. Then, an affine transform applied to each cepstral coefficient is found out by minimizing the distance between synthetic cepstral vectors, and those derived from natural speech. This adaptation procedure turns out to reduce the distance between natural and synthetic cepstral vectors very substantially.

6.1.1.3. Acoustic-to-articulatory inversion using a generative episodic memory

We have developed an episodic based inversion method. Episodic modeling is interesting for two reasons. First, it does not rely on any assumption about the mapping relationship between acoustic and articulatory, but rather it relies on real synchronized acoustic and articulatory data streams. Second, the memory structurally embeds the naturalness of the articulatory dynamics as speech segments (called episodes) instead of single observations as for the codebook based methods. Estimating the unknown articulatory trajectories from a particular acoustic signal, with an episodic memory, consists in finding the sequence of episodes, which acoustically best explains the input acoustic signal. We refer to such a memory as a concatenative memory (C-Mem) as the result is always expressed as a concatenation of episodes. Actually a C-Mem lacks from generalization capabilities as it contains only several examples of a given phoneme and fails to invert an acoustic signal, which is not similar to the ones it contains. However, if we look within each episode we can find local similarities between them. We proposed to take advantage of these local similarities to build a generative episodic memory (G-Mem) by creating inter-episodes transitions. The proposed G-Mem allows switching between episodes during the inversion according to their local similarities. Care is taken when building the G-Mem and specifically when defining the inter-episodes transitions in order to preserve the naturalness of the generated trajectories. Thus, contrary to a C-Mem the G-Mem is able to produce totally unseen trajectories according to the input acoustic signal and thus offers generalization capabilities. The method was implemented and evaluated on the MOCHA corpus, and on a corpus that we recorded using an AG500 articulograph. The results showed the effectiveness of the proposed G-Mem which significantly outperformed standard codebook and C-Mem based approaches. Moreover similar performances to those reported in the literature with recently proposed methods (mainly parametric) were reached.

6.1.2. Using Articulography for Speech production

Since we have at our disposal an articulograph (AG500, Carstens Medizinelektronik), we can easily acquire articulatory data required to study speech production. The articulograph is used to record the movement of the tongue (this technique is called electromagnetography - EMA). The AG500 has a very good time resolution (200Hz), which allows capturing all articulatory dynamics. The articulograph was used in a study about inversion (see the previous section) and to investigate pharyngealization.

Pharyngealized phonemes are commonly described as having the same place of articulation (dental) as their non-pharyngealized counterparts, but differ by the presence of a secondary articulation involving mainly the back of the tongue.

To study pharyngealized phonemes in Arabic from an articulatory point of view, our articulograph was used to record the movement of the tongue. Although EMA is not known as an optimal technique to cover the back of the tongue, good placement of the sensors and good interpretation of their positions can help to define pharyngealization relevantly. In fact, it is important to set one sensor as far as possible on the tongue (in our case, at 7cm from the tongue tip).

A corpus of several CVCVCVs was recorded using this articulograph, then phonetically labeled, and analyzed. The main finding of this work is that the coarticulation effect of the pharyngealized phonemes extends the immediate surrounding phonemes to influence the phonemes up to four-phoneme distance from the pharyngealized phoneme. The pharyngealization affects indifferently the previous and the following vowels and consonants.

6.1.3. Labial coarticulation

Results show that protrusion is a fragile cue to the rounding feature. Although we observe for each speaker a clear (but not large) separation between vowels /i/ and /y/ produced in isolation, many realizations of /i/ and /y/ come very close together and even overlap in few cases for vowels in contexts. The efficiency of the parameter depends on speakers and contexts. The distance between the corners is probably the most fragile cue to vowel roundedness. Many overlapping areas are observed for vowels in context. This is not good news for speech specialists since this parameter is easy to measure (with cameras and markers painted on the speaker's face) and its evaluation can be fully automatic. Each of the three lip opening parameters constitutes a very efficient cue to the rounding feature. For vertical opening, the opposition between /i/ and /y/ in initial position appears to be endangered in bilabial context, due to the anticipation of lip closing during /i/. Nevertheless, the temporal variations of lip opening during the initial /i/ are very important, and more analyses, taking into account these variations, will be necessary to analyse /i/ versus /y/ phonetic distinction more thoroughly.

6.1.4. Speech synthesis

6.1.4.1. Text-To-Speech

This year, we focused on the weighting of features to drive the selection step. As previously explained in 3.2.4, the selection is done from numerous linguistic features for the target cost and acoustic features for the overall cost. In order to add visual features (6.1.4.2), we normalized the weighting of linguistic features and the duration feature (introduced last year) and we added the derivative of the F0 in the concatenation cost. The system is now ready to add a visual part in the TTS process.

Moreover, we continued to study the concatenation of units. The algorithm developed last year to eliminate most of the large potential dephasings by choosing the right time point of concatenation has been refined. A correlation between potential instants of concatenation (two for one boundary: positive and negative major pitchmarks) is computed between the two units to concatenate (one correlation around the two positive marks of units and one around the two negative marks). The better couple was chosen. This method has been applied for all the concatenation instants and removes large and slight dephasings without any digital signal processing: the signal was not modified (excepted during one period - some milliseconds - by a classical slight smoothing).

As we will acquire a new corpus for the acoustic-visual TTS project (ANR ViSAC), we also refined our algorithm of selection of sentences to build the new corpus. Thanks to the experience of the first recording (in 2008), we have improved the selection of “good” sentences by a better filtering of input sentences. Indeed, the sentences need to be easily pronounced (in a comfortable manner) by the speaker to obtain a natural prosody. And as the analysis of over 3 million of sentences takes a lot of time to compute (about 400h), the program can now be launched on a cluster of computers. In particular, we used the cluster of the TALC CPER project.

6.1.4.2. Acoustic-Visual synthesis

At this stage of the project, we acquired a small corpus to develop and test our approach of acoustic-visual synthesis. Visual data acquisition was performed simultaneously with acoustic data recording, using an improved version of a low-cost 3D facial data acquisition infrastructure. The system uses two fast monochrome cameras, a PC, and painted markers, and provides a sufficiently fast acquisition rate to enable an efficient temporal tracking of 3D points. The recorded corpus consisted of the 3D positions of 252 markers covering the whole face. The lower part of the face was covered by 70% of all the markers (178 markers), where 52 markers were covering only the lips so as to enable a fine lip modeling. The corpus was made of 319 medium-sized French sentences uttered by a native male speaker and corresponding to about 25 minutes of speech.

We designed a first version of the text to acoustic-visual speech synthesis based on this corpus. The system uses bimodal diphones (an acoustic component and a visual one) and unit selection techniques (see 3.2.4). We have introduced visual features in the selection step of the TTS process. The result of the selection is the path in the lattice of candidates found in the Viterbi algorithm, which minimizes a weighted linear combination of three costs: the target cost, the acoustic joined cost, and the visual joined cost. The target and acoustic join costs are the ones used in the acoustic TTS process: they exploit linguistic and acoustic features (see 3.2.4). Similarly, the visual join cost is defined as the visual distance between the units to be concatenated [26], [25]. Currently, we are working in extending and modifying the target cost to include visual and quantitative articulatory information. Although the database used for this system is small, the preliminary results seem to be very promising.

We also studied a method to segment the visual signal (PCA trajectories). The segmentation was performed using an HMM-based forced alignment mechanism widely used in automatic speech recognition. The idea is based on the assumption that using visual speech data alone for the training might capture the uniqueness in the facial component of speech articulation, asynchrony (time lags) in visual and acoustic speech segments and significant coarticulation effects. This should provide valuable information that helps to show the extent to which a phoneme may affect surrounding phonemes visually [21].

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

6.1.5.1. Phonemic segmentation in reading and reading-related skills acquisition in dyslexic children and adolescents

This year, in the field of reading and reading related-skills identification in dyslexic children, our computerized tool EVALEC was published [33]. It is the first compurerized battery of tests in french language assessing reading and related skills (phonemic segmentation, phonological short term memory) comparing results both to chronological age controls and reading level age control in order to diagnostic Dyslexia. Both processing speed and accuracy scores are taken into account. This battery of tests is used by speech and langage therapists.

The goals of our longitudinal data in dyslexic adolescentswere twofold: 1) to assess dyslexics'reading and reading related skills compared to chronological age controls and 2) examine the predictors of their reading level from skills assessed at age 8 and 17. The results indicate 1) the persistance over time of reading deficits 2) that phonological skills largely explain reading success or failure; 3) the importance of processing time, included for the assessment of phonological reading-related skills. This research was supported by a grant from the French Ministery of Health (Contrat 17-02-001, 2002-2005).

6.1.5.2. Langage acquisition and langage disabilities (deaf chidren, dysphasic children)

Providing help for improving French language acquisition for hard of hearing (HOH) children or for children with language disabilities was another project: ADT (Action of Technological Developpement) Handicom. The originality of this project was to combine psycholinguistical and speech analyses research. Agnès Piquard-Kipffer established and coordinated a network including researchers, speech and language therapists, cued-speech coders, illustrator, graphic designer...New ways to learn to speak/read were developed. A collection of three digital books has been written by Agnès Piquard-Kipffer for both 2-6, 5-9, 8-12 year old children (kindergarten, 1-4th grade) to train speaking and reading acquisition regarding their relationship with speech perception and audio-visual speech perception. A web interface has been created in order to create others books for language impaired children. A series of three related studies (simple cases studies) were proposed to investigate the linguistical, audio-visual processing... presumed to contribute to language acquisition by deaf children. Publications are submitted.

6.1.6. Enhancement of esophageal voice

6.1.6.1. Detection of F0 in real-time for audio: 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 the 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 [15]. At the same time, we improved the voiced / unvoiced decision by using a clever majority vote algorithm electing the actual F0 index candidate. A second publication describing this new result was published at the ISCIT 2010 conference [14].

6.1.6.2. 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), the GMM parameters are determined by minimizing a mean squared distance between the transformed vectors and target vectors. In the second method (Kain), 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. This interesting result was published and presented at the ISIVC 2010 conference [27].

6.1.7. Perception and production of prosodic contours in L1 and L2

In the framework of the “Intonale” project (see 7.2.2), we have collected a corpus recorded by 34 French speakers and made up of sentences with different modalities: assertions, questions, major and minor continuations. French speakers uttered these sentences both in French (their native language) and in English (the “targeted” non native language). The corpus will be used to study the perception and the production of prosodic contours. One of the goals is to improve our language learning tools in order to propose more efficient visual and acoustic feedback to help the learner to be aware of the difference between the L1 and L2 prosody.

In order to analyse the corpus in terms of acoustic data, a tool has been developed. It proposes the visualization of INTSINT-MOMEL model (and computation of statistics). This well-known model gives a spline interpolation of the intonation contour and relative frequency targets. It should allow us to extract intonative information.

6.2. Automatic Speech Recognition

Participants: Christophe Cerisara, Sébastien Demange, Dominique Fohr, Christian Gillot, Jean-Paul Haton, Irina Illina, Denis Jouvet, David Langlois, Odile Mella, Kamel Smaïli, Frederick Stouten, Frédéric Tantini.

6.2.1. Core recognition platform

6.2.1.1. Broadcast News Transcription

In the framework of the Technolangue project ESTER, we have developed a complete system, named ANTS, for French broadcast news transcription (see section 5.7).

Extensions of the ANTS system have been studied, including the possibility to use the Sphinx recognizers http://cmusphinx.sourceforge.net/. Training scripts for building acoustic models for the Sphinx recognizers are now available and take benefit of the computer cluster for a rapid optimization of the model parameters. The Sphinx models are also used for speech/text alignment. A new speech decoding program has been developed for efficient decoding on the computer cluster, and easy modification of the decoding steps (speaker segmentation and clustering, data classification, speech decoding in one or several passes, ...). It handles both the Julius and Sphinx (versions 3 and 4) decoders.

This year, for acoustic features, we investigated PLP (Perceptual Linear Perception) coefficients, variance normalization speaker by speaker. We improved training of acoustic models by removing the badly recognized sentences from the training corpus. For recognition, extra speech events are better taken into account using specific models: one for breathing, two for noises, one for hesitation (euh in French), and one for pauses between words.

We are beginning a study about non native speech (mainly African radio) of Ester corpus. The main problems are non canonical pronunciations, prosody and specific vocabulary.

6.2.2. Robustness of speech recognition

Robustness of speech recognition to multiple sources of speech variability is one of the most difficult challenge that limits the development of speech recognition technologies. We are actively contributing to this area via the development of the following advanced approaches:

6.2.2.1. Spontaneous speech

Spontaneous speech is characterized by:

- insertions (hesitations, repetitions, pauses, false start...),
- pronunciation variations (word contractions),
- variations of the speed of speech,
- noisy background (superimposed speech, laughter...).

During the master internship of Panpan Zhang, we focused on word repetitions in spontaneous speech. As the language model of ANTS is mainly trained with newspaper articles, repetitions are not taken into account in our system. So repetitions give recognition errors. The analysis of the training corpus shows that repeated words are mainly function words (de, le, des, et, les...). The chosen approach is based on “multiword”: a new word is created by the concatenation of repeated words, for instance the creation of the multiword “de_de” for the repetition of “de”. These multiwords are added to the lexicon. Three methods for estimating the language model probabilities of these multiwords were evaluated. Best results were obtained by re-estimation of the language model using the modified training corpus where multiwords were included. A slight improvement was observed on the small test corpus.

6.2.2.2. Detection of Out-Of-Vocabulary words

One of the key problems for large vocabulary continuous speech recognition is the occurrence of speech segments that are not modeled by the knowledge sources of the system. An important type of such segments are so-called Out-Of-Vocabulary (OOV) words (words are not included in the lexicon of the recognizer). Mostly OOV words yield more than one error in the transcription result because the error can propagate due to the language model.

We have investigated, with Frederik Stouten (postdoctoral), to what extent OOV words can be detected. For this we used a classifier that makes a decision about each speech frame whether it belongs to an OOV word or not. Acoustic features for this classifier are derived from three recognition systems. The first one is a word recognizer constrained by the lexicon. This recognizer builds a word lattice which is used to calculate frame-based word posterior probabilities. The second system is a phone recognizer constrained by a grammar. This system was used for calculating approximations to the phoneme posteriors. The third system is a phoneme recognizer (a free phoneme loop) from which we extracted frame-based phoneme posterior probabilities. The difference between these probabilities is assumed to give an indication about speech frames that belong to words that are not included in the lexicon of the word recognizer. On top of the acoustic features we also used four language model features: the ngram probability, the order of the gram that was used to calculate the language model probability, the unigram probability for the current word and a binary indicator that takes the value one if the word is preceded by a first name.

We propose to exploit the fact that 38% of the OOV word observations in the broadcast news data are pronounced more than one time in a time period of less than 1 minute. To improve the detection of repeated OOV words, we design a clustering module working on the detected OOV word segments. This algorithm is based on the estimation of the entropy. The proposed incremental clustering algorithm has been evaluated on the broadcast news corpus ESTER and gave better performance than a classical baseline incremental clustering algorithm based on a distance threshold [28].

6.2.2.3. Pronunciation variants

Modeling pronunciation variation is an important topic for automatic speech recognition. It has been widely observed that speech recognition performance degrades notably on spontaneous speech, and more precisely, that the word error rate increases when the degree of spontaneity increases. The rate of speech is also an important variability source which impacts notably on the acoustic realization of the sounds as well as on the pronunciation of the words, and consequently affects recognition performance. Large increases in word error rates are observed when speaking rate increases. And, it should be noted that rate of speech and spontaneous speech are not completely independent as the rate of speech is an important cue for detecting spontaneous speech.

Consequently, we have investigated the modeling of the probabilities of pronunciation variants for large vocabulary continuous speech recognition, and evaluated it on broadcast news transcriptions. Because of the optional schwa (mute-e) and of the liaisons which are important pronunciation variants in French, there is, on average, more than 2 pronunciation variants per word. For words that are observed frequently enough on the training data, the probability of the pronunciation variants are usually estimated as the frequency of occurrences of the pronunciation variants on the training data. However many words are not observed on the training data. Thus generic phenomena (such as pronunciation or not of the optional schwa in a n-syllable word) were defined, and their frequency estimated from the training data. The frequency of these generic phenomena was used to derived an a priori probability for the pronunciation variants of the words, and finally, the probability of each pronunciation variant was determined through a MAP (maximum a posteriori) estimation using these a priori values and the actual frequency observed on the training data [20].

It was also observed that the frequency of pronunciation of the optional schwa depends on the speaking rate. To take this fact into account a variability dependent modeling of the probability of the pronunciation variants was proposed; where the variability considered was the speaking rate. Moreover, as the actual realization of the liaisons depends on the following word (roughly, whether it starts with a vowel or not), the following word contextual influence was also introduced in the modeling. This leads to a detailed modeling of the pronunciation variants (following word-context influence, speaking rate dependence), which was trained using the MAP approach described above, and provided a significant improvement on broadcast news transcription [20].

6.2.3. Speech/text alignment

Speech to text alignement is a research objective that is derived from speech recognition. While it seems easier to solve at first sight, expectations are also higher and new problems appear, such as how to handle very large audio documents, or how to handle out-of-vocabulary words. Another important challenge that motivated our work in this area concerns how to improve our results and meet the user expectation by exploiting as much as possible the interactions and feedback loop between the end-user and the system. We developed the open-source JTrans software platform for this task (see section 5.9). We continued our work in this area in collaboration with linguists from the ATILF laboratory, and we are currently extending this collaboration with the Paris 3 University on the one hand, and on the other hand with a company named "Timecode" (see section 7.4.2).

6.2.4. Computing and merging linguistic information on speech transcripts

The raw output of speech recognition is difficult to read for humans, and difficult to exploit for further automatic processing. We thus investigated solutions to enrich speech recognition outputs with non-lexical information, such as dialog acts, punctuation marks and syntactic dependencies. Computing such a linguistic information requires a corpus to train stochastic models, and we also worked out new semi-supervised training algorithms for building a French corpus dedicated to syntactic parsing of oral speech. The creation of this corpus is realized in collaboration with the TALARIS team. Finally, we designed a new solution to improve our core language models by integrating into them lexical semantic distances.

An important information for post-processing speech transcripts concerns dialog acts and punctuation marks. We initiated some work in this area several years ago with the PhD thesis of Pavel Kral. Since then, we continued our collaboration in this domain by successively investigating specific challenges, such as finding the most relevant features, models and testing the adaptation of our approaches in two languages, Czech and French [11]. We recently focused our efforts on recovering commas, which constitute the most difficult punctuation marks to detect. We further started integrating relevant syntactic information for this task, and we submitted a paper on this work.

Infering syntactic dependencies is an extremely important step towards structuring the text and an absolute prerequisite for working with relations between words and next interpreting the utterance. Yet, no state-of-the-art solutions designed for parsing written texts can be reliably adapted to parsing speech, and even less transcribed speech. The lack of such methods and resources is especially blatant in French. We started, in collaboration with the TALARIS team, to address this issue by building a new French treebank dedicated to speech parsing [16], as well as a software platform dedicated to working with this corpus (see section 5.8). We further investigated a new active learning approach to speed up the corpus creation process [23] and started to design a new semantic role labelling approach on top of this syntactic treebank [17].

While a large part of our work is dedicated to enriching the output of our speech recognition system, we also tried integrating within the speech decoding process itself new information coming from the higher levels. We proposed in this area to integrate lexical semantic distances in order to smooth and improve the estimation of the speech recognition language models [18].

6.3. Speech-to-Speech Translation and Langage Modeling

Participants: Kamel Smaïli, David Langlois, Caroline Lavecchia, Sylvain Raybaud.

This year, Caroline Lavecchia defended her PhD Thesis [8] on inter-lingual triggers for Machine Translation.

Moreover, we pursued our work on Confidence Measures for Machine Translation yet presented in [53], [54], [52]. We have synthesized our work in a journal paper which is currently submitted. In order to measure the impact of the Confidence Measures, we integrated them into a multilingual tool. To this end we implemented a post editing tool with confidence measures and let users correct machine translated sentences, with and without the help of confidence measures. This experiment requires an amount of financial and human resources beyond our team's capacity, therefore we had to do with a restricted number of volunteers during a restricted time span. The results are therefore more qualitative than quantitative. This experiment showed that confidence measures are not mature enough to be helpful in such a setting. However, the limited number of volunteers and the lack of long term observations make the results somewhat difficult to interpret. But the knowledge gained from this experiment and users feedback will help us improve confidence measures in a way that is really helpful for users.

Last year, we have started a work on multilingual summarization. In this work, we want to provide a translation of a document content. For that, we address in the same work the summarization field and the machine translation field. In order to prevent from the difficulty of producing a true syntactically correct summary of a document, we propose a graph representation of the document, and we propose a method to translate the nodes of the graph by taking into account the neighbours. Our first results encourage us to continue in this direction by defining a measure of the correctness of a graph versus the initial document. This work started during a research Master 2 training period. This year, we published a paper presenting our first results [13].

Moreover, we pursued our collaboration with Chiraz Latiri from the URPAH Team, University of Tunis. In this scope we want to compare two methods in order to build translation tables for Machine Translation: the inter-lingual approach from us, and an approach inspired from the association rules mining technique, well-known in data mining. In [12], we extend the both approaches to phrase based translation, and we combine them.

Last, in the scope of statistical language modeling for automatic speech recognition, we work with Christian Gillot and Christophe Cerisara [18] on a method for classifying histories.

7. Contracts and Grants with Industry

7.1. Introduction

Our policy in terms of technological and industrial partnership consists in favoring contracts that quite precisely fit our scientific objectives. We are currently involved in an ANRs project about audiovisual speech synthesis (VISAC), another about acoustic-to-articulatory inversion of speech (ARTIS), another about the processing of articulatory data (DOCVACIM), one Eurostar European project called Emospeech about spoken interfaces in serious games and virtual worlds, and one Interreg European project called Allegro about second language learning.

In addition, we are involved in several regional or industrial projects and bilateral cooperations.

7.2. Regional Actions

7.2.1. CPER MISN TALC

The team is involved in the management of the Contrat Plan E'tat-Région (CPER) contract. In particular, Christophe Cerisara is co-responsible, with Claire Gardent, of the CPER MISN TALC, whose objective is to leverage collaborations between regional academic and private partners in the domain of Natural Language Processing and Knowledge engineering. The TALC action involves about 12 research teams and 40 researchers for a budget of about 240,000 euros per year.

In addition to the co-management of this project, our team is also involved in three scientific collaborative operations:

- An operation about text-to-speech alignement, in collaboration with the TALARIS research team and the ATILF laboratory. This operation aims at proposing semi-supervised solutions to facilitate the transcription and processing of large bimodal text and speech corpora. The main outcomes of this operation are the JTRANS software described in section 5.9, and a concordancer that was developped in Java by two BSc students in the framework of their final year project.
- An operation about syntactic analysis of speech transcripts, in collaboration with the TALARIS research team and the ATILF laboratory. This operation aims at adapting state-of-the-art stochastic parsers to the specificities of manual and automatic transcriptions of speech, and at building a French treebank of broadcast news speech transcripts. The main outcome of this operation is the J-Safran software, described in section 5.8.
- An operation about speech translation, which aims at improving translation approaches thanks to triggers. The main outcome of this operation is a browsing and annotation toolkit for the creation of aligned multilingual subtitle corpus, available at http://talc.loria.fr/Subweb.html.

7.2.2. "Intonale": Perception and production of prosodic contours in L1 and L2

This action, launched by the CCOSL, aims at developping collaboration between academic partners from Lorraine laboratories and universities. It has started in September 2009 and should last until the end of 2010. The speech team from LORIA is associated with the laboratory ATILF (Mathilde Dargnat). The project deals with the perception and production of prosodic contours in the first language (L1) and in a second language (L2). We have chosen two radically different languages with respect to prosody: French and English. We have collected a corpus recorded by 34 French speakers and made up of sentences with different modalities: assertions, questions, major and minor continuations. French speakers uttered these sentences both in French (their native language) and in English (the "targeted" non native language). The English part of the corpus is used by the project ALLEGRO, presented hereafter. The French part of the corpus is currently segmented, whilst its English part is segmented under the framework of the INTERREG project ALLEGRO. A tool for prosody analysis, devoted to "Intonale", this action, has been developed by a master student.

7.3. National Contracts

7.3.1. ADT Handicom

An ADT (Action of Technological Development), was led from 2008 till 2010, managed by Agnès Piquard-Kipffer. The aim of this project is to provide help for improving French Language Acquisition for hard of hearing (HOH) chidren or for chidren with language disabilities.

A collection of three digital books has been written by Agnès Piquard-Kipffer and a web interface has been created in order to create others books for language impaired children.

A workflow which transforms a text and an audio source in a video of digital head has been developed. This workflow includes:

- An automatic speech alignment has been integrated. This process can retrieve from an acoustic signal and a text transcription, the length and the position of each phoneme and of each word. This allows a synchronization of the articulation of the head with acoustic signal and text display. This technology is a recognition engine, result of a previous work called ESPERE from EPI Parole.
- A phonetic transcription designed in the EPI Parole has been integrated and adapted.
- A speech synthetizer has been integrated. This technology can create an artificial voice from a text. It is a part of tools provided to make a digital book. Several software programs are tested in order to find the best result.
- A French cued speech coding and talking head has been improved in order to generate videos on a server. The animation consists in animating a 3D talking head, in association with a 3D hand which can code cued speech. This technology was created from a previous RIAM project called LABIAO.

A digital book written in FLASH has been developed. It integrates videos of the digital head, which are synchronized with texts displayed for each page. Digital books can be created manually with a text editor (to create XML file) or automatically with software which can be easily used to add all necessary multimedia elements in pages.

Data (audio source and text) are provided from a web interface. This web site allows users to create digital books. Through this interface, the books can be easily modified, shared and read. This website has been developed with Symfony (PHP 5 web framework) and AJAX (Dojo toolkit API) technologies. A linguistical study and a case study analysis of the current version of the talking head and of the digital books were conducted in collaboration with the Ecole d'Orthophonie of Nancy and with 6 students from the Speech Therapy School of Nancy (Ecole d'Orthophonie de Nancy : Floriane Jacques, Amélie Dumont, Sophie Bardin, Elodie Racine, Claire Nostrenoff and Anaïs Laurenceau.

7.3.2. ANR DOCVACIM

This contract, coordinated by Prof. Rudolph Sock from the Phonetic Institute of Strasbourg (IPS), addresses the exploitation of X-ray moving pictures recorded in Strasbourg in the eighties. Our contribution is the development of tools to process X-ray images in order to build articulatory models.

7.3.3. ANR ARTIS

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 more appropriate articulatory models of the tongue, the development of a lifter distance for cepstral data and the synthesis of speech signal from articulatory contours outlined from X-ray moving pictures.

7.3.4. ANR ViSAC

This ANR Jeunes Chercheurs started in 2009, in collaboration with Magrit group. The main purpose of ViSAC (Acoustic-Visual Speech Synthesis by Bimodal Unit Concatenation) is to propose a new approach of a text-to-acoustic-visual speech synthesis which is able to animate a 3D talking head and to provide the associated acoustic speech. The major originality of this work is to consider the speech signal as bimodal (composed of two channels acoustic and visual) "viewed" from either facet visual or acoustic. The key advantage is to guarantee that the redundancy of two facets of speech, acknowledged as determining perceptive factor, is preserved. An important expected result is a large bimodal speech corpus offering a high linguistic coverage which will be used to build the acoustic-visual speech synthesis system, and allows to study coarticulation in depth.

At this stage of the project, we designed a first version of the text to acoustic-visual speech synthesis based on this corpus. The system is using bimodal diphones (an acoustic component and a visual one) and it is using unit selection techniques. Although the database for the synthesis is small, however the first results seem to be very promising. For the next year, we will acquire a larger corpus of 2 to 3 hours of audiovisual speech. This will be used for the final synthesis system, as it should cover all the language characteristic needed by the synthesis method. This corpus will allow also refining our algorithms used for selection and concatenation.

7.4. Grants with Industry

7.4.1. Spinal Images

We have begun a collaboration with The Picture Factory about the indexation of rushes. The automatic transcription of French dialogs contained in the rushes would automatically allow the rush indexing.

In this framework, during the master internship of Panpan Zhang, we focused of word repetitions in spontaneous speech.

7.4.2. Timecode

We begin a collaboration with the Timecode company that works in dubbing (recording and replacing voices on a motion picture or television soundtrack). We want to use tools developped in our team to speed up the process of making a rythmo band (or "lip-sync band"). The band is actually a clear 35 mm film leader on which the dialogue is written, along with numerous additional indications for the actor (laughs, cries, length of syllables, mouth sounds, breaths, mouth openings and closings, etc.). The rythmo band is projected in the studio and scrolls in perfect synchronization with the picture. We have designed a tool for automatic alignment of the rythmo band and the audio file.

7.5. International Contracts

7.5.1. Allegro

Allegro is an Interreg project (in cooperation with the Department of COmputational LInguistics (COLI) 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 segmentating of a corpus made up of English sentences uttered by French speakers and we analyzed specific problems encountered by French speakers when speaking English.

7.5.2. EMOSPEECH

The Emospeech project is a Eurostar project started on 1st June 2010 in cooperation with SMEs Artefacto (France) and Acapela (Belgium). This project comes within the scope of serious games and virtual worlds. If existing solutions reach a satisfying level of 3D physical immersion, they do not provide satisfactory natural language interactions. The objective is thus to add spoken interactions via automatic speech recognition and speech synthesis. EPI Parole and Talaris take part in this project and the contribution of Parole will be about the interaction between the virtual world, automatic speech recognition and the dialogue management.

7.5.3. CMCU - Tunis University

This cooperation involves the LSTS (Laboratoire des systèmes et Traitement du Signal) of Tunis University headed by Prof. Noureddine Ellouze and Kais Ouni. This new project involves the investigation of automatic formant tracking, the modelling of peripheral auditory system and more generally speech analysis and parameterization that could be exploited in automatic speech recognition.

7.5.4. The Oesovox Project 2009-2011

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. In order to answer the INRIA Euromed 3+3 Mediterranean 2006 call, the INRIA Parole group (Joseph Di Martino, LORIA senior researcher, Laurent Pierron, INRIA engineer and Pierre Tricot, Associated Professor at INPL-ENSEM) associated with the following partners:

- **Spain**: Begoña Garcia Zapirain, Deusto University (Bilbao-Spain), Telecommunication Department, PAS-"ESOIMPROVE" research group.
- **Tunisia**: Sofia Ben Jebara, TECHTRA research group, SUP'COM, Tunis.
- **Morocco**: El Hassane Ibn-Elhaj, SIGNAL research group, INPT, Rabat.

This project named LARYNX has been subsidized by the INRIA Euromed program during the years 2006-2008. Our results have been presented during the INRIA 2008 Euromed colloquium (Sophia Antipolis, 9-10 October 2008). During this international meeting, the French INRIA institute decided to renew our project with the new name "OESOVOX". This new project will be subsidized during the years 2009-2011.

In the framework of the European COADVISE-FP7 program, two PhD students have been assigned to the Euromed 3+3 Oesovox project. These students are: Miss Fadoua Bahja from INPT-Rabat (Morocco) whose PhD thesis title is "Detection of F0 in real-time for audio: application to pathological voices" and Mr. Ammar Werghi from SUP'COM-Tunis (Tunisia) whose PhD thesis title is "Voice conversion techniques applied to pathological voice repair".

8. Dissemination

8.1. Animation of the scientific community

- The members of the team frequently review articles and papers for Journal of Phonetics, JASA, Acta Acoustica, CSL, Speech communication, TAL, IEEE Journal of Selected Topics in Signal Processing, IEEE Transaction of Information Theory, Signal Processing, Multimedia Tools, Pattern Recognition Letters, ICASSP, INTERSPEECH, EURASIP, JEP.
- Member of editorial boards :
  - Speech Communication (J.P. Haton, D. Jouvet)
  - Computer Speech and Language (J.P. Haton)
  - EURASIP Journal on audio, Speech, and Music Processing (Y. Laprie)
- Member of scientific commitee of conference :
  - TAIMA, SIIE (K. Smaïli)
- Chairman of French Science and Technology Association (J.P. Haton)
- Member of “Association Française pour la Communication Parlée” (French Association for Oral Communication) board (I. Illina)
- Member of the Lorraine network on specific language and Learning disabilities and in charge of the speech and language therapy expertise in the Meurthe-et-Moselle House of Handicap (MDPH) (A. Kipffer-Piquard)
- The members of the team have been invited as lecturer:

  - Anne Bonneau, Vincent Colotte & Yves Laprie; “Tools devoted to the acquisition of the oral of a second language” at Department computational linguistics and Phonetics of the of Saarland University on July 2.
  - Slim Ouni; “Tête parlante : outils pour l’étude de la parole audiovisuelle” at Institut de Phonétique de Strasbourg.
  - Slim Ouni; “Parole Audiovisuelle” at Master of Cognitive Sciences of Nancy.
  - Anne Bonneau, Workshop «Natal» Natural language processing and Computer aided Language Learning, LORIA
  - Anne Bonneau, Workshop organised by “E’cole d’orthophonie de Nancy”
  - Agnès Piquard-Kipffer, Finnish Center of Excellence in Learning and Motivation - (Finlande, Jyväskylä).
  - Agnès Piquard-Kipffer, EHESP, E’cole des Hautes E’tudes de Santé Publique - Rennes, Sorbonne Paris Cité Université.

- For the second time, the team organised a Statistical Machine Translation Day. We invited four researchers in our domain from main French-speaking laboratories : LIMSI (Paris), LINA (Nantes), LIUM (Le Mans), LIG (Grenoble). During this day, we presented our work to the community. Students of the Erasmus Mundus Master were present during this day. See http://wikitalc.loria.fr/dokuwiki/doku.php?id=operations:journee_traduction_2010 for details and slides.

8.2. Invited lectures

- Catherine Pelachaud (LTCI, Telecom ParisTech), IPAC seminar.
- Robert Ladd Professor of Linguistics, Head of School of Philosophy, Psychology, and Language Sciences (PPLS) at the University of Edimburg.
- François Yvon (LIMSI, Univ. Parix XI), Translation seminar, “n-gram based Statistical Machine Translation: principles and some recent improvements”
- Emmanuel Morin (LINA), Translation seminar, “Bilingual Lexicon Extraction from Comparable Corpora”
- Yannick Estève (LIUM, Univ. of Maine), Translation seminar, “Coupling ASR and SMT systems for automatic speech translation”
- E’ric Gaussier (LIG, Univ. Joseph Fourier), Translation seminar, "Improving Corpus Comparability for Multilingual Access”
- Hugo Van Hamme (Katholieke Universiteit Leuven), TALC "Vocabulary acquisition by machines”

8.3. Higher education

- A strong involvement of the team members in education and administration (University Henri Poincaré, University Nancy 2, INPL): Master of Computer Science, IUT, MIAGE, Speech and Language Therapy School of Nancy,
- Coordinator of C2i (Certificat Informatique et Internet) at Henri Poincaré University (V. Colotte),
- Head of MIAGE Maroc (students of University Nancy 2 but having their courses in Morocco) (K. Smaïli),
- Head of Networking Speciality of University Henri Poincaré Master of Computer Science until 1st September (O. Mella),
- co-Director of DU, « Troubles du Langage et des Apprentissages », Université de Nancy 1, Faculté de Médecine (Agnès Piquard-Kipffer).

8.4. Participation to workshops and PhD thesis committees

- Members of Phd thesis committees I. Illina, D. Fohr, J.-P. Haton, M.-C. Haton, Y. Laprie, K. Smaïli, D. Jouvet, C. Cerisara, D. Langlois;
- Members of HDR committees J.-P. Haton, D. Jouvet;
- All the members of the team have participated to workshops and have given talks.

9. 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, p. 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, p. 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, p. 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), p. 444–460, PACS numbers: 43.70.h, 43.70.Bk, 43.70.Aj [DOS], http://hal.archives-ouvertes.fr/hal-00008682/en/.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[8] C. Lavecchia. *Les Triggers Inter-langues pour la Traduction Automatique Statistique*, Université Nancy II, June 2010, http://hal.inria.fr/tel-00545463/en.

Articles in International Peer-Reviewed Journal

[9] E. Didiot, I. Illina, D. Fohr, O. Mella. *A wavelet-based parameterization for speech/music discrimination*, in "Computer Speech & Language / Computer Speech and Language", April 2010, vol. 24, n^{o} 2, p. 341-357, http://hal.inria.fr/hal-00435076/en.

[10] I. Jemaa, K. Ouni, Y. Laprie. *Evaluation of Automatic Formant Tracking Method Using Fourier Ridges*, in "Cognitive Computation", 2010, vol. 2, p. 170-179, http://hal.inria.fr/inria-00544221/en.

[11] P. KRAL, C. CERISARA. *Dialogue act recognition approaches*, in "Computing And Informatics", 2010, vol. 29, n^{o} 2, p. 227–250, http://hal.inria.fr/inria-00431396/en.

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

International Peer-Reviewed Conference/Proceedings

[13] R. ANDRÉ-LOVICHI, K. SMAÏLI, D. LANGLOIS. *Utilisation de graphes sémantiques pour l'extraction et la traduction des idées essentielles d'un texte*, in "Extraction et Gestion des Connaissances (EGC)", Tunisia Hammamet, February 2010, http://hal.inria.fr/inria-00545488/en.

[14] F. BAHJA, J. DI MARTINO, E. H. IBN ELHAJ, D. ABOUTAJDINE. *An improvement of the eCATE algorithm for F0 detection*, in "10th International Symposium on Communications and Information Technologies - ISCIT 2010", Japan Tokyo, 2010, http://hal.inria.fr/inria-00545441/en.

[15] 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", Morocco Rabat, 2010, http://hal.inria.fr/inria-00545435/en.

[16] C. CERISARA, C. GARDENT, C. ANDERSON. *Building and Exploiting a Dependency Treebank for French Radio Broadcast*, in "TLT9 – the ninth international workshop on Treebanks and Linguistic Theories", Estonia Tartu, November 2010, http://hal.inria.fr/inria-00537147/en.

[17] C. GARDENT, C. CERISARA. *Semi-Automatic Propbanking for French*, in "TLT9 - The Ninth International Workshop on Treebanks and Linguistic Theories", Estonia Tartu, November 2010, http://hal.inria.fr/inria-00537148/en.

[18] C. GILLOT, C. CERISARA, D. LANGLOIS, J.-P. HATON. *Similar N-Gram Language Model*, in "INTERSPEECH 2010", Japan Tokyo, September 2010, p. 1824-1827, http://hal.inria.fr/inria-00540428/en.

[19] I. JEMAA, O. REKHIS, K. OUNI, Y. LAPRIE. *Evaluation d'une nouvelle méthode de suivi de formants sur un corpus Arabe*, in "XXVIIIèmes Journées d'Etude sur la Parole - JEP'10", Belgium Mons, May 2010, http://hal.inria.fr/inria-00544361/en.

[20] D. JOUVET, D. FOHR, I. ILLINA. *Detailed pronunciation variant modeling for speech transcription*, in "INTERSPEECH", Japan Makuhari, ISCA, September 2010, http://hal.inria.fr/inria-00528225/en.

[21] U. MUSTI, A. TOUTIOS, S. OUNI, V. COLOTTE, B. WROBEL-DAUTCOURT, M.-O. BERGER. *HMM-based Automatic Visual Speech Segmentation Using Facial Data*, in "Interspeech 2010", Japan Makuhari, Chiba, ISCA, September 2010, p. 1401-1404, http://hal.inria.fr/inria-00526776/en.

[22] B. POTARD, Y. LAPRIE. *Automatic adaptation of a vocal tract model*, in "Proceedings of the 18th European Signal Processing Conference - EUSIPCO-2010", Denmark Aalborg, August 2010, http://hal.inria.fr/inria-00544363/en.

[23] F. TANTINI, C. CERISARA, C. GARDENT. *Memory-Based Active Learning for French Broadcast News*, in "INTERSPEECH 2010", Japan Tokyo, September 2010, p. 1377-1380, http://hal.inria.fr/inria-00540423/en.

[24] F. Tantini, A. Terlutte, F. Torre. *Sequences Classification by Least General Generalisations*, in "10th International Colloquium on Grammatical Inference", Spain Valencia, J. M. Sempere, P. Garcia (editors), Lecture Notes in Artificial Intelligence, Springer, September 2010, vol. 6339, p. 189-202, The original publication is available at www.springerlink.com [*DOI :* 10.1007/978-3-642-15488-1_16], http://www.springerlink.com/content/e6270247p5048l21/, http://hal.inria.fr/inria-00524707/en.

[25] A. Toutios, U. Musti, S. Ouni, V. Colotte, B. Wrobel-Dautcourt, M.-O. Berger. *Setup for Acoustic-Visual Speech Synthesis by Concatenating Bimodal Units*, in "Interspeech 2010", Japan Makuhari, Chiba, ISCA, September 2010, p. 486-489, http://hal.inria.fr/inria-00526766/en.

[26] A. Toutios, U. Musti, S. Ouni, V. Colotte, B. Wrobel-Dautcourt, M.-O. Berger. *Towards a True Acoustic-Visual Speech Synthesis*, in "9th International Conference on Auditory-Visual Speech Processing - AVSP2010", Japan Hakone, Kanagawa, September 2010, p. POS1-8, http://hal.inria.fr/inria-00526782/en.

[27] A. Werghi, J. Di Martino, S. Ben Jebara. *On the Use of an Iterative Estimation of Continuous Probabilistic Transforms for Voice Conversion*, in "5th International Symposium on I/V Communications over fixed and Mobile Networks - ISIVC 2010", Morocco Rabat, 2010, http://hal.inria.fr/inria-00545428/en.

National Peer-Reviewed Conference/Proceedings

[28] F. Stouten, I. Illina, D. Fohr. *Regroupement des occurrences des mots hors-vocabulaire répétés en vue de leur modélisation pour la transcription d'émissions radio*, in "28ème Journées d'étude sur la parole - JEP'10", Belgium Mons, Université de Mons, May 2010, http://hal.inria.fr/inria-00544140/en.

Scientific Books (or Scientific Book chapters)

[29] Y. Laprie. *Inversion acoustique articulatoire*, in "Le livre blanc de l'acoustique en France en 2010", SFA (Société Française d'Acoustique), December 2010, p. 91, http://hal.inria.fr/inria-00545066/en.

[30] S. Sidhom, K. Smaïli, M. Ghenima. , I. Tunis (editor)*Systèmes d'information et Intelligence économique (SIIE'2010)*, IHE Tunis, February 2010, vol. 1, Dépôt de la notice bibliographique seulement., http://hal.inria.fr/inria-00549757/en.

Patents and standards

[31] J. Di Martino, L. Pierron. *Synthétiseur numérique audio amélioré*, 2010-06-25, n^{O} 10/02674, Oesovox, http://hal.inria.fr/inria-00546967/en.

Other Publications

[32] A. Piquard-Kipffer, D. Lelarge, L. Pierron, F. Monnay. *Création de livres numériques pour enfants présentant des troubles du langage*, September 2010, http://hal.inria.fr/inria-00545856/en.

[33] L. Sprenger-Charolles, P. Colé, A. Piquard-Kipffer, G. Leloup. *EVALEC, Batterie informatisée d'évaluation diagnostique des troubles spécifiques d'apprentissage de la lecture.*, 2010, http://hal.inria.fr/inria-00545950/en.

References in notes

[34] 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^{O} 4, p. 85–89.

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

[36] 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, p. 201–206.

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

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

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

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

[41] 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, I.: Computing Methodologies/I.2: ARTIFICIAL INTELLIGENCE, I.: Computing Methodologies/I.5: PATTERN RECOGNITION, http://hal.inria.fr/inria-00105908/en/.

[42] 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, Ouvrage disponible à l'ANRT : http://www.anrtheses.com.fr/ Nom de l'auteur : Agnès Piquard-Kipffer. Reproduction de la thèse de Linguistique soutenue à l'Université de Paris 7 - Denis Diderot., http://hal.inria.fr/inria-00185312/en/.

[43] 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/.

[44] 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.

[45] 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, p. 115-224.

[46] J. KUPIEC. *Robust part-of-speech tagging using a hidden markov model*, in "Computer Speech and Language", 1992, vol. 6, p. pp. 225–242.

[47] Y. LAPRIE. *A concurrent curve strategy for formant tracking*, in "Proc. Int. Conf. on Spoken Language Processing, ICSLP", Jegu, Korea, October 2004.

[48] 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/.

[49] 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/.

[50] S. MAEDA. *Un modèle articulatoire de la langue avec des composantes linéaires*, in "Actes 10èmes Journées d'Etude sur la Parole", Grenoble, Mai 1979, p. 152-162.

[51] 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, p. 440–447.

[52] S. RAYBAUD, D. LANGLOIS, K. SMAÏLI. *Efficient Combination of Confidence Measures for Machine Translation*, in "10th Annual Conference of the International Speech Communication Association - INTERSPEECH 2009", Royaume-Uni Brighton, 2009, http://hal.inria.fr/inria-00417546/en/.

[53] S. RAYBAUD, C. LAVECCHIA, D. LANGLOIS, K. SMAÏLI. *New Confidence Measures for Statistical Machine Translation*, in "International Conference On Agents and Artificial Intelligence - ICAART 09", Portugal Porto, 2009, http://hal.inria.fr/inria-00333843/en/.

[54] S. RAYBAUD, C. LAVECCHIA, D. LANGLOIS, K. SMAÏLI. *Word- and sentence-level confidence measures for machine translation*, in "13th Annual Meeting of the European Association for Machine Translation - EAMT 09", Espagne Barcelona, 2009, http://hal.inria.fr/inria-00417541/en/.

[55] L. SPRENGER-CHAROLLES, P. COLÉ, D. BÉCHENNEC, A. KIPFFER-PIQUARD. *French normative data on reading and related skills from EVALEC, a new computerized battery of tests (end Grade 1, Grade 2, Grade 3, and Grade 4)*, in "Revue Européenne de Psychologie Appliquée", 2005, p. 157-186, http://hal.inria.fr/inria-00184979/en/.