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

2009

Table of contents

PARODLE

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 Scientist

Yves Laprie [Team Leader, Research Director CNRS, HdR]
Anne Bonneau [Research scientist CNRS]
Christophe Cerisara [Research scientist CNRS]
Dominique Fohr [Research scientist CNRS]
Denis Jouvet [Research Director INRIA, since 1st September 2009, HdR]

Faculty Member

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, HdR]
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, INPL]

Technical Staff

Fabian Monnay [INRIA]

PhD Student

Christian Gillot [MENRT grant, thesis to be defended in 2011]
Caroline Lavecchia [ATER I.U.T Charlemagne, Nancy 2 University, thesis to be defended in 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]
Farid Feïz [CNRS grant (ASPI contract), thesis to be defended in 2010]
Imen Jemaa [Joint supervision with ENIT Tunis, since April 2009]
Nadia Amar [Joint supervision with ENIT Tunis, thesis to be defended in 2011]
Alexander Sepulveda-Sepulveda [National University of Colombia, invited from January to August]

Post-Doctoral Fellow

Asterios Toutios [University Nancy 2, since 1st November 2009]
Frederik Stouten [INRIA]
Frédéric Tantini [since October, 1st 2009]

Administrative Assistant

Martine Kuhlmann [CNRS]

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.
- **Automatic speech recognition.** Our works are concerned with stochastic models (HMM[1], bayesian networks and missing data models), adaptation of a recognition system to a new or non-native speaker, to the communication channel or the environment, and with language models. 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 7.2.3), 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 [50], [51] 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 [61].

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 allows 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 [59].

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 for instance. Indeed at present, only vowels and sequences of vowels can be inverted, and only some attempts concern fricatives sounds. Moreover, most of the inversion techniques use formant frequencies as input data although formants cannot be extracted from speech easily and reliably.

3.2.3. Strategies of labial coarticulation

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

In the long term, our goal is to determine a method of prediction of labial coarticulation adaptable to a virtual speaker. Predicting labial coarticulation is a difficult problem that gave rise to many studies and models. To predict the anticipatory coarticulation gestures (see [46] for an overall presentation of labial coarticulation), three main models have been proposed: the look-ahead model, the time-locked model and the hybrid model.

These models were often compared on their performance in the case of the prediction of anticipation protrusion in VCV or VCCV sequences where the first vowel is unrounded, the consonant(s) is neutral with respect to labial articulation and the last vowel is rounded. There is no general agreement about the efficiency of these models. More recent models have been developed. The one of Abry and Lallouache [36] advocates for the theory of expansion movements: the movement tends to be anticipated when no phonological constraint is imposed on labiality. Cohen and Massaro [42] 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 [41] 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 [43] 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, we introduced a chunker (shallow parser) to give us information on a potential rhythmic structure. Moreover, to perform the selection, we developed an algorithm to automatically weight the linguistic features given by the NLP. Our method relies on acoustic clustering and entropy information [43]. 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 progressivelly made our research works evolve towards stochastic approaches, both for acoustic and language models. Regarding acoustic models, we have especially investigated HMM (Hidden Markov Models), STM (Stochastic Trajectory Models), multi-band approach and BN (Bayesian Networks). Regarding language models, our main interest has concerned ngram approaches (word classes, trigger, impossible ngram, etc).

The main challenge of automatic speech recognition is its robustness to multiple sources of speech variability [48]. Among them, we have been focusing on acoustic environment, inter- and intra-speaker variability, different speaking styles (prepared speech, spontaneous, etc.) and non-native pronunciations.

Another specifity of automatic speech recognition is the necessity to combine efficiently all the research works (in acoustic modeling, langage modeling, speaker adaptation, etc.) into a core platform in order to evaluate them.

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 developped 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

The core 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 (improvements of Parallel Model Combination, such as Jacobian adaptation). 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, missing data recognition, denoising and integration of high-level and contextual information to constraint speech decoding. Missing data recognition is a natural extension of the multi-band paradigm that we have chosen to develop. It is based on the assumption that speech and noise are separable in the feature domain, so that noise contribution can be marginalized out during decoding. The main challenge is then to accurately locate and separate both signal sources. A dual paradigm to handle speech variabilities is to acknowledge that part of the acoustic information will always be missing or corrupted, which implies to exploit 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 speech team has been working on this domain for a long time.

3.4. Speech to Speech Translation and Langage 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 [39] 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 [58]. 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, deals with linguistic units which consists in more than one word. The model supporting phrase-based machine translation uses reordering concept and additional feature functions. In order to retrieve phrases, several approaches have been proposed in the litterature. Most of them require word-based alignments. For example, Och and al. [60] collected all phrase pairs that were consistent with the word alignment provided by Brown's models.
We developed a phrase based algorithm which is based on finding first an adequate list of phrases. Then, we find out the best corresponding translations by using our concept of inter-lingual triggers. A list of the best translations of sequences is then selected by using simulated annealing algorithm.

3.4.3. Language model

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

3.4.4. Decoding

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

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

The international community uses either PHARAOH [53] or MOSES [52] 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 [47] 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. Software

5.1.1. WinSnoori

Snorri 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 Snorri 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 [55]. 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.

5.1.2. LABCORP

contacts : David Langlois (langlois@loria.fr) and Kamel Smaïli (smaili@loria.fr).

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 set 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 [54]. 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 5.1.11).

5.1.3. Automatic lexical clustering

contacts : David Langlois (langlois@loria.fr) and Kamel Smaïli (smaili@loria.fr).

In order to adapt language models in ASR applications, we use a toolkit we developed in 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.

5.1.4. SUBWEB

contacts : David Langlois (langlois@loria.fr) and Kamel Smaïli (smaili@loria.fr).

We published in 2007 a method which allows to align sub-titles comparable copora [57]. 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 and a graphical representation of the course of the algorithm. This work has been supported by CPER/TALC/SUBWEB[2].

5.1.5. ESPERE

contact : Dominique Fohr (fohr@loria.fr).

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.

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

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.

5.1.6. SELORIA

contact : Odile Mella (Odile.Mella@loria.fr).

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 results 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.

5.1.7. ANTS

contact : Dominique Fohr (fohr@loria.fr).

The aim of the Automatic News Transcription System (ANTS) is to transcribe radio broadcast news. ANTS is composed of four stages: broad-band/narrow-band speech segmentation, speech/music classification, detection of silences and 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.

5.1.8. JSynATS

contact : Christophe Cerisara (Christophe.Cerisara@loria.fr).

JSynATS is the “Java Syntactic parser of Automatically Transcribed Speech”. 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. It departs from the other existing French syntactic parsers from the fact that it aims at efficiently handling transcription errors produced by every automatic transcription systems. JSynATS will participate to the Passage evaluation campaign in november 2009.

5.1.9. JTRANS

contact : Christophe Cerisara (Christophe.Cerisara@loria.fr).

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, but an English version may be released in the future.

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

5.1.10. STARAP

contact : Dominique Fohr (fohr@loria.fr).

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.

5.1.11. TTS SoJA

contact : Vincent Colotte (Vincent.Colotte@loria.fr).

TTS SoJA (Speech synthesis platform in Java) is a software of text-to-speech synthesis 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 algorithm. 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 launch with a graphical user interface or directly integrated in a Java code or by following the client-server paradigm.

The software license should easily allow associations of impaired people to use the software.

6. New Results

6.1. Speech Analysis and Synthesis

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

6.1.1. Acoustic-to-articulatory inversion

Our approach of acoustic-to-articulatory inversion is an advanced table lookup method. The table is built by synthesizing speech spectra from a set of articulatory configurations generated by the articulatory model, which thus plays an important role. The current articulatory model is that designed by Maeda [59] by using a semi-polar grid. Some articulatory configurations, especially those corresponding to back vowels, are not well interpolated since the tongue contour does not intersect all the grid lines. We thus evaluated new strategies by using an adaptive grid which is attached to the jaw and compared it to the standard semi-polar grid [23] The main advantage is to get a better deformation model of the tongue in the front part of the mouth part. We also substantially improved the base of articulatory contours outlined from X-ray images recorded at the IPS laboratory in the eighties by using MRI images recorded in the framework of the ASPI project by the same speaker.

Our objective is to elaborate inversion algorithms that work on standard spectra data, i.e. cepstral vectors for instance, in real time. Despite the great theoretical interest of the codebook approach, mainly the possibility of potentially exploring the entire articulatory space, it is hard to imagine using it in a real time context. We thus developed a new iterative approach using a neutral articulatory trajectory which is then deformed in order to account for formant trajectories extracted from speech. The algorithm is inspired by the variational approach we previously designed to improve the acoustic proximity with data extracted from natural speech. The main strong point is that the convergence is guarantied even if the initial neutral curve is not in the vicinity of the expected inverse trajectory[27]. Even if the current version of the algorithm does not reach real time since it requires the derivation of the articulatory to acoustic synthesis we envisage to use the articulatory codebook as a fast computation of the synthesis.

Using standard spectra data as input of inversion requires a distance able to compare synthetic speech spectra together with natural speech spectra. This distance thus should minimize distorsions due to the influence of the source since synthetic spectra do not invole the source characteristics. The idea investigated was to design a lifter (i.e. the filtering of the cepstral coefficients) which minimizes a perceptual distance based on formants (spectral prominences affiliated to resonance cavities of the vocal tract) between a corpus of formant data extracted from natural speech and entries of the articulatory codebook. The optimal lifter has been derived by minimizing the average perceptual distance. Preliminaries results are very encouraging and this new distance will be exploited in our inversion framework.

6.1.2. Using Articulography for Speech production

The recent purchase of the articulograph AG500 allowed acquiring almost unlimited quantity of articulatory data and thus several speech production studies are possible. Electromagnetic articulography (EMA) is a current technique to record articulatory data with a very good temporal resolution as movement signals are sampled at 200 Hz. This allows capturing very fine speech movement. The system uses 12 sensors that can be glued on the tongue and lips for instance.

- **Mapping EMA data to an articulatory model.** Acoustic-to-articulatory maps based on articulatory models have typically been evaluated in terms of acoustic accuracy, that is, the distance between mapped and observed acoustic parameters. Since last year we have been developing a method that allows the evaluation of such maps in the articulatory domain. The proposed method estimates the parameters of Maeda's articulatory model on the basis of electromagnetic articulograph data, thus producing full midsagittal views of the vocal tract from the positions of a limited number of sensors attached to articulators. The match between the EMA data and the articulatory model is good. However, some improvements need to be done to take into account the larynx position (which cannot be covered by EMA). This method will allow a direct comparison of articulatory trajectories derived by inversion against those corresponding to the actual vocal tract dynamics, as recorded by EMA [64].

- **Studying pharyngealization using EMA.** Pharyngealization is an important characteristic of a set of consonants in Arabic and it has an important coarticulation effects on the neighboring vowels. Studying articulatory aspects of pharyngealization is currently accessible using articulography. One way to study the coarticulation effect of pharyngealization is to compare the dynamics of the articulation of sequences containing pharyngealized phonemes with similar sequences containing their non-pharyngealized cognates. We highlighted the differences between pharyngealized phonemes and non-pharyngealized ones in addition to pharyngeals vs. pharyngealized phonemes. The articulation of the tongue was tracked by four sensors glued on the tongue. A corpus of Arabic words uttered by a male speaker was recorded using AG500, labeled and analyzed. The main finding of this work is that the secondary articulation of moving the tongue back can be observed, while the main articulation of the tongue is the forward movement toward alveolar and dental positions. The dynamics observation showed that in pharyngealized context a backing of the tongue starts earlier than the production of the pharyngealized phoneme or pharyngeal. This anticipatory coarticulation is in accordance with earlier studies. We also showed that the phoneme context has an influence on backing of the tongue during the articulation of pharyngealized phonemes and there exists a mutual influence between pharyngealized phonemes and pharyngeals [18].

6.1.3. Labial coarticulation

We investigated the effect of “adverse” contexts, especially that of the consonant /ʃ/ and the “transconsonantal” vowel /i/, on labial parameters for French /i/ or /y/. Five parameters were analysed: the height, width and area of lip opening, the distance between the corners of the mouth, as well as lip protrusion. Ten speakers uttered a corpus made up of isolated vowels, syllables and logatoms. A special procedure has been designed to evaluate lip opening contours.

Results showed that the carry-over effect of the consonant /ʃ/ can have drastic consequences on lip protrusion for /i/, impeding, for about half of the speakers involved in this study, the distinction between /i/ and /y/ in this dimension. The (labial) opposition between these vowels was nevertheless ensured by other labial parameters, such as the height of lip opening. Results also put in light the existence of large variations among speakers’ coarticulatory habits. More experiments appeared necessary to find out the various influences of consonantal contexts and transconsonantal vowels on visual cues for vowels [21].

6.1.4. Speech synthesis

6.1.4.1. Text-To-Speech

This year, the development of the software platform TTS SoJA (Synthesis platfOrm in JAva) has continued (after 2 years of Associate Engineer INRIA grant, 2006-2008). Some corrections and improvements were made (more numerous than expected) what explains the delay of the expected official release.

Meanwhile, we have studied and proposed two improvements of our synthesis method. As previously explained in 3.2.4, the originality of our approach is that the selection is made from linguistic features without using a prosodic model. With a prosodic model, we can constrain the selection to be sure that the result has a good prosodic behavior. For our approach, the lack of strong constrain is an advantage for prosodic variability, but occasionally, can be a drawback because the selection can chose an unit without the right prosodic features (for instance the unit is too short for its position in the sentence). Duration is important in French, because the perception of stress is partly expressed by the lengthening of the last syllable of a word. A “mistake” in the length of a selected unit can completely disrupt the word from its accentuation and give an unnatural effect to the whole sentence. To constrain target selection slightly, we proposed to penalize a unit during the selection by comparing its length with the distribution of the duration of the phoneme (according to several significant positions in the sentence). The computation of this distribution was made on the same corpus to reflect the prosody (for the duration) of the speaker and not that of a standard model. The incorporation of this new feature gives good preliminary results.

The second improvement deals with the concatenation of units. The concatenation is a slight smoothing by standard OverLap and Add technique (pitch-synchronized). To avoid the complexity of GCI (Glottal Closure Instant) algorithms, we chose to put pitch marks on important negative or positive peaks of the speech signal [44]. Unfortunately, the positions of these marks are not consistent from one period to another period somewhere else in the audio corpus. This can result in a large dephasing at the time of unit concatenation, and consequently in acoustic artifacts (perceived as a glottal sound) in the signal obtained. Experimentally, this dephasing occurs either with positive peaks or with negative peaks but not with these both kinds of peaks for the same period. To eliminate the large potential dephasings, we have proposed to slightly change the concatenation by adding an algorithm based on the computation of correlations to choose the right kind of peak (positive or negative). This method removes large dephasings and the effects of slight remaining dephasings disappear with the OLA smoothing.

6.1.4.2. Acoustic-Visual synthesis

This year, we started our work on acoustic-visual speech synthesis within the framework of the ANR Jeunes Chercheurs ViSAC. This new challenge is a natural extension of our work from purely acoustic speech synthesis to acoustic-visual speech synthesis. In addition, this allows having a tied link with our ongoing work on speech production. Our main goal is to develop an acoustic-visual speech synthesis system using bimodal unit concatenation. This is a new approach of a text-to-acoustic-visual speech synthesis, which allows animating a 3D talking head and providing 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. We keep this association during the different processes. The key advantage is to guarantee that the redundancy of two facets of speech, acknowledged as determining perceptive factor, is preserved. As this work is done in collaboration with the Magrit team, they started by setting up the acquisition system of acoustic and stereo-visual data (the main challenge is to perform a real-time acquisition and the synchronization of the acoustic and the video entries). The Magrit team processed the visual data to provide 3D data. We performed some testing sessions to verify the quality of the recording to continue to tune up the content of the corpus, and the recording conditions (synchronization, noise, etc.). We started also studying the visual alignment of the corpus and how to keep it linked with the acoustic alignment for which existing alignment methods were used. We expect to finish developing a visual alignment algorithm during first quarter of next year.

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

The evaluation of phonemic discrimination has been based on the test specially made by [50] for her longitudinal study. 36 pairs of pseudowords, similar or different were presented to the child who must say if he heard the same item or not.

Concerning dyslexia and normal acquisition of reading, a group study has been conducted. The 85 children of our population (age 5.6) were separated in a group "at risk" for dyslexia (39 children) and a control group (45 children). The results have been analysed to characterize the performance pattern of these subjects, as a group. Three different types of oppositions have been examined (voicing, place of articulation, intervertions and insertions). Statistical analyses have been conducted. Publications are submitted.

Concerning dyslexia, a multiple case study has been conducted in collaboration with the CNRS (Paris-Descartes University, Savoie University and University Hospital Paris-Bicêtre). The results indicates that the deficit of phonemic awareness is more prevalent than the deficit in short term memory or in rapid naming in the 15 french-speaking dyslexics than to those of reading level controls. This research was supported by a grant from the ACI ’Cognitique’ (COG 129, French Ministry of Research). Publication is in revision [62].

For dysphasia, a multiple case study has been started in September 2007. 3 dysphasic children will be tested, matched with 3 children who are simply retarded in reading. A speech and language therapist student, Margaud Martin, is working on this project.

6.1.6. Enhancement of esophageal voice

6.1.6.1. Detection of F0 in real-time for audio: application to pathological voices

The PhD thesis subject of Fadoua Bahja is: "Detection of F0 in real-time for audio: application to pathological voices". To achieve this goal, the first step has consisted in optimizing the CATE algorithm developed by Joseph Di Martino and Yves Laprie [45]. 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. At the beginning, we studied the different steps implemented in the CATE algorithm. Then, we tuned new parameters and made tests on various thresholds in order to find the most pertinent. Finally, we compared the results obtained with the eSRPD method (Enhanced Super Resolution Pitch Determination) developed by P. Bagshaw in 1993. The results obtained are satisfactory. Then we decided to make a bibliographical study in order to learn the various existing methods.

6.1.6.2. Voice conversion techniques applied to pathological voice repair

The subject of Ammar Werghi's thesis is the improvement of the esophageal voice using voice conversion techniques. To do this, we need to implement techniques similar or better than those described in the literature. 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 [63], 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 [49], 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. The preliminary results obtained until now are quite promising.

6.2. Automatic Speech Recognition

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

6.2.1. 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:

6.2.1.1. Missing data recognition

The objective of Missing Data Recognition (MDR) is to handle “highly” non-stationary noises, such as musical noise or a background speaker. These kinds of noise can hardly be tackled by traditional adaptation techniques, like PMC. Two problems have to be solved: (i) find out which spectro-temporal coefficients are dominated by noise, and (ii) decode the speech sentence while taking into account this information about noise.

We published a journal paper [14] that summarizes our work on context-dependency modeling of missing data masks. The context considered here is the whole frequency band along with the preceding mask. The paper presents extensive evaluation of our model on the noisy Aurora2 and Aurora4 numbers and large vocabulary speech recognition tasks. Furthermore, additional experimental results are given for concurrent speech, which is a very difficult task that has received specific attention from the missing data speech recognition community. The proposed models are analyzed both in terms of strengths and weaknesses, such as the dependency of mask models to the environment and the robustness of the mask clustering process.

6.2.1.2. 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, 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.

The detection experiments were carried out on the ESTER corpus using the segmentation and transcription tool ANTS developed in our Team. We evaluated the detection at the segment level. The detection results were represented as precision vs.recall (EER of 35%) [31].

6.2.2. Core recognition platform

6.2.2.1. 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 5.1.7).

Two versions of ANTS were implemented: the first one gives better accuracy but is slower (10 times real time), the second one is real-time (1 hour of processing for 1 hour of audio file).

This year, we included a tool for automatic speaker diairization (speaker segmentation and clustering): SELORIA (cf. 5.1.6). For acoustic features, we did not use first and second derivatives, but we concatenated MFCC parameters of 9 consecutive frames. We further reduced the number of parameters to 40 using HLDA (Heteroscedastic Linear Discriminant Analysis). For acoutic models, in order to be more robust to speaker variability, we used SAT (Speaker Adaptive Training). We also increased the size of the lexicon and for language model we moved form 3-gram to 4-gram.

We integrated the “liaison” phenomenon into the recognition engine. We evaluated the effect of the number of acoustic models for phonemes or allophones which are acoustically close.

Moreover, we tried to integrate linguistic knowledge using the random indexing technique for computing "semantic" distances between words.

Furthermore, we rewrited the training scripts in order to take advantage of the new cluster TALC. The training of the acoustic models was speeded up, from 1 week to less than 8 hours.

We presented our new version of ANTS (with HLDA and SAT) at the workshop ESTER 2009 at Paris.

6.2.2.2. Speech/music segmentation

We adressed the speech/music segmentation problem using a new parameterization based on wavelets. We studied different decompositions of the audio signal based on wavelets (Daubechie, Coiflets, symlets) which allow a better analysis of non stationary signals like speech or music. We computed different energy types in each frequency band. Results on an audio broadcast corpus gave significant improvement compared to classical MFCC features for music/non music segmention [15].

6.2.3. Speech/text alignment

Speech and text alignment is an old research area that can be considered as solved in constrained situations (relatively clean speech, limited size audio streams). However, we started the ALIGNE project in 2008 (see section 7.2.2) to answer a request from linguist researchers, who need to align long and noisy speech corpora with independent manual transcriptions. In contrast with recent state-of-the-art solutions to this problem, which basically automatically compute distant anchors with a large vocabulary speech transcription system, we have focused our work on the interactive control of the automatic algorithms by the user. Our objective is thus to help the user to work with semi-automatic algorithms rather than completely unsupervised batch processing. A Master internship (Josselin Pierre) has contributed in 2008 to the implementation of the first release of the jtrans software (see section 5.1.9). A first set of evaluations have been performed in 2009 by linguist researchers from the ATILF laboratory. The results of this user evaluation have shown some usability and related speed issues. We have then designed and proposed a new interaction paradigm, which is largely inspired by the reference Transcriber software [38], but which extends the functionalities proposed by Transcriber to a great extend, thanks to the smooth integration of semi-automatic alignment algorithms. The novel interaction model largely improves the alignement accuracy while further reducing the processing time and cognitive efforts as compared to Transcriber [24]. In addition, another Master internship (Jean-René Courtault) has greatly improved the phonetizer of JTrans in 2009 by implementing a classifer-based grapheme-to-phoneme converter for out-of-vocabulary words. More generally, the proposed JTrans software compares favorably to the other existing software for text and speech alignement, as it is the only one that integrates semi-automatic algorithms within an application GUI and proposes a smooth integration paradigm to reliably align corpora in faster than real-time.

6.2.4. Integration of linguistic information in speech recognition

One of most striking weakness of nowadays speech recognition systems is their total lack of understanding faculty, whereas everybody agrees that human processing of speech is largely guided by the semantic content of speech, and what can be understood out of it. A promising research area is then to investigate new research directions in the integration of higher-level information, typically related to syntax and semantic, into the speech decoding process.

The first information we have been interested in are dialog acts. Dialog acts represent the meaning of an utterance at the level of illocutionary force. This can be interpreted as the role of an utterance (or part of it), in the course of a dialog, such as statements or questions. The objective of our work is to automatically identify dialog acts from the user's speech signal. This is realized by considering both prosodic and lexical cues, and by training discriminant models that exploit these cues. This work, which has begun with Pavel Kral's PhD thesis, has been recently summarized in [16].

The second information we have investigated is the topic of the speech, which is a coarse semantic information that relates to the discourse thematics. In the past, research on thematic recognition have already been carried on in the team, for instance in Armelle Brun's Ph.D. thesis [40]. However, the current work differs from this previous studies because it addresses the specific case of speech input without any explicit linguistic or textual knowledge. The main advantage of this approach is its portability and its independance to the language. The basic principle proposed here consists first in extracting acoustic repetitions from the speech stream: the most frequent of these repetitions are then associated to a lexical entry. Then, the distributional hypothesis is applied to cluster the lexical entries into a hierarchy of clusters that are associated to the main thematics discussed in the corpus, leading to the building of a semantic lexicon. A system implementing this approach has been evaluted on two very different tasks, without any adaptation to the task, in order to show the robustness of the system that results from the lack of initial constraints. The first task is spontaneous telephone speech from the OGI corpus, while the second task is French broadcast news transcription. These experiments are described in details in [13].

The third work regarding the computation of syntax and semantic information for speech recognition has taken place within the context of the INRIA ARC RAPSODIS project, which has begun in feburary 2008. This project is described in section 7.3.1: it is a place of collaboration between specialists of different domains (lexical semantic, computational linguistic, speech recognition, ...). We have in particular investigated two aspects in the PAROLE team, respectively concerning lexical semantics and syntactic parsing.

For lexical semantics, we have based our work on the distributional hypothesis, which assumes that the meaning of a word can be deduced from its usage in context. We have collaborated in these aspects with the team CEA-LIST in Paris, which is specialized in the related aspects of Latent Semantic Analysis. We have hence exploited Random Indexing approaches, which are incremental dimensionality reduction techniques, based on the Johnson-Lindenstrauss theorem [37], that support very large corpora. We have used this approach to process “Le Monde” corpus and derive semantic distances between words that have an interesting potential for several future research works that may need to generalize lexical models without falling back to the (too) broad part-of-speech tags. Langage models or syntactic analysis are such potential applications.

Regarding syntactic parsing, we have decided in 2009 to invest a large amount of efforts in the development of a new syntactic parser dedicated to transcribed speech. This objective was motivated by the lack of existing parsing solutions for erroneously transcribed speech, and by the very important requirement of exploiting such a parsing in order, in the long term, to be able to compute semantic information beyond lexical semantics. Before taking this decision, we have spent more than one year trying to use existing parsers such as Syntex for this purpose, but none of them was efficient and adaptable enough. We have then strenghtened our collaboration with the TALARIS team, which is specialized in computational linguistics, to design and develop such a parser. The first step, achieved in 2009, has mainly consisted in focusing on the problem of parsing oral speech with stochastic dependency parsers, in order to more easily adapt and integrate the parsing model within our own stochastic framework. We have further decided to participate to the french syntactic parsing PASSAGE evaluation campaign organized in November 2009 (http://atoll.inria.fr/passage/eval2.en.html). The following joint paper [33] describes the resulting JSynATS parser. Other details can also be found in section 7.3.1.

All these works are described in details in the 2009 RAPSODIS project report available in the web site (http://rapsodis.loria.fr).

6.3. Speech-to-Speech Translation and Langage Modeling

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

The objective of our team is to provide an entire speech to speech system. Currently, the results presented are on text-to-text translation:

- **Phrase-based machine translation.** We pursued our effort to build a phrase based machine translation system. Last year, we retrieved the best phrases in a language; then, we translated them by using the concept of inter-lingual triggers and we selected the best ones by using simulated annealing algorithm [56]. This year, we systematically added n-gram of length 2, or 3, and 4 with their best inter-lingual triggers of length 2 or 3. This led to better results in terms of BLEU.
- **Confidence Measures for machine translation.** In machine translation, errors obviously happen. In order to estimate which confidence we give to the obtained translation,last year we decided to

develop several confidence measures based on mutual information, n-gram language model and lexical features language model. This year we pursued our efforts in this direction. This had led to three new publications describing the following results: (i) the combination of our measures yields a classification error rate as low as 25.1% with an F-measure of 0.708 [29]; (ii) the introduction of another standard confidence measures (backward n-gram) [30]; (iii) the design of a method to automatically build corpora containing realistic errors [28] (errors are introduced into reference translation with the supervision of Wordnet); (iv) the use of SVM to combine the confidence measures: the combination outperforms by 14% (absolute) our best single word-level confidence measure.
We are currently writing a journal paper on this work, and now, as the confidence measures show interesting discriminating power, we will integrate them in a more general process of discriminative training.

- **A decoder for Machine Translation.** We developed a first version of a machine translation decoder based on genetic algorithms. Dealing with genetic algorithms allows to use a search space composed of whole sentences. Then, in the future, it will be possible to use sentence-level confidence measures, or sentence-level evaluation such as syntactic correctness, in order to pilot the search algorithm. This decoder needs now to be systematically tuned and used with state of the art models (translation models, distortion models...). This work started during a research Master 2 training period supported by the CPER/TALC/TATI operation (http://wikitalc.loria.fr/dokuwiki/doku.php?id=operations:tati).
- **Language modelling for Arabic.** Always in the multi-lingual scope, but more on the language modelling aspect, we conducted works on Arabic Languages. In a first work, we used Multi-Category Support Vector Machines for topic identification [19]. Second, we studied the difference of modelisation (smoothing methods, order of n-gram) between French and Arabic [26].
- **Multi-lingual summarization.** In this work, we want to provide a translation of a document content. For that, we abord 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.

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 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.

In addition, we are involved in several regional projects.

7.2. Regional Actions

The team is involved in the management of the regional CPER contract. In particular, Christophe Cerisara is co-responsible, with Claire Gardent, for the CPER MISN TALC.

7.2.1. Investigation of speech production (MODAP)

The acquisition of articulatory data represents a key challenge in the investigation of speech production. These data could be used either to improve the naturalness of talking heads, or to add further information in automatic speech recognition. We thus initiated cooperation with Equipe IMS from SUPELEC and EPI Magrit to capture and exploit articulatory data.

This project relies on an articulograph AG 500 (base on the tracking of electromagnetic sensors) developed by Carstens. This equipment is available since August 2008 and has been complemented by real time software to make the acquisition of articulatory data easier. We already acquired several corpus of data which are used to evaluate inversion algorithms in a first time and some coarticulations effects, for instance those corresponding to pharyngealization in Arabic. Other acquisitions are scheduled in order to collect a corpus sufficiently vast to evaluate speech recognition algorithms, and also to study speech production dynamics.

7.2.2. Semi-automatic speech/text alignement (ALIGNE)

An active collaboration between the INRIA Parole and Talaris teams and researchers from the ATILF laboratory has started in 2008, in the framework of the ALIGNE project of the regional CPER MISN TALC contract. The objective of this collaboration is to design and develop an interactive software to help linguistic researchers in the process of aligning speech corpora, and for the manipulation of these corpus (e.g. for the purpose of anonymisation).

The main result of this collaboration is the release of the open-source JTrans software, which is distributed in the INRIA gforge. This project, which is lead by the Parole team, should continue at least until the end of 2009, and shall probably be continued in 2010, with a shift of focus towards syntactic processing of speech.

7.2.3. 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. The French intonation is characterized by recurrence of well-marked “continuation rises” whereas the English intonation is shaped by the recurrence of prominent peaks. The object of the action is to investigate the impact of a language characteristics on the perception and production of the other language. To that purpose, we will use speech processing tools developped in our team, especially those devoted to the automatic modification of prosody (fundamental frequency, durations, and energy).

7.3. National Contracts

7.3.1. RAPSODIS project

The RAPSODIS project is an “INRIA Action de Recherche Concertée” that has started in 2008 and will end in 2009. It is lead by the Parole team, and further involves three other INRIA teams (Talaris in Nancy and TexMex and Metiss in Rennes) and the CEA-LIST research team in Paris.

The main objective of this project is to study and analyze solutions for the integration of syntactic and semantic information within speech recognition systems. This project thus defines a pluridisciplinary framework that integrates researches from two main research areas: automatic speech recognition and natural language processing, including syntax parsing, semantic lexicon and distributional semantics.

The members of this project have chosen to address two main challenges:

1. Design and computation of syntactic and semantic features that may prove useful for speech recognizers;
2. Integration of these features into state-of-the-art speech recognition systems.

The work realized so far has mainly focused on exploring several types of syntactic and semantic information, such as dependency graphs derived from rule-based syntax parser, and thematic recognition computed from Random Indexing frequency matrices grabbed from the Web, and on investigating the possibility to rescore the speech recognizer n-best outputs with such information. One of the main difficulty to deal with is the lack of syntactic parser for oral speech processing (as opposed to written texts) in French. The current main research objectives concern the exploitation of stochastic syntactic parsers that can compute parsing probabilities for different sentences. A post-doctoral researcher has been hired in october 2009 to improve the training of such parsers with active learning.

More details can be found at http://rapsodis.loria.fr

7.3.2. ESTER project

As, in USA, NIST organizes every year an annual evaluation of the systems performing an automatic transcription of radio and television broadcast news, the French association AFCP (Association Francophone de la Communication Parlée) has initiated such an evaluation for the French language, in collaboration with ELRA (European Language Resources Association) and DGA (Délégation Générale pour l'Armement). The ESTER (Évaluation des Systèmes de Transcriptions Enrichies des émissions Radiophoniques) project evaluate different tasks: segmentation, as speech/music segmentation, speaker tracking system and orthographic transcription.

We have developed a fully automatic transcription system (Automatic News Transcription System: ANTS) containing a segmentation module (speech/music, broad/narrow band, male/female) and a large vocabulary recognition engine (see section 5.1.7). The first evaluation was conducted in January 2005. The next one took place from november 2008 until march 2009. We participated to this Ester 2 evaluation campaign for the broadcast news transcription task. The aim of this campaign was to evaluate automatic radio broadcasts rich transcription systems for the French language. This campaign was organised by DGA (Direction générale de l'armement) and AFCP (Association Francophone de la Communication Parlée).

We presented the new version of our Automatic News Transcription System (ANTS), including HLDA and SAT implementations, at the workshop ESTER 2009 at Paris.

7.3.3. 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 model.

7.3.4. 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.

7.3.5. 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, aknowledged 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.

7.3.6. 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.

During the first phase of the project (until sept 2009) we evaluated the contribution of automatic recognition of speech for indexing rushes in order to identify interesting topics of research. To be more precise, we used our Automatic News Transcription System, ANTS, to highlight the specific problems posed by the automatic transcription of a rush, The main problems are: speech/non-speeech segmentation, language identification and spontaneous speech recognition. In the second phase, a thesis funded by the project will begin about spontaneous speech recognition.

7.4. International Contracts

7.4.1. ASPI–IST FET STREP

The ASPI (Audiovisual to Articulatory Speech Inversion) project is funded by the European Commission in the framework of the 6th PCRD. The ASPI project, started on November 2005, aims at recovering the vocal-tract shape (from vocal folds to lips) dynamics from the acoustical speech signal, supplemented by image analysis of the speaker's face. The partners of this project are KTH (Stockholm), ULB (Brussels), ENST LTCI (Paris), and NTUA-ICCS (Athens). Together with the Magrit project we are involved in this project.

The final review of the project hold successfully in January 2009.

The main contributions of Parole were about inversion algorithm, especially inversion from standard spectral data (MFCC for instance), the inversion of fricatives, the incorporation of constraints and the development of software to analyze and exploit X-ray moving pictures.

7.4.2. 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.

7.4.3. The Oesovox Project 2009-2011: 4 international groups associated...

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 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". The activity reports of these two students for the year 2009 is described in 6.1.6.

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, Interspeech, CSL, Speech communication, TAL, IEEE Transaction of Information Theory, Signal Processing, Integration the VLSI journal, Pattern Recognition Letters, ICASSP, EURASIP.
- 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 :
  - Interspeech (J.P. Haton)
  - RFIA (D. Jouvet)
  - 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 lorrain 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 at :
  - TAIMA (Traitement et Analyse de l'Information : Méthodes et Applications[3]) Conference (K. Smaïli)
  - University of Annaba, Algeria (K. Smaïli)
  - Blaise Pascal University of Clermont-Ferrand (A. Kipffer-Piquard)
  - IUFM of Amiens (A. Kipffer-Piquard)
  - Universität des Saarlandes, Gipsa-lab and Telecom Paris (Y. Laprie)
  - International Workshop on Pharyngeals and Pharyngealisation, March 2009 at Newcastle University (S. Ouni and Y. Laprie)
- The team organised a Statistical Machine Translation Day. We invited four researchers in our domain from main French-speaking laboratories : GETALP (Grenoble, France), LIMSI (Paris, France), LIUM (Le Mans, France), RALI (Montréal, Canada). During this day, we presented our work to the community. Students of the Erasmus Mundus Master were present during this day.

[3] Processing and Analysis of Information: Methods and Applications

8.2. Invited lectures

- Bogdan Minescu
- Chiraz Latiri
- Rafael Laboissière

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 University Henri Poincaré (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).

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;
- 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", vol. 41, 2009.

[2] A. BONNEAU, Y. LAPRIE. *Selective acoustic cues for French voiceless stop consonants*, in "The Journal of the Acoustical Society of America", vol. 123, 2008, 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", vol. 21, n^{o} 3, 2007, p. 443-457.

[4] C. CERISARA, D. FOHR. *Multi-band automatic speech recognition*, in "Computer Speech and Language", vol. 15, n^{o} 2, April 2001, p. 151-174.

[5] C. CERISARA, L. RIGAZIO, J.-C. JUNQUA. *α-Jacobian environmental adaptation*, in "Speech Communication", vol. 42, n^{o} 1, January 2004, p. 25–41, Special Issue on Adaptation Methods for Automatic Speech Recognition.

[6] K. DAOUDI, D. FOHR, C. ANTOINE. *Dynamic Bayesian Networks for Multi-Band Automatic Speech Recognition*, in "Computer Speech and Language", vol. 17, 2003, p. 263-285.

[7] 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/.

[8] D. Langlois, A. Brun, K. SmaÃ¯li, J.-P. Haton. *Ãvénements impossibles en modélisation stochastique du langage*, in "Traitement Automatique des Langues", vol. 44, n^{O} 1, Jul 2003, p. 33-61.

[9] 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/.

[10] 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)", vol. 118 (1), 2005, p. 444–460, http://hal.archives-ouvertes.fr/hal-00008682/en/, PACS numbers: 43.70.h, 43.70.Bk, 43.70.Aj [DOS].

[11] I. Zitouni, K. Smaïli, J.-P. Haton. *Statistical Language Modeling Based on Variable-Length Sequences*, in "Computer Speech and Language", vol. 17, n^{O} 1, Jan 2003, p. 27-41.

Year Publications

Articles in International Peer-Reviewed Journal

[12] J. Cai, G. Bouselmi, Y. Laprie, J.-P. Haton. *Efficient likelihood evaluation and dynamic Gaussian selection for HMM-based speech recognition*, in "Computer Speech & Language / Computer Speech and Language", vol. 23, n^{O} 2, 2009, p. 147-256, http://hal.inria.fr/inria-00432533/en/.

[13] C. Cerisara. *Automatic discovery of topics and acoustic morphemes from speech*, in "Computer Speech & Language / Computer Speech and Language", vol. 23, n^{O} 2, 2009, p. 220-239, http://hal.inria.fr/inria-00330698/en/.

[14] S. Demange, C. Cerisara, J.-P. Haton. *Missing data mask estimation with frequency and temporal dependencies*, in "Computer Speech & Language / Computer Speech and Language", vol. 23, n^{O} 1, 2009, p. 25-41, http://hal.inria.fr/inria-00338397/en/.

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

[16] P. Kral, C. Cerisara. *Dialogue act recognition approaches*, in "Computing And Informatics", 2009, http://hal.inria.fr/inria-00431396/en/.

[17] L. Sprenger-Charolles, P. Colé, A. Kipffer-Piquard, F. Pinton, C. Billard. *Reliability and prevalence of an atypical development of phonological skills in French-speaking dyslexics*, in "Reading and Writing", vol. 22, n^{O} 7, 2009, p. 811-842, http://hal.archives-ouvertes.fr/hal-00414110/en/.

Invited Conferences

[18] S. Ouni, Y. Laprie. *Studying pharyngealisation using an articulograph*, in "International Workshop on Pharyngeals and Pharyngealisation, Royaume-Uni Newcastle", 2009, http://hal.archives-ouvertes.fr/hal-00431829/en/.

International Peer-Reviewed Conference/Proceedings

[19] M. Abbas, K. Smaïli, D. Berkani. *Multi-category support vector machines for identifying Arabic topics*, in "10th International Conference on Intelligent Text Processing and Computational Linguistics - CICLing 2009, Mexique Mexico", vol. 41, 2009, http://hal.inria.fr/inria-00403102/en/.

[20] M. Aron, A. Toutios, M.-O. Berger, E. Kerrien, B. Wrobel-Dautcourt, Y. Laprie. *Registration of Multimodal Data for Estimating the Parameters of an Articulatory Model*, in "IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Taiwan Taipei", 2009, http://hal.inria.fr/inria-00350298/en/.

[21] A. Bonneau, J. Busset, B. Wrobel-Dautcourt. *Contextual effects on protrusion and lip opening for /i,y/*, in "10th Annual Conference of the International Speech Communication Association - Interspeech 2009, Royaume-Uni Brighton", ISCA, 2009, http://hal.inria.fr/inria-00433381/en/.

[22] M. Cadot, A. Lelu. *Massive Pruning for Building an Operational Set of Association Rules: Metarules for Eliminating Conflicting and Redundant Rules.*, in "International Conference on Information, Process, and Knowledge Management - eKnow09, Mexique Cancun", A. Kusiak, Sang-goo. Lee (editors), 2009, p. 90-98, http://hal.inria.fr/inria-00337067/en/.

[23] J. Cai, Y. Laprie, J. Busset, F. Hirsch. *Articulatory Modeling Based on Semi-polar Coordinates and Guided PCA Technique*, in "10th Annual Conference of the International Speech Communication Association - INTERSPEECH 2009, Royaume-Uni Brighton", 2009, http://hal.inria.fr/inria-00433067/en/.

[24] C. Cerisara, O. Mella, D. Fohr. *JTrans, an open-source software for semi-automatic text-to-speech alignment*, in "Proceedings of the 10th Annual Conference of the International Speech Communication Association - Interspeech 2009, Royaume-Uni Brighton", 2009, http://hal.inria.fr/inria-00431398/en/.

[25] I. Jemaa, O. Rekhis, K. Ouni, Y. Laprie. *An Evaluation of Formant Tracking methods on an Arabic Database*, in "10th Annual Conference of the International Speech Communication Association - INTERSPEECH 2009, Royaume-Uni Brighton", 2009, http://hal.inria.fr/inria-00433057/en/TN.

[26] K. Meftouh, K. Smaïli, M. T. Laskri. *Comparative study of Arabic and french statistical language models*, in "International Conference On agents and Artificial Intelligence - ICAART'09, Portugal Porto", INSTICC, 2009, http://hal.inria.fr/inria-00352927/en/DZ.

[27] B. Potard, Y. Laprie. *A robust variational method for the acoustic-to-articulatory problem*, in "10th Annual Conference of the International Speech Communication Association - INTERSPEECH 2009, Royaume-Uni Brighton", 2009, http://hal.inria.fr/inria-00433053/en/.

[28] 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/.

[29] 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/.

[30] 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/.

[31] F. STOUTEN, D. FOHR, I. ILLINA. *Detection of OOV words by combining acoustic confidence measures with linguistic features*, in "The eleventh biannual IEEE workshop on Automatic Speech Recognition and Understanding (ASRU), Italie Merano", 2009, p. 1-4, http://hal.archives-ouvertes.fr/hal-00435087/en/.

National Peer-Reviewed Conference/Proceedings

[32] J. BUSSET. *Utilisation d'une grille polaire adaptative pour la construction d'un modèle articulatoire de la langue*, in "Rencontres Jeunes Chercheurs en Parole - RJCP 2009, France Avignon", 2009, 72, http://hal.archives-ouvertes.fr/hal-00433299/en/.

[33] C. CERISARA, C. GARDENT. *Analyse syntaxique du français parlé*, in "Journée ATALA, France Paris", 2009, http://hal.inria.fr/inria-00432754/en/.

Scientific Books (or Scientific Book chapters)

[34] M. ARON, M.-O. BERGER, E. KERRIEN, Y. LAPRIE. *Acquisition multimodale de données articulatoires*, in "L'imagerie médicale pour l'étude de la parole", A. MARCHAL, C. CAVÉ (editors), Hermes Science Publications, 2009, p. 175-196, http://hal.inria.fr/inria-00429585/en/.

[35] M. EMBARKI, S. OUNI, M. YEOU, C. GUILLEMINOT, S. AL MAQTARI. *Acoustic and EMA study of pharyngealization : Coarticulatory effects as index of stylistic and regional distinction*, in "Instrumental Studies in Arabic Phonetics", M. HASSAN, B. HESELWOOD (editors), Benjamins, 2009, p. 1-56, http://hal.archives-ouvertes.fr/hal-00348775/en/.

References in notes

[36] 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", vol. 3, n^{o} 4, 1995, p. 85–89.

[37] D. ACHLIOPTAS. *Database-friendly random projections*, in "PODS '01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, New York, NY, USA", ACM, 2001, p. 274–281.

[38] C. BARRAS, E. GEOFFROIS, Z. WU, M. LIBERMAN. *Transcriber: development and use of a tool for assisting speech corpora production*, in "Speech Communication", 2001, p. 5–22.

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

[40] A. BRUN. *Détection de thème et adaptation des modèles de langage pour la reconnaissance automatique de la parole*, Université Henri Poincaré - Nancy I, 2003, Ph. D. Thesis.

[41] 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.

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

[43] 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/.

[44] V. COLOTTE, Y. LAPRIE. *Higher precision pitch marking for TD-PSOLA*, in "XI European Signal Processing Conference EUSIPCO, Toulouse, France", vol. 1, September 2002, p. 419-422.

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

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

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

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

[49] A. KAIN, M. MACON. *Spectral voice conversion for text-to-speech synthesis*, in "International Conference on Acoustics, Speech, and Signal Processing", May 1998, p. 285–288.

[50] 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, http://hal.inria.fr/inria-00185312/en/, 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..

[51] 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/.

[52] 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.

[53] 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.

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

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

[56] C. LAVECCHIA, D. LANGLOIS, K. SMAÏLI. *Discovering Phrases in Machine Translation by Simulated Annealing*, in "INTERSPEECH 2008 - 9th Annual Conference of the International Speech Communication Association, Brisbane Australie", 2008, p. 2354-2357, http://hal.inria.fr/inria-00331327/en/.

[57] 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/.

[58] 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/.

[59] 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.

[60] 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.

[61] 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/.

[62] L. SPRENGER-CHAROLLES, P. COLÉ, A. KIPFFER-PIQUARD, F. PINTON, C. BILLARD. *Reliability and prevalence of an atypical development of phonological skills in french-speaking dyslexics*, in "Reliability and prevalence of an atypical development of phonological skills in french-speaking dyslexics".

[63] Y. STYLIANOU, O. CAPPÃ©, E. MOULINES. *Continuous probabilistic transform for voice conversion*, in "IEEE Transactions on Speech and Audio Processing", vol. 6, n^{o} 2, March 1998, p. 131–142.

[64] A. TOUTIOS, S. OUNI, Y. LAPRIE. *Protocol for a Model-based Evaluation of a Dynamic Acoustic-to-Articulatory Inversion Method using Electromagnetic Articulography*, in "The eighth International Seminar on Speech Production - ISSP'08, Strasbourg France", 2008, http://hal.archives-ouvertes.fr/inria-00336380/en/.