INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team Parole

Analysis, Perception and Speech Recognition

Nancy - Grand Est

THEME COG

Table of contents

PAROLE

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, HdR]
Anne Bonneau [Research scientist CNRS]
Christophe Cerisara [Research scientist CNRS]
Dominique Fohr [Research scientist CNRS]

Faculty Member

Martine Cadot [PRAG, Henri Poincaré University]
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, Henri Poincaré University, HdR]
Irina Illina [Assistant Professor, I.U.T Charlemagne, Nancy 2 University, HdR]
David Langlois [Assistant Professor, IUFM (University Institute for Teacher Training)]
Agnès Kipffer-Piquard [Assistant Professor, IUFM (University Institute for Teacher Training)]
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]
Jun Cai [Associate Professor, Xiamen University]

External Collaborator

Joseph Razik [PostDoc ENST]

Technical Staff

Jonathan Ponroy [INRIA, CPER, 50% (collaboration with Magrit) since 1st February]
Alexandre Lafosse [INRIA (associate engineer)]
Benjamin Husson [CNRS and then INRIA, since 1st February]

PhD Student

Ghazi Bouselmi [TA, ATER until september 2008, thesis to be defended in november 2008]
Christian Gillot [MENRT grant, thesis to be defended in 2011]
Blaise Potard [MENRT grant, thesis defended in oct. 2008]
Vincent Robert [High school teacher, thesis to be defended in 2008]
Caroline Lavecchia [EADS foundation grant, thesis to be defended in 2008]
Farid Feïz [CNRS grant (ASPI contract), thesis to be defended in 2008]
Marina Piat [MENRT grant until july 2008]

Post-Doctoral Fellow

Frederic Stouten [from october 1st]
Luz Garcia [University of Granada - july-october]
Asterios Toutios [INRIA]

Visiting Scientist

Konstantin Markov [invited professor ATR Japan]

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 interfaces and necessitates works in analysis, perception and automatic speech recognition (ASR).

Our activities are structured in three topics:

- **Speech analysis.** Our works are concerned with automatic extraction and perception of acoustic cues, acoustic-to-articulatory inversion and speech synthesis. These themes give rise to a number of ongoing and future applications: vocal rehabilitation, improvement of hearing aids and language learning.
- **Modeling speech for automatic recognition.** Our works are concerned with stochastic models (HMM[1], bayesian networks and missing data models), multiband approach, adaptation of a recognition system to a new speaker or to the communication channel, and with language models. These topics give also rise to a number of ongoing and future applications: automatic speech recognition, text-to-speech alignment and audio indexing.
- **Speech to Text Translation** 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 and multiband approaches 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 several cooperations with companies using automatic speech recognition, for instance, the one with TNS Sofres about word spotting. We have a cooperation with EDF and Audivimedia in the form of an RIAM project. We recently had a contract with Ninsight and Thales Avionics. The latter gave rise to a European project in the field of non-native speech recognition in a noisy environment. We are also involved in the 6th PCRD projects MUSCLE, AMIGO, HIWIRE and more recently ASPI as the coordinator team, and in a regional project with foreign language teachers in Nancy within the framework of a Plan État Région project.

3. Scientific Foundations

3.1. Introduction

Keywords: *Digital signal processing*, *acoustic cues*, *automatic speech recognition*, *health*, *language learning*, *language modeling*, *lipsync*, *perception*, *phonetic*, *speech analysis*, *speech synthesis*, *stochastic models*, *telecommunications*.

[1] Hidden Markov Models

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

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

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

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

These two approaches are complementary; the latter borrows theoretical results on speech from the former, which, in its turn, borrows some numerical methods. Spectral analysis methods are undoubtedly the domain where exchanges are most marked. The simultaneous existence of these two approaches is one of the particularities of speech research conducted in Nancy and we intend to enhance exchanges between them. These exchanges will probably grow in number since automatic speech recognition systems (especially those for automatic dictation) are now available for every consumer: their acoustical robustness (against noise and speaker variation) and their linguistic reliability have to be increased.

3.2. Speech Analysis

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

3.2.1. Acoustic cues

The notion of strong (selective) and weak cues has been introduced to palliate a weakness of ASR systems: the lack of confidence. Indeed, due to the variability of speech signals, acoustical regions representing different sounds overlap one with another. Nevertheless, we know from previous perceptual experiments [39], that some realizations of a given sound can be discriminated with a high level of confidence. That is why we have developed a system for the automatic detection of selective cues, devoted to the reliable recognition of stop place of articulation. Selective cues identify or eliminate a feature of a given sound with certainty (almost no error is allowed). Such a decision is possible in few cases, when the value of an acoustic cue has a high power of discrimination. During selective cue detection, we must fulfill two requirements: to make no error on the one hand, and to obtain a relatively high firing rate, on the other hand. The notion of selective cue must not be merged into the one of “robust” cue or landmark which are systematically fired and can make some errors. On a corpus, made up of approximately 2000 stops, we obtained a firing rate for stop bursts and transitions in more than one case out of three [11].

Selective cues can be exploited either to improve speech intelligibility (through the enhancement of the most reliable cues), with application to language learning or hearing impairment, or to provide “confidence islands” so as to reduce the search space during the lexical access, in automatic speech recognition.

3.2.2. 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).

3.2.2.1. Automatic detection and correction of a learner’s second language oral realizations

Within the framework of a project concerning language learning, more precisely the acquisition of the prosody of a second language, we are studying automatic detection and correction of prosodic deviations. 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 model. The identification and correction tasks use speech analysis and modification tools developed in our team. We started our project with the automatic detection of the lexical accent of "transparent" words. For more complex identification tasks, we plan to implement a prosodic model.

3.2.2.2. Phonemic discrimination in language acquisition and language disabilities

This year, 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 seem 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 [48], [49] 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 [60].

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 ?

3.2.3. 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 [57].

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.

3.2.4. 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 [38] 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.

3.3. Automatic speech recognition

Automatic speech recognition aims at reproducing the cognitive ability of humans to recognize and understand oral speech. This difficult challenge cannot be solved globally, and a reasonable approach consists of decomposing it into simpler problems and related technologies. At the broadest scale, we identify two classes of problems: the first one is called “acoustic features and models”. It relates to the processing of speech signal. The second one is called “language modeling”, and it addresses the problem of modeling and understanding natural language. Both these research problems are further analyzed and decomposed in the next sections. Despite this artificial (but necessary) division of the task, our ambition is to merge all these approaches to solve the problem globally. The dependencies between these research areas are thus favored whenever our research work and applications make it possible. These connections are facilitated in our team, thanks to the common statistical basis we share, i.e. stochastic and Bayesian modeling approaches.

3.3.1. Acoustic features and models

3.3.1.1. Acoustic features

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 recently used and 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 have further developed a new robust front-end, which is based on wavelet-decomposition of the speech signal. This front-end generalizes the Frequency filtered coefficients. Finally, we also largely exploit the standard ETSI advanced front-end [45], which is famous for its robustness to noise.

3.3.1.2. 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) and Bayesian Networks (BN). HMMs are mainly used to model the acoustic units to be recognized (usually triphones) in all of our recognition engines (ESPERE, ANTS...), while BNs constitute powerful investigation tools to develop new research ideas by explicitly representing the random variables and their independence relationships. For example, BNs can be used to model the relations between clean and noisy speech in denoising, or between the environment classes and the mask models in missing data recognition. We do not do research on BN, but we rather exploit them to work on the important statistical properties of robust speech recognition.

3.3.1.3. Robustness and invariance

The core of our research activities about ASR aims at improving the robustness of recognizers to the different kinds of variabilities 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 have 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). These technologies constitute the foundations of our recent developments in this area, such as non-native speaker adaptation, missing data recognition and denoising. 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.

3.3.1.4. 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 is often based on the acoustic differences between both kinds of sounds. So discriminative acoustic cues are investigated (FFT, zero crossing rate, spectral centroid, wavelets ...). Except the selection of acoustic features, another point is to find the best 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 smaller 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.

Our research has been applied to large vocabulary dictation machine, news transcription, automatic categorization of mails, dialog systems, vocal services...

3.4. Speech to Speech Translation

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 was 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 [40] proposes 5 statistical translation models which became inexorable 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 [56]. 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. [58] 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 [51] or MOSES [50] 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 (cf. 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. Furthremore, 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, automatic transcription, and keyword spotting.

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. The automatic formant tracking that has been implemented is an improved version of the concurrent curve formant tracking [54]. 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 four years, formant tracking enables copy synthesis. The current version of WinSnoori is available on http://www.winsnoori.fr.

5.1.2. Labelling corpora

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 [53]. That's why we began to develop another tagger based on a unsupervised tagging algorithm.

The TTS platform currently under development (see 5.1.9) uses the latter tagger with the class set of the former tagger.

5.1.3. Automatic lexical clustering

In order to adapt language models in ASR applications, we have developed a new toolkit to automatically create 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. ESPERE

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

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

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

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

5.1.5. SELORIA

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.

5.1.6. ANTS

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.

Recently, a real time version of ANTS have been developped. The transcrption is done in real time on a quad-core PC.

5.1.7. JTRANS

JTrans is an open-source software for semi-automatic alignement of speech and textual corpus. It is written 100% in JAVA and exploits the JHTK library 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 should be available soon.

JTrans is developed in the context of the CPER MISN TALC 2008-2009 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

5.1.8. STARAP

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.

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

5.1.9. TTS SoJA

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, ...).

The major parts of modules have been developed in Java. Some modules are in C. The platform has been developed to take 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 principle.

The official release is planned to the end of the year. The software license should easily allow, for associations working with impaired people, to use the software.

6. New Results

6.1. Speech Analysis

Keywords: *Signal processing*, *acoustic cues*, *articulatory models*, *health*, *hearing help*, *learning language*, *perception*, *phonetics*, *speech analysis*, *speech synthesis*.

Participants: Anne Bonneau, Vincent Colotte, Dominique Fohr, Yves Laprie, Joseph di Martino, Slim Ouni, Jacques Feldmar, Agnes Kipffer, Martine Cadot, Blaise Potard, Vincent Robert, Alexandre Lafosse, Jonathan Ponroy, Benjamin Husson.

6.1.1. Acoustic-to-articulatory inversion

Our approach of acoustic-to-articulatory inversion is an advanced table lookup method. The table is an articulatory codebook, each entry associating a vector of articulatory parameters together with its acoustic image given by the first three formant frequencies.

The codebook is thus a key component of our analysis-by-synthesis inversion method since it represents the synthesis facet of the algorithm. Therefore it needs to offer a homogoneous articulatory density. This year we thus designed a new method [28] for sampling the null space of the acoustic-to-articulatory mapping, which is considerably faster and more accurate than the previous method. The first aspect improved is the time to decide whether a hypercuboid contains solutions. Although the property of the initial solution found via SVD – to be the closest to the center of the hypercuboid – is indeed quite strong, it is not appropriate to decide whether a hypercuboid contains solutions. It is however straightforward to determine an approximate solution that "almost" minimizes the dH_c norm ($|x|_r = \max_{1\leq i\leq N} \frac{|x_i|}{r_i}$ where r_i is the edge of the coordinate direction i in the hypercuboid) using a stochastic exploration of the space of solutions, starting from the point found by SVD, iteratively sampling random solutions, and keeping it as a new starting position when closer to the center with regards to dH_c. This yields a new initial point P_{H_c} used to generate solutions. A hypercuboid H_c will be considered to contain solutions if and only if $d_{H_c}(PH_c - P_0) \leq 1$. After having decided that a hypercuboid contains solutions, the volume of the space of solutions v_{H_c} is evaluated in order to determine the number of solutions to be generated. Then, solutions are obtained doing random moves in the solution space from P_{H_c} and keeping only points belonging to H_c.

The work conducted in the framework of the European ASPI project has focused the incorporation of constraints coming either from phonetic knowledge or audiovisual data. These constraints and the corresponding strategies are addressed in several papers [14], [22].

6.1.2. Articulatory Speech Production

6.1.3. 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. This year we developed 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). In the near future, 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 [32].
- **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. This year, we highlighted the differences between pharyngealized phonemes and non-pharyngealized ones. The articulation of the tongue was tracked by four sensors glued on the tongue. A corpus of several words in Arabic 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 [26].

6.1.4. Coarticulation

Our approach of coarticulation represents transitions in the form of sigmoid curves which can be easily concatenated. Since the concatenation algorithm would require a huge corpus (which is not available yet) to cover all the transitions CV (Consonant Vowel), VCV, VCCV, a completion algorithm has been developed to derive all transitions from those recorded in the corpus.

This year we conducted a complete evaluation of the approach [31]. It turns out that the implementation of the Cohen & Massaro algorithm gives slightly better results than the concatenation method. It should be noted that Beskow, with a similar corpus for Swedish, obtained results not as good as ours. This probably originates in the slight improvements implemented in our version of this algorithm. We also noticed that the amplitude registration intended to ensure the syntagmatic consistency of the labial parameters is more efficient for protrusion which presents a larger degree of freedom. On the other hand, the paradigmatic axis seems to be more important for other labial parameters and the Cohen & Massaro algorithm implicitly favors the paradigmatic axis since it is based on dominance functions attached to each of the phonemes. This probably explains that performances are better for lip opening and stretching, and jaw opening. A finer analysis showed that results obtained by the concatenative approach is markedly better that the Cohen & Massaro algorithm for the prediction of the protrusion in consonant clusters (20% more for the correlation and 4% less for the RMSE).

6.1.5. Text-to-Speech synthesis

We are developing (2006-2008) a software platform to Text-To-Speech synthesis : TTS SoJA (Synthesis platfOrm in JAva). An INRIA associate engineer, Alexandre Lafosse worked on this project (with Vincent Colotte). The aim of this project is to obtain a system as a toolkit for TTS or speech recognition applications (Natural Langage Processing (NLP) tools) and as free TTS system to be used by associations for hearing impaired or partially sighted (or blind) people.

During the last year, the NLP and the Non Uniform Unit (NUU) selection steps have been built. The NLP part tags and analyzes the input text to obtain a feature set (phoneme sequence, word grammar categories, syllables...). From NLP information, the second part selects, in a recorded corpus, units which can be diphone, syllable or part of words. The selection is based on a Viterbi algorithm. The originality of our approach is that the selection is piloted by linguistic features without acoustic model of prosody (see [43]). At the end of the last year, we presented a demo version at the 40th aniversary celebration of INRIA (http://parole.loria.fr/Demo/Demo40INRIA/flash.html , synthÃ¨se de la parole). A small corpus has been built for this event (<1h).

During this year, we built our corpus for the selection. Indeed, a TTS system need a corpus generally with 3-5 hours of speech recorded by one speaker. We developed a greedy algorithm to extract N textual sentences from M textual sentences with N $<<$ M. The N sentences must cover a set of required features (for instance, one criterion is that all diphones must be represented in the corpus). The iterative algorithm is based on the seek of the best sentence (among the M sentences) which contains a maximum of units/features. This iterative algorithm is used until all features are covered. The major point of the algorithm is the choice of features to cover. It's obvious than we can not seek all diphones in all positions in the sentence to hope a corpus with only about 2000 sentences. We focused our approach on diphone positions into the rhythm group. The rhythm group is one of the major predictor of prosody in french. We worked from 6 millions of sentences extracted from LEMONDE corpus (about 10 years). We successively launched 3 times our greedy alogrithm to obtain a selection of sentences with different length (long, medium and short). At the end, we kept 1600 sentences and we supplemented this set with 200 sentences (semi-automatically extracted) to be such that some major diphones are not missing. These sentences have been recorded in 5 days (not conseccutively) in a quiet room (soundproofed for the kind of work). The obtained corpus is about 3 hours long. We developed some tools to take easy the recording for the speaker and for the automatic transcription alignment. We also recorded about 800 interrogative sentences to supplement the corpus, but these sentences have not already been included in the system.

The official release of this platform is planned at the end of the year. This platform will allow us to continue researches about corpus building for synthesis systems based on NUU selection and to end up in project to elaborate a system of acoustic-visual synthesis (ANR Jeunes Chercheurs ViSAC).

6.1.6. Automatic correction of the prosody of English as a second language

A preliminary experiment, with a small corpus and five subjects, has been conducted in order to verify whether speech modifications could improve the perception of English prosody by French speakers. The modifications deeply change the learner's prosodic realisation so as to make it very close to that of an English human reference. The exercise focused on the realisation of the English lexical accent, which is strongly marked with respect to the French one. The session was unfolding in the following manner: the learner uttered a sentence, then heard and visualized the sentence after modification, and finally was asked to repeat the sentence with the expected prosody. The evaluation consisted in verifying whether the learner's production has improved. Preliminary results showed that the learners tended to mark the lexical accent in a more satisfactory way (one criterion being the presence of peak on the correct syllable, i.e. the one bearing the lexical accent). More experiments should be conducted to confirm these first results and show the interest of a tutor to improve English prosody.

Another important difference of English prosody with respect to the French one is the frequent use of a "focus" accent (an accent which gives information on the sentence topic). A study was conducted to examine the acoustic cues of the focus accent in interrogative sentences (its realization in declarative sentences have been the object of numerous studies). A set of English sentences was pronounced by a native speaker of English and

a native speaker of French. Results showed the importance of some acoustic cues, such as the velocity of the fundamental frequency movement near the peak or the peak height. They also showed that the French speaker encountered very strong difficulties to realize this accent, arguing for the necessity of a special learning of “focus”.

Both these experiments stand for the interest of a tutor to improve the learning of foreign language prosody and we will continue our research in this direction.

6.1.7. 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 [48] 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 [61].

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.

6.1.8. Development of robust data mining for speech data

The specificity of our data (huge amount of observations, and numerous variables, quantitative and qualitative with numerous modalities) has invited us to develop a data mining method based upon the "Association Rule Extraction". In despite of the robustness of our method, the produced rule set may contain incoherencies. A method of building an operational rule set is currently under development in our team [17].

6.2. Automatic Speech Recognition

Keywords: *acoustic models*, *automatic speech recognition*, *language models*, *robustness*, *stochastic models*, *telecommunications*, *training*.

Participants: Ghazi Bouselmi, Jun Cai, Christophe Cerisara, Dominique Fohr, Jean-Paul Haton, Irina Illina, Pavel Kral, David Langlois, Odile Mella, Marina Piat, Joseph Razik, Kamel Smaïli, Luz Garcia.

6.2.1. Robustness of speech recognition

Robustness of speech recognition to noise and to speaker 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 [44] 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.

We have continued our efforts in this domain by proposing a completely new solution to adapt the missing data paradigm to new parameterization domains, which are more robust to noise than the spectral domain. In this approach, the mask concept is generalized to represent any kind of undesired speech variability, and not only additive noise. The damaging coefficients are now localised from a speech confidence measure, and a second speech recognition pass that exploits the resulting masks is realized that prevents the speech decoder to take decisions based on these coefficients. This method is evaluated in [20].

6.2.1.2. Non-native speakers

The performance of automatic speech recognition (ASR) systems drastically drops with non native speech. The main aim of non-native enhancement of ASRs is to make available systems tolerant to pronunciation variants by integrating some extra knowledge (dialects, accents or non-native variants).

Our main motivation is to develop a new approach for non-native speech recognition that can automatically handle non-native pronunciation variants without a significant loss in recognition time performance. As non-native speakers tend to realize phones of the spoken language as they would do with similar phones from their native language, we claim that taking into account the acoustic models of the native language in the modified ASR system may enhance performance. We automatically extracted association rules between non-native and native phones models from an audio corpus recorded by non native speakers. Then, new acoustic models were built according to these rules.

This year, we investigate the feasibility of multi-accent non-native speech recognition without any accent detection layer. Tests on the *HIWIRE* corpus show that multi-accent pronunciation modeling and acoustic adaptation reduce the WER by up to 76% compared to results of canonical models of the target language. We also investigate accent-independent approaches in order to assess the robustness of the proposed methods to unseen foreign accents. Experiments show that our approaches correctly handle unseen accents and give up to 55% WER reduction, compared to the models of the target language. Finally, the proposed pronunciation modeling approach maintains the recognition accuracy on canonical native speech as assessed by our experiments on the *TIMIT* corpus [16].

We propose an automatic approach for foreign accent identification. Knowledge of the speaker's origin allows to adapt acoustic models for non-native speech recognition. In this study, we use a statistical approach based on prosodic parameters. This approach relies on the fact that prosody is different between languages, and has been done within the framework of the HIWIRE (*Human Input that Works In Real Environments*) European project. The corpus is composed of English sentences pronounced by French, Italian, Greek and Spanish speakers. Results obtained with duration and energy are promising for foreign accent identification: 67.1% correct L1 identification with duration and 68.6% with energy. These two parameters combined with MFCC achieve a 87.1% correct foreign accent identification rate [27], [35].

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 5.1.6).

In order to adapt acoustic models to the speaker, we have added two new modules: one for speaker turn detection and speaker clustering and another one for MLLR-MAP adaptation. The clustering process is based on the Bayesian Information Criterion (BIC).

Two versions of ANTS have been 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 improved the training of acoustic models:

- HLDA (Heteroscedastic Linear Discriminant Analysis): the goal of HLDA is to generalize LDA under ML (Maximum Likelihood) framework.
- SAT (Speaker Adaptive Training): in SAT, speaker characteristics are modeled explicitly as linear transformations of the SI acoustic parameters.

Moreover, we increased the size of the lexicon (72000 words) and the order of the n-gram Language model (from 3 to 4).

6.2.2.2. Confidence measures

In automatic speech recognition, confidence measures aim at estimating the confidence we can give to a result (phone, word, sentence) provided by the speech recognition engine; for example, the contribution of the confidence measure allows to highlight the misrecognized or out-of-vocabulary words. In this framework, we proposed several word confidence measures which are able to provide this estimation for applications using large vocabulary and on-the-fly recognition, as keyword indexation, broadcast news transcription, and live teaching class transcription for hard of hearing children. We defined two types of confidence measures. The first, based on likelihood ratio, are frame-synchronous measures which can be computed simultaneously with the recognition process of the sentence. The second ones are based on an estimation of the posterior probability limited to a local neighborhood of the considered word, and need only a short delay before being computed. We evaluated these measures according to Equal Error Rate (EER) criterion in an automatic transcription task using the French broadcast news corpus ESTER [30], [36]. Our local measures achieved performance very close to a state-of-the-art measure which requires the recognition of the whole sentence (23.% EER vs. 22%). This year we studied how our local measure can increase the understanding of an automatic transcription by hard of hearing. For that, we conducted a preliminary experiment with 20 subjects in order to study two visual modalites to highlight low-confidence words and an experimental protocol to assess the improvement of the transcription comprehension by the subjects [37], [29].

6.2.2.3. Speech and 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 to the implementation of the first release of the jtrans software (see section 5.1.7).

6.2.3. 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 continued in 2008 with a publication in an IEEE Workshop [21].

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 [41]. However, the current work realized in 2008 deeply 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 [12].

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, ...). The first work realized has been a study of the state-of-the-art. The *semantic map* model, first developped in the CEA, has also been shared amongst the partners of the project and studied in terms of usefulness for speech recognition: this study has shown several potential issues. A Master student (Omar Bouras) has also studied during his internship in our team a semantic model derived from the Random Indexing approach and inspired by the semantic map. The semantic map has also been exploited to desambiguate terms in an internship (Myriam Rakho) co-supervised by the CEA and the Parole team. All these works are described in details in an internal report and in two internship reports available in the RAPSODIS web site (http://rapsodis.loria.fr). The current focus of the project is on exploiting syntax for speech recognition: 400 millions of words from the newspaper "Le Monde" have thus been parsed with the rule-based Syntex syntax parser. We plan to further process additional text and audio corpora, with the objective of training a stochastic dependency parser, from which parsing probabilities can be computed for alternative speech recognition words sequences and integrated into the speech decoder itself.

6.3. Language Modeling for Speech-to-Speech Translation

Keywords: *machine translation*, *statistical models*.
Participants: Kamel Smaïli, David Langlois, Caroline Lavecchia.

All the experiments we conduct on machine translations are achieved on two corpora EUROPARL [52] and on a movie subtitles corpus [55].
The results of this year consists in:

- **Inter-Lingual Triggers**: We extend the idea of triggers used in statistical language models to bilingual corpus. This principle allows us to find out relationships between a word and its co-occurred words in a target language. This original idea allows us to propose an alternative method to the classical IBM's translation models. This method achieves better results than the third model of IBM [34].
- **A phrase-based statistical machine translation**: A realistic machine translation system must be phrase-based. This year we achieved a new version of statistical machine translation. It is based on phrases which have been learnt by using an original method. We first start retrieving the best phrases in a language; then, we translate them by using the concept of inter-lingual triggers and then we select the best ones by using simulated annealing algorithm. The first results are very encouraging but unfortunately they are less powerful than those obtained by the state of art [24], [23]. We are working on improving our method.
- **Confidence Measures in machine translation**: In machine translation, errors are obviously possible. In order to estimate which confidence we give to the yielded translation, we decided to develop several confidence measures [59]. These measures show interesting discriminating power, that is why we will integrate them in a more general process of discriminative training.

7. Contracts and Grants with Industry

7.1. Introduction

Our policy in terms of industrial partnership consists in favoring contracts that quite precisely fit our scientific objectives. We are involved in several cooperations with companies using automatic speech recognition, for instance, the one with TNS Sofres about word spotting. We have a cooperation with EDF and Audivimedia in the form of an RIAM project. We recently had a contract with Ninsight and Thales Aviation. The latter gave rise to a European project in the field of non-native speech recognition in a noisy environment. We are also involved in the 6th PCRD projects MUSCLE, AMIGO, HIWIRE and more recently ASPI as the coordinator team.

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 and will be complemented by real time software to make the acquisition of articulatory data easier. We already acquired a small corpus of data which will be used to evaluate inversion algorithms in a first time. Further 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 has been the first release of the JTrans software, which shall be deployed in the near future in the TALC platform and in the ATILF laboratory. This project, which is lead by the Parole team, should continue at least until the end of 2009.

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 should last until the end of 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. A Ph.D. thesis has begun in october 2008 on this specific topic. Another research track currently explored concerns the exploitation of stochastic syntactic parsers that can compute parsing probabilities for different sentences. A post-doctoral researcher might also be hired in 2009 to design algorithmic solutions to integrate syntax and semantic information within speech recognizers.

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 5.1.6). The first evaluation has been conducted in January 2005. The next one will take place from november 2008 until march 2009

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.

7.4. International Contracts

7.4.1. Amigo

Amigo is an Integrated Project funded by the European Commission, whose main topic is "Ambient intelligence for the networked home environment". It is leaded by Philips Research Eindhoven and includes 15 European partners. In this project, we have mainly continued the efforts we have begun in OZONE, with a focus on multimodality (speech, 2D and 3D gestures with VTT), and on adapting our speech technologies to handle implicit user interactions.

7.4.2. Muscle

Due to the convergence of several strands of scientific and technological progress we are witnessing the emergence of unprecedented opportunities for the creation of a knowledge driven society. Indeed, databases are accruing large amounts of complex multimedia documents, networks allow fast and almost ubiquitous access to an abundance of resources and processors have the computational power to perform sophisticated and demanding algorithms. However, progress is hampered by the sheer amount and diversity of available data. As a consequence, access can only be efficient if based directly on content and semantics, the extraction and indexing of which is only feasible if achieved automatically.

MUSCLE aims at creating and supporting a pan-European Network of Excellence to foster close collaboration between research groups in multimedia datamining and machine learning. Our contribution will be on the development of acoustic-to-articulatory inversion and the improvement of the robustness of ASR through the use of Bayesian networks.

Muscle is a Network of Excellence funded by the European Commission.

Our contribution concerns speech analysis, improvement of automatic speech recognition robustness and language models.

This year we have completed a demonstrator that incorporates real time phonetic speech recognition and the talking head developed by the speech group of KTH. Automatic speech recognition is used to pilot the talking head. In addition, the real time computation of F0 (fundamental frequency) enables the improvement of the rendering.

7.4.3. 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 HIWIRE 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.

This year, the first main achievement concerns the acquisition of two multimodal corpora with the ASPI system, one in French and the second in Swedish. The ASPI system enables the synchronized acquisition of ultrasound images, electromagnetic sensors, speech and face stereovision images to be captured. The second main achievement concerns inversion. We improved the exploration of the null space in order to get a continuous sampling density. We also designed an inversion algorithm which does not need the codebook to initialize inverse solutions. We also processed two corpora of X-ray data in order to design articulatory models offering a better precision in the front region of the mouth cavity. Indeed this region is important for the production of fricatives, which is a key objective of the team.

The acquisition of MRI data has been continued in order to elaborate a 3D model of the vocal tract.

The main contributions of Parole are 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.

7.4.4. 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.

7.4.5. The LARYNX Project 2006-2008

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, Associat Professor at INPL-ENSEM) associated with the following partners:
Spain: Begona Garcia Zapirain, Deusto University, Bilbao
Tunisia: Sofia Ben Jebara, TECHTRA research group, SUP'COM, Tunis.
Morocco: Â 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.

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, Eurospeech, CSL, Speech communication, TAL, IEEE Transaction of Information Theory, Signal Processing, Integration the VLSI journal, Pattern Recognition Letters.
- Member of editorial boards :
  - Speech Communication (J.P. Haton)
  - Computer Speech and Language (J.P. Haton)
  - EURASIP Journal on audio, Speech, and Music Processing (Y. Laprie)
  - IJDMMM International Journal of Data Mining, Modelling and Management (M. Cadot)
- Member of scientific commitee of conference :
  - ICSLP (J.P. Haton)
  - LREC (Y. Laprie)
  - JEP (I. Illina)
  - ISSP 2008 (Y. Laprie)
  - QDC 2008 (M. Cadot)
  - 6ème Atelier Fouille de données complexes dans un processus de d'extraction de connaissance (M. Cadot)
- 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 the ACOUSTICS'08 conference (C. Cerisara)
  - at TAIMA (Traitement et Analyse de l'Information : Méthodes et Applications[2]) Conference (K. Smaïl)
  - by the University of Annaba (K. Smaïl)
  - by the Blaise Pascal University of Clermont-Ferrand (A. Kipffer)
  - by the IUFM of Amiens (A. Kipffer)
  - by Telecom Paris (Y. Laprie)
  - at the CNRS summer school about medical imaging applied to speech production organized by A. Marchal from LPL (Y. Laprie)

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

8.2. Invited lectures

- Konstantin Markov has been an invited professor for a one-month stay in the team in November 2008,
- Jean Schoentgen, National Fund for Scientific Research, Belgium
- Philippe Langlais - University of Montreal, Canada
- Nathalie Camelin,University of Avignon
- Imen Jmeaa, ENIT, Tunisia
- Olivier Pietquin, SUPELEC

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;
- Head of MIAGe department (K. Smaïli),
- Head of Networking Speciality of University Henri Poincaré Master of Computer Science (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;
- Y. Laprie is the co-chair of the 8th International Seminar on Speech Production (ISSP 2008) in Strasbourg. INRIA Lorraine is one of the sponsors of this conference. S. Ouni is the organizing chair of ISSP.
- 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] F. BIMBOT, M. EL-BÈZE, S. IGOUNET, M. JARDINO, K. SMAÏLI, I. ZITOUNI. *An alternative scheme for perplexity estimation and its assessment for the evaluation of language models*, in "Computer Speech and Language", vol. 15, n^{o} 1, Jan 2001, p. 1-13.

[2] 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.

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

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

[5] 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.

[6] 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/.

[7] D. LANGLOIS, A. BRUN, K. SMAÏLI, J.-P. HATON. *Événements impossibles en modélisation stochastique du langage*, in "Traitement Automatique des Langues", vol. 44, n^{o} 1, Jul 2003, p. 33-61.

[8] 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/.

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

[10] 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

[11] 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/.

[12] C. CERISARA. *Automatic discovery of topics and acoustic morphemes from speech*, in "Computer Speech and Language", 2008, http://hal.inria.fr/inria-00330698/en/.

[13] S. DEMANGE, C. CERISARA, J.-P. HATON. *Missing data mask estimation with frequency and temporal dependencies*, in "Computer Speech & Language / Computer Speech and Language", 2009, http://hal.inria.fr/inria-00338397/en/.

[14] B. POTARD, Y. LAPRIE, S. OUNI. *Incorporation of phonetic constraints in acoustic-to-articulatory inversion*, in "The Journal of the Acoustical Society of America", vol. 123, 2008, p. 2310-2323, http://hal.inria.fr/inria-00112226/en/.

[15] 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: An Interdisciplinary Journal", 2008, XX, http://hal.archives-ouvertes.fr/hal-00329672/en/.

International Peer-Reviewed Conference/Proceedings

[16] G. BOUSELMI, D. FOHR, I. ILLINA. *Multi-Accent and Accent-Independent Non-Native Speech Recognition*, in "INTERSPEECH, Australie Brisbane", 2008, http://hal.inria.fr/inria-00327709/en/.

[17] 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", 2008, http://hal.inria.fr/inria-00337067/en/.

[18] J. CAI, G. BOUSELMI, D. FOHR, Y. LAPRIE. *Dynamic Gaussian Selection Technique for Speeding Up HMM-Based Continuous Speech Recognition*, in "ICASSP, États-Unis d'Amérique Las Vegas", 2008, http://hal.inria.fr/inria-00327703/en/.

[19] J. CAI, J. FELDMAR, Y. LAPRIE, D. FOHR, J.-P. HATON. *Transcribing Southern Min Speech Corpora with a Web-Based Language Learning System*, in "International Conference on Audio, Language and Image Processing - ICALIP 2008, Chine Shangai", IEEE, 2008, http://hal.inria.fr/inria-00336375/en/.

[20] C. CERISARA. *Exploiting confidence measures for missing data speech recognition*, in "Proceedings on Acoustics'08, France Paris", 2008, http://hal.inria.fr/inria-00330726/en/.

[21] P. KRAL, T. PAVELKA, C. CERISARA. *Evaluation of dialogue act recognition approaches*, in "IEEE International Workshop on Machine Learning for Signal Processing, Mexique Cancun", 2008, http://hal.inria.fr/inria-00330719/en/.

[22] Y. LAPRIE, P. MARAGOS, J. SCHOENTGEN. *How can acoustic-to-articulatory maps be constrained?*, in "16th European Signal Processing Conference - EUSIPCO 2008, Suisse Lausanne", 2008, http://hal.inria.fr/inria-00335958/en/.

[23] C. LAVECCHIA, D. LANGLOIS, K. SMAÏLI. *Discovering Phrases in Machine Translation by Simulated Annealing*, in "INTERSPEECH 2008, Australie Brisbane", 2008, http://hal.inria.fr/inria-00331327/en/.

[24] C. LAVECCHIA, D. LANGLOIS, K. SMAÏLI. *Phrase-Based Machine Translation based on Simulated Annealing*, in "Sixth international conference on Language Resources and Evaluation - LREC 2008, Maroc Marrakech", 2008, http://hal.inria.fr/inria-00285277/en/.

[25] D. MASSARO, S. BIGLER, T. CHEN, M. PERLMAN, S. OUNI. *Pronunciation Training: The Role of Eye and Ear*, in "Interspeech 2008, Australie Brisbane", ISCA (editor), 2008-09-22, p. FriSe3.O4-2, http://hal.archives-ouvertes.fr/hal-00327687/en/.

[26] S. OUNI. *Aspects of Pharyngealized Phonemes in Arabic Using Articulography*, in "Interspeech, Australie Brisbane", ISCA (editor), 2008-09-23, p. FriSe3.P4-6, http://hal.archives-ouvertes.fr/hal-00327676/en/.

[27] M. PIAT, D. FOHR, I. ILLINA. *Foreign accent identification based on prosodic parameters*, in "INTERSPEECH, Australie Brisbane", 2008, http://hal.inria.fr/inria-00327706/en/.

[28] B. POTARD, Y. LAPRIE. *Improving the Sampling of the Null Space of the Acoustic-to-Articulatory Mapping*, in "The eighth International Seminar on Speech Production - ISSP'08, France Strasbourg", 2008, http://hal.inria.fr/inria-00336381/en/.

[29] J. RAZIK, O. MELLA, D. FOHR, J.-P. HATON. *Comprehension Improvement using Local Confidence Measure: Towards Automatic Transcription for Classroom*, in "Workshop on Child, Computer and Interaction WOCCI08, ICMI'08 post-conference workshop, Grèce Chania", 2008, 5, http://hal.inria.fr/inria-00335558/en/.

[30] J. RAZIK, O. MELLA, D. FOHR, J.-P. HATON. *Frame-Synchronous and Local Confidence Measures for on-the-fly Automatic Speech Recognition.*, in "InterSpeech, Australie Brisbane", 2008, http://hal.inria.fr/inria-00325519/en/.

[31] V. ROBERT, J. FELDMAR, Y. LAPRIE. *Comparison between two predicting methods of labial coarticulation*, in "The eighth International Seminar on Speech Production - ISSP'08, France Strasbourg", INRIA (editor), 2008, http://hal.inria.fr/inria-00336382/en/.

[32] 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, France Strasbourg", INRIA, 2008, http://hal.inria.fr/inria-00336380/en/.

[33] C. ZAKARIA, O. CURÉ, K. SMAÏLI. *Conflict Ontology Enrichment Based on Triggers*, in "The 2nd International workshop on Ontologies and Information Systems for the Semantic Web, États-Unis d'Amérique Napa Valley, California", ACM 17th Conference on Information and Knowledge Management, 2008, http://hal.inria.fr/inria-00315300/en/.

National Peer-Reviewed Conference/Proceedings

[34] C. LAVECCHIA, K. SMAÏLI, D. LANGLOIS. *Une alternative aux modèles de traduction statistique d'IBM : Les triggers inter-langues*, in "15eme conférence sur le Traitement Automatique des Langues Naturelles - TALN'08, France Avignon", 2008, http://hal.inria.fr/inria-00285275/en/.

[35] M. PIAT, D. FOHR, I. ILLINA. *Identification de l'origine des locuteurs non natifs en utilisant des paramètres prosodiques*, in "XXVIIèmes Journées d'Étude sur la Parole - JEP 08, France Avignon", AFCP (editor), 2008, http://hal.inria.fr/inria-00327696/en/.

[36] J. RAZIK, O. MELLA, D. FOHR, J.-P. HATON. *Mesures de confiance locales et trame-synchrones*, in "XXVIIèmes Journées d'Etude sur la Parole - JEP 2008, France Avignon", 2008, http://hal.inria.fr/inria-00289905/en/.

[37] J. RAZIK, O. MELLA, D. FOHR, J.-P. HATON. *Transcription automatique pour malentendants : amélioration à l'aide de mesures de confiance locales*, in "XXVIIèmes Journées d'Etude sur la Parole - JEP 2008, France Avignon", 2008, http://hal.inria.fr/inria-00289904/en/.

References in notes

[38] 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.

[39] A. BONNEAU, L. DJEZZAR, Y. LAPRIE. *Perception of the Place of Articulation of French Stop Bursts*, in "Journal of the Acoustical Society of America", vol. 100, n^{o} 1, Jul 1996, p. 555-564.

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

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

[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] 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, 2009, p. 25-41, http://hal.inria.fr/inria-00338397/en/.

[45] ETSI ES 202 050 v1.1.1. *Distributed speech recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms*, 2002.

[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] 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.*, 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., ARNT - Lille, 2006, http://hal.inria.fr/inria-00185312/en/.

[49] 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/.

[50] 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.

[51] 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.

[52] P. KOEHN. *Europarl: A Multilingual Corpus for Evaluation of Machine Translation*, in "MT Summit, Thailand", 2005.

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

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

[55] C. LAVECCHIA, K. SMAÏLI, D. LANGLOIS. *Building Parallel Corpora from Movies*, in "The 4th International Workshop on Natural Language Processing and Cognitive Science - NLPCS 2007, Funchal, Madeira/Portugal", 06 2007, http://hal.inria.fr/inria-00155787/en/.

[56] 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/.

[57] 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.

[58] 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.

[59] S. RAYBAUD, C. LAVECCHIA, D. LANGLOIS, H. SMAÏLI. *New confidence measures for statistical machine translation*, in "Proceedings of the International Conference on Agents and Artificial Intelligence, ICAART 2009.".

[60] 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/.

[61] 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".