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

1. Team

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/.

Head of project-team

Yves Laprie [Research scientist HDR, CNRS, HdR]

Administrative Assistant

Martine Kuhlmann [CNRS]

CNRS Research scientist

Anne Bonneau [Research scientist]
Christophe Cerisara [Research scientist]
Dominique Fohr [Research scientist]

Faculty member

Armelle Brun [Assistant Professor, Nancy 2 University]
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 working for INRIA until 2006, September, 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]

Phd Students

Ghazi Bouselmi [TA, ATER since september 2007, thesis to be defended in 2008]
Sébastien Demange [TA, thesis defended on 8th November 2007]
Emmanuel Didiot [CIFRE grant, ATER since november 2006, thesis defended on 13th November 2007]
Guillaume Henry [TA, until August 2007]
Pavel Král [Czech coPhD, thesis defended on 12th November 2007]
Blaise Potard [MENRT grant, thesis to be defended in 2007]
Joseph Razik [INRIA grant, ATER since october 2006, defended on 9th October 2007]
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, thesis to be defended in 2010]

Project technical staff

Jonathan Ponroy [INRIA, CPER, 50% (collaboration with Magrit) since 1st February]
Alexandre Lafosse [INRIA (associate engineer)]
Julien Maire [CNRS, until 30th January]
Benjamin Husson [CNRS and then INRIA, since 1st February]
Viet-Bac Le [CNRS, until 28th February]

Invited professors

Jun Cai [Associate Professor, Xiamen University, since 13th March]
Jean Schoentgen [CNRS, 15th October-15th December]

Specialist engineer

Jacques Feldmar [Specialist engineer]

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 two topics:

- **Speech analysis.** Our works are concerned with automatic extraction and perception of acoustic cues, acoustic-to-articulatory inversion and speech analysis. 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, automatic speech to speech translation, text-to-speech alignment and audio indexing.

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 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, and in a regional project with teachers of foreign languages 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*.

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

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

[1] Hidden Markov Models

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

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

These two approaches are complementary; the latter borrows theoretical results on speech from the former, which, in its turn, borrows some numerical methods. Spectral analysis methods are undoubtedly the domain where exchanges are most marked. The simultaneous existence of these two approaches is one of the particularities of speech research conducted in Nancy and we intend to enhance exchanges between them. These exchanges will probably grow in number 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

3.2.1.1. Automatic detection of selective 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 [53], 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.

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.

3.2.1.2. Automatic detection of "well realized" sounds

The detection of selective cues confirms that a same sound, depending on its realization, can be identified with a very different level of confidence. Sounds that are identified with confidence are probably well realized and clearly pronounced. We made the hypothesis that the enhancement of well realized sounds in a sentence gives

listeners some islands of confidence during the acoustic decoding stage and improves speech intelligibility. Previous studies have shown that such an enhancement as well as the slowing down of some classes of sounds (fricatives and stops, in particular) improve the perception of a second language as well as that of the first language for hearing impaired people.

But the detection of these well realized sounds in an automatic manner is not obvious. It is possible to find well realized features with a speech recognition system based upon phonetic knowledge, through the use of "selective cues". But this method cannot be entirely automatic, especially due to segmentation problems. Stochastic methods, such as Hidden Markov Models (HMM), can recognize sentences in an entirely automatic way. But, if these systems obtained very high overall recognition scores, they do not give any indication about the way one sound in particular has been realized.

To solve this problem, we made the hypothesis that systematically well identified sounds are also well realized sounds and we forced HMMs to model those well identified sounds in the following way. First, on a training corpus, the system models the phonemes. Then, after a recognition test on the training corpus, the well identified sounds are set apart, and the system is trained to recognize these sounds. After three or four iterations, the system learns to recognize only systematically well identified sounds. First results with stop consonants show that the “well realized” models of sounds have high firing rate (about 30-60%, depending on the class) and make fewer errors.

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. Speech signal transformation

In order to improve oral comprehension, we use PSOLA (Pitch Synchronous Overlap and Add), a speech signal transformation method. PSOLA is based on the decomposition of the speech signal into overlapping pitch synchronous frames. Signal modifications consist in manipulating analysis marks to generate new synthesis marks. PSOLA is well known for its easy implementation and the quality of the slowed down signals. However, temporal discrepancies may appear in the region of the synthesis marks and PSOLA may generate noise between harmonics. In order to reduce the loss of quality, the method was improved in the two following ways. First, we have proposed a pruning algorithm to seek analysis marks (for pitch synchronization). It increased the robustness of pitch marking for speech segments with strong formant variation. Second, we improved the localization of analysis and synthesis marks. During the analysis stage, we can either oversample the signal or use F0 detection algorithm which gives an accuracy better than one sample. During the synthesis stage, the improvement is based on a dynamical re-sampling of the speech signal so as to accurately replace the frame on synthesis marks. Both improvements strongly reduced the level of noise between harmonics and we obtained a speech signal of high quality [59].

3.2.2.2. 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 starting a study on the 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.

3.2.2.3. Talking head

The aim of this project, which started within the RIAM framework (LABIAO, led by EDF), is to provide hard of hearing (HOH) people with an artificial talking head piloted by automatic speech recognition (ASR). The talking head can be lip read and, optionally, can produce cued speech (including a disambiguating hand). Textual subtitles can also be presented to the user. The "talking head" project is developed under Jacques Feldmar and Yves Laprie's supervision.

3.2.2.4. 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 [66], [67] 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 [78].

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 [73].

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 [21].

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 [62] 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 [52] advocates for the theory of expansion movements: the movement tends to be anticipated when no phonological constraint is imposed on labiality. Cohen and Massaro [57] 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, one requirement of missing data recognition is to build masks in a frequency-like domain. Furthermore, depending on the marginalization technique, different properties of the time-frequency feature domain are required. Hence, we have developed two additional feature domains: the first one is the simple Mel-scale filterbank energies, and the second one, called "Frequency filtered coefficients", decorrelates the frequency coefficients to justify the use of diagonal covariance marginalization approaches. Both these feature domains are exploited in the context of missing data recognition. 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 [61], 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) Maximum Likelihood Linear Regression adapts the acoustic models to noisy conditions or to a new speaker in the cepstral domain. The method estimates the linear regression parameters associated with Gaussian distributions of the models. The Maximum Likelihood criterion is used for the estimation of the regression parameters.
- MAP and MAPLR (Maximum A Posteriori - Linear Regression) This adaptation is based on Maximum A Posteriori training of HMM parameters, which uses some data from the target condition. This approach uses both the adaptation data and the prior information. The flexibility in incorporating the prior information makes MAP efficient for handling the sparse training data problem.
- PMC (Parallel Model Combination) is an algorithm to adapt the clean speech models to a noisy environment. It basically converts the models back to the power-spectral domain where speech and noise are assumed to be additive. Unlike the two previous methods, it does not require a large amount of adaptation data - about one second speech signal is enough to estimate the noise model.
- CMN (Cepstral Mean Normalization) is an algorithm to compensate for channel mismatch (differences in microphones for example). It is quite effective and very simple to implement, which explains why it is now used in nearly every recognition system.

- Spectral Subtraction subtracts a noise estimated from the incoming signal in the power spectral domain. This "denoising" algorithm is not extremely efficient when used as a pre-processor to a recognition engine.
- Jacobian Adaptation is a linear version of PMC that operates only in the features domain. It is one of the fastest model adaptation algorithms. The original models do not need to be trained in a clean environment. The method works actually better when the models are already slightly noisy.

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.

3.3.2. Language modeling

Apart from the challenges related to acoustic modeling (see preceding section), some problems due to the complexity of natural language remain without any satisfactory solution. State-of-the-art language models, as n-gram models, are very simple and reach relatively high performance in ASR. Such characteristics explain their predominance in recognition systems. n-gram models assign a probability to the current word, using the only $n-1$ preceding words. For example, given the beginning of sentence "*les pommes que j'ai*", a 4-gram model considers only the three previous words "*que j'ai*" to compute the probability of the current word. Due to long distance dependencies, such a model cannot be efficient. However, it is impossible to systematically increase the value of n due to computational constraints and probability reliance.

Our group, as other groups through the world, makes an increased effort in order to design more efficient language models. Most of language models we propose rest on information theory and statistics, some of them include linguistic knowledge to improve statistical modelization.

Combining several language models increases performance but, unfortunately, this is not sufficient. That is why we are still developing new methods to deal with the complexity of language modeling. One way to reach this objective is to deal with a larger history and to consider words as complex units. Henceforth, we redefine a word as a compound entity represented by a list of features including the orthographic form of the word. To do so, we work in several directions, which can be clustered into two main parts:

- Retrieving useful information in history, and designing *ad-hoc* models:
  - **Feature vector models**. A word is considered not only as an orthographic form but as a complex unit which contains a class tag, a gender and number features, a semantic tag... This makes a statistical language model more realistic and in harmony with linguistic theory. Some tracks were explored and confirmed the feasibility of this approach [60].
  - **Phrase-based models**. Another way to deal with complex linguistic units consists in using phrase-based models [80]. The idea is to include phrases in the dictionary. These phrases bring up more information for the decoding process. In our works, we retrieve automatically the useful phrases from large corpora.
  - **Language model adaptation using topic identification**. The objective is twice: firstly we want to find out the topic of the uttered sentences. Secondly, we want to adapt the baseline language model using the one that corresponds to the retrieved topic. Our researches concerns both identification and adaptation [51].

  - **Dealing with agreement problem in speech recognition** [77]. We introduced an original model called Features-Cache (FC) to estimate the gender and the number of the word to predict. It is a dynamic variable-length Features-Cache. The size of the FC window is determined in accordance to syntactic delimiters. However, this model does not need any syntactic parsing; we use it as any other statistical language model.

- Efficiently combining all models:
  - **Selecting the best language model in accordance to the history**. Combining language models is not sufficient to deal with the complexity of natural language. The best way to improve the performance of a speech recognition or translation system is to select dynamically the best language model depending on a history as in the SHP principle [71].
  - **Bayesian Networks**. Bayesian network is a powerful formalism that modelizes the relationship between several events. We exploit this concept in order to construct a more consistent combination of language knowledge. We have developed a unifying approach that processes each bit of knowledge in a unique model and constructs new data-driven language models with improved performances. The principle of this approach is to construct Dynamic Bayesian Networks (DBNs) in which a variable (word, class or any other linguistic unit) may depend on a set of context variables. The details and evaluation of this approach using several datasets are reported in [60].
  - **Crossing context n-grams.** We studied distant language models in the scope of using them in an intelligent prediction framework (non distant prediction) [56]. Classical models are, in some cases, unable to predict the correct contiguous word. However, classical bigram models may be able to predict distant words. Our works let us think that this behavior depends on history.

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 speech signal into another speech signal. The input signal is recorded in a Source language and the output is provided in a Target language.

Our purpose is to build a complete speech-to-speech translation system based on phrases.

This issue deals with several skills in our group: speech recognition for transcription, language modeling for outputting well formed sentences in Target language, and synthesis for producing final speech signal. Therefore, numerous works in our group deal with Speech-to-Speech Translation.

Moreover, Speech-to-Speech Translation requires a translation model. A translation model describes how to translate a sentence from Source language to Target language. In this scope, we used our experience in statistical language modeling (see section 3.3.2). We proposed the notion of Inter-Lingual Triggers, adapted from the so-called triggers. This allows automatically detecting the connections between linguistic elements in Source and Target sentences. These elements may be words, linguistic features, phrases, etc. Moreover, in accordance with the research community, we developed methods in order to detect bilingual phrases.

In sake of comparison, we implemented both models 1 and 2 proposed by Brown in [55] to train a baseline translation component.

Another important issue in our work is that training machine translation systems requires a huge quantity of bilingual aligned corpora. Even if this kind of corpora becomes increasingly available, there may be a coverage problem for a specific need. Several French-English applications use either the Canadian Hansard corpus or corpora extracted from the proceedings of European Parliament [69]. One way to enrich the existing parallel corpora is to catch the important amount of free available movie subtitles. Several websites (http://divxsubtitles.net) provide files used for subtitling movies. Aligned movie subtitles corpora may be

useful for realistic machine translation applications because the raw data may contain formal, informal, vulgar or coarse words, common expressions, hesitations, etc. For one movie, two subtitle files for two different languages are not necessary aligned because different human translators independently make the different files. The raw subtitle corpora cannot be used without pre-processing. In order to make these files convenient for use, it is first necessary to align bilingual versions of the same movie at paragraph, sentence or phrase level. In this scope, we automatically built such a corpus for movies subtitles. In our method, the alignment between subtitles is handled by Viterbi algorithm [42]. We consider this work as a first stage towards a real time subtitling machine translation system.

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 [64] 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. Snorri and 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 [72]. 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 [40].

This year we added a phase unwrapping algorithm which will be exploited to detec glottal closure instants in the future. The current version of WinSnoori is available on http://www.winsnoori.fr.

5.1.2. PhonoLor

PhonoLor is a phonetizer enabling word translations into a sequence of phonemes. This software exploits phonetization rules by a decision tree automatically learnt from a corpus of examples. The phonetizer is used in the TTS platform currently under development (see 6.1.3).

5.1.3. 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 [70]. That's why we began to develop another tagger based on a unsupervised tagging algorithm.

The TTS platform currently under development (see 6.1.3) uses the latter tagger with the class set of the former tagger.

5.1.4. 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.

5.1.5. SALT (Semi-Automatic Labelling Tool)

Given speech signal and the orthographic transcription of a sentence, this labelling tool provides a sequence of phonetic labels with associated begin-end boundaries. It is composed of two main parts: a phonetic transcription generator and an alignment program. The phonetic transcription generator provides a graph of a great number of potential phonetic realizations from the orthographic transcription of a sentence. The second part of the labelling tool performs a forced alignment between all the different paths of the phonetic graph and the speech signal. The path giving the best alignment score is accepted as the labelling result.

5.1.6. LIPS (Logiciel Interactif de Post-Synchronisation)

The lipsync process or post-synchronization is a step in the animation production pipelines of 2D and 3D cartoons. It consists in generating the mouth positions of a cartoon character from the dialogue recorded by an actor. The result of this step is a sequence of time stamps which indicate the series of mouth shapes to be drawn. Until now, the lipsync phase has been done by hand: experts listen to the audio tape and write mouth shapes and their timing on an exposure sheet. This traditional method is tedious and time consuming. LIPS (lipsync interactive software) is a tool that, from the speech signal and the orthographic transcription of a dialogue, semi-automatically generates the series of mouth shapes to be drawn. LIPS performs the post-synchronization for French and English cartoons.

5.1.7. 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.

5.1.8. SELORIA

SELORIA is a toolbox for speaker diarization [44] .

The system contains the following steps:

- Speech activity detection: an audio stream may consist of some acoustic activities like speech, noise, music, background conversation, advertisement. Therefore, non-speech regions must be detected and removed from the audio stream.
- Speaker change detection: inside every speech region, a speaker change (or speaker turn) detector is used to find points in the audio stream which are candidates for speaker change points. To do this, 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 if its distance value is higher than a predefined threshold.
- 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. The BIC threshold value must be tuned on a development corpus in order to reduce the number of false alarms without increasing the number of new missed detection errors.
- 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.

5.1.9. 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.

5.1.10. BNTK

The Bayesian Network ToolKit (BNTK) is an open-source toolkit for developing and testing Bayesian networks. It is written in C++. It supports multidimensional continuous and discrete random variables. Continuous variables are assumed to be linear conditional Gaussians, and cannot be parent-nodes of discrete variables. Both inference and training steps of the network parameters were implemented. Exact inference is based on the junction tree and message passing algorithms. Training can only be realized for now in the complete case.

The objective of this toolkit is to help researchers to quickly implement, train and test the graphical models they may need for their research. With this toolkit, they can thus compare different sets of variables and network topologies, and choose the best one for their problem. Then, they can implement their own optimized algorithms for the chosen network topology. This toolkit is quite general and can be used for a wide range of research areas, but our primary goal is to use it for automatic speech recognition. It is distributed under the LGPL license on the GForge INRIA Web Site.

5.1.11. HTK-compliant recognition tools

HTK is a widely used standard toolkit to train HMMs. For example, the Julius recognition engine, which is used in our broadcast news, OZONE and Amigo platforms, exploits HTK acoustic models. We have developed our own set of additional recognition tools that support this format, and that can interface with HTK. These tools are described next:

- The HMM parallel training tool distributes the training process over the 25 computers of the PAROLE PC cluster.
- The HMMModelConv toolkit is an extensible software that converts acoustic models between different formats: HTK, ESPERE and Sphinx3.
- The stochastic speech library (old GMMlib) is a JAVA library that integrates most of our recent work on stochastic speech processing. It is highly modular, and supports JUnit testing for most of its functionalities, as well as non-regression tests for speech recognition on standard databases (Aurora4 for now). Its most visible functionalities include:
  - Support load and save formats for HTK models, HTK parameter files and HTK label files.
  - Support visualization and tagging of speech spectrograms, parameter files, missing data masks and label files.
  - It is based on the stochastic processing “pull” paradigm, which allows to easily plug-in and chain several processing modules, while ensuring stream synchronization in the case of lattice-like modules chains.
  - Support advanced Gaussian Mixture Models training and editing (LBG, cross-correlation, marginalization, etc.).

- Include beta modules for Bayesian denoising of speech signals.
- Include beta modules for accuracy-based missing data recognition masks inference and training (see section 6.2.1).

We plan to integrate on a regular basis our old and new research algorithms about robust speech recognition into this toolkit.

5.1.12. 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 (see section 7.3.1).

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, Jean-Paul Haton, Yves Laprie, Joseph di Martino, Slim Ouni, Jacques Feldmar, Agnès Kipffer, Martine Cadot, Guillaume Henry, Blaise Potard, Vincent Robert, Alexandre Lafosse, Jonathan Ponroy, Benjamin Husson.

6.1.1. Acoustic-to-articulatory inversion

The strength of our inversion method lies on the quasi-uniform acoustic resolution of the articulatory table. The originality is based on the generation method that evaluates the linearity of the articulatory-to-acoustic mapping at each step. Articulatory parameters of Maeda's model vary between -3 σ and 3 σ where σ is the standard deviation. Thus, the codebook inscribes a root hypercube. Sampling the articulatory space amounts to finding reference points that limit linear regions. The inversion procedure then retrieves articulatory vectors corresponding to acoustic entries from the hypercube codebook. A non-linear smoothing algorithm together with a regularization technique is then used to recover the best articulatory trajectory. The inversion ensures that retrieved articulatory parameters produce original formant trajectories accurately and a realistic sequence of the vocal tract shapes [75].

The articulatory codebook is a key component of our analysis-by-synthesis inversion method since it represents the synthesis facet of the algorithm. Therefore it needs to be compact while offering a very good acoustic precision. This year we improved the construction of codebooks by using a representation of the articulatory-to-acoustic mapping more general than that presented in [75]. The mapping is approximated by multivariable polynomials [47]. The second major improvement concerns the subdivision process which finds out the most efficient subdivision of the articulatory space. In the previous version the subdivision was applied to all the parameters as soon as the linearity test failed. This thus gave rise to 2^7 new cubes at each subdivision. In order to minimize the size of the codebook while guarantying a very good acoustic precision the subdivision is now applied only on the parameters which gives rise to nonlinearities. For now, a simple heuristic is being used, which tries to find the direction in which the irregularity appears to be the strongest: for all directions, i.e. for all the articulatory parameters,, a score is computed from the acoustic distance between synthesized vectors and the approximation obtained using Taylor's approximation in the center.

Experiments carried out show that the size of the codebook can be divided by a factor of 20, and simultaneously, the acoustic precision can improved by a factor of 2 by using second order polynomials together with this new construction strategy.

This year, we complemented the investigation of the phonetic constraints. First, we checked the consistency of phonetic constraints. Indeed each constraint applies for a specific vowel, or equivalently to a specific region in the formant space, which thus has to be found beforehand. Even if these regions do not need to be known very precisely there is a risk of mismatching constraints, i.e. applying a constraint dedicated to a given vowel to invert acoustic parameters corresponding to a very different vowel from an articulatory point of view. Experiments showed that phonetic constraints are specific to vowels they were designed for [20]. Based on this observation we evaluated a strategy for triggering constraints which does not require the prior vowel identification. Indeed, if the articulatory domains and constraints are correctly defined, the relevant constraint, i.e. the one to use, is the one which yields the highest average phonetic score over all inverted VT shapes from one 3-tuple of formants.

We also started works about the inversion of fricatives sounds. This comprises the construction of an adapted codebook for fricatives what requires using two extra parameters, one for specifying the location of the frication source downstream from the constriction, and the second the amplitude of the frication. Lastly, the inversion framework has been substantially improved by adding the possibility of using scripts to control inversion experiments.

6.1.2. Talking head

6.1.2.1. Coarticulation

We implemented Cohen and Massaro's coarticulation algorithm which utilizes dominance function and articulatory targets learnt from the corpus acquired through the stereovision tracking system previously developed by the Magrit EPI. This coarticulation is used to pilot lip movements of the talking head. A dominance function is attached to each viseme. The learning consisted in learning five parameters for each viseme from the labial corpus. Tests realized on a test corpus gave a correlation of 0.88 with original data and a root mean square error of 8.54% what is better than comparable implementations, in particular with respect to the correlation criterion what means that parameters variations are well taken into account.

We also tested the concatenative algorithm which tries to predict coarticulation from small segments of the labial corpus [49]. Coarticulation transitions are represented in the form of sigmoïd curves which can be easily concatenated. Since this 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. Since these transitions are influenced by their phonetic and prosodic contexts we added a registration algorithm that reduces jumps between neighbouring sigmoïds. First tests carried out showed not as good a correlation (0.73) as that obtained through Cohen and Massaro's algorithm and a RMSE of 10%. However, results could be improved in two directions. The first is lip aperture dynamics which has not been captured correctly by sigmoïds. The second concerns the completions of transitions which have not been observed. Indeed the choice of transitions used could be improved by using more appropriate distance criteria.

6.1.2.2. Talking head for people with special needs

A statistical and a case study analysis of the current version of the talking were conducted in collaboration with the INSHEA (Suresnes, ex CNEFEI [35]) and with two students from the Speech Therapy School of Nancy (Lorène Mourot and Marie Rovel) [50].

The main conclusion is that, even though messages can be understood, the involved cognitive effort is too important to make it usable in the classroom as it is today. A better understanding of cued speech production and perception resulted from these studies and a number of recommendations were provided. They have given new research leads:

- A robust speech/non speech detector has been developed based on GMMs (Benjamin Husson and Jonathan Ponroy).

- A coarticulation prediction algorithm has been integrated.
- Prosody elements such as fundamental frequency (F0), energy and speech rate can now be computed in real time and will be used to integrate some visual prosody cues.
- A highly realistic geometry and texture is being built in collaboration by the Magrite Team (Marie-Odile Berger, Jonathan Ponroy)

New talking heads are being integrated with our ASR:

- Synface (KTH, Sweden) shows articulators such as tongue and jaw (Muscle European Network of Excellence)
- Baldi, (UCSC, USA) shows some basic visual prosody (Slim Ouni)
- MyMultimediaWorld (INT, Evry) makes our talking head compatible with the Mpeg4 format
- Greta (Linc, Montreuil) shows facial expression but requires manual XML tagging of the ASR output.

The notion of "cognitive effort for understanding" does not have an accurate definition even though it is central for quality assessment of our project. We are working on trying to better understand this notion and design experiments to get qualitative and quantitative measurement of it:

- Lexical access computation models have been investigated and Electro-Encephalogram (EEG) experiments have been conducted in collaboration with the Cortex team (Laurent Bougrain, Louis Mayaud, Enrique Sidhoum) [24]
- Different combinations of textual orthographic and phonetic transcriptions of the ASR output have been evaluated, making use of confidence measures (Joseph Razik). Results are encouraging and could lead new applications, especially for the aging population.

Another work deals with technical assistance for people with special needs [34]. This work is not directly linked to speech processing yet. But it could be the beginning of such kind of technical help in which speech processing (as dialog interaction) could take place in the future.

6.1.2.3. Intelligibility of Visual speech

An important challenge for the talking head development is to evaluate the effectiveness in terms of the intelligibility of visible speech. In collaboration with Dominic Massaro's team (University of California at Santa Cruz), we introduced a new metric to describe the benefit provided by a synthetic animated face (a talking head) relative to the benefit provided by a natural face [19]. This metric allows direct comparisons across different experiments and gives measures of the benefit of a synthetic animated face relative to a natural face (or indeed any two conditions) and how this benefit varies as a function of the type of synthetic face, the test items (e.g., syllables versus sentences), different individuals, and applications.

This study and others showed that human talkers outperform synthetic talker in perceptual experiments. This is probably due to a lacking of finer modeling of the face or very likely because we still did not capture some important aspects of visual speech. For instance, some parts of the face that we do not consider relevant to visual speech or they are noticeably visible from outside, might actually provide some information which human are capable of decoding. This was the outcome of a perceptual study that we performed for Arabic. We showed that pharyngeal phonemes in Arabic, not known as visible from outside, provide actually some information to perceivers [45], [46]. This result will be extended to other languages.

6.1.3. Text-to-Speech synthesis

We are developing (since September 2006) a software platform to Text-To-Speech (TTS) synthesis. An INRIA associate engineer, Alexandre Lafosse, has been hired on this project. The aim is to obtain a TTS system as a toolkit for speech recognition applications (with Natural Language Processing of the platform) and for TTS system building.

6.1.3.1. Natural Language Processing

The NLP is the first step of a TTS system. This part tags and analyzes the input text to obtain a feature set (phoneme sequence, word grammar categories, syllables...). During this year, it is built on the one hand from tools which already exist in the PAROLE team (syntactic tagger, phonetizer, syllabification). A objective is to obtain a unified set of tools within the Parole team for NLP. For instance, this toolkit shall be re-used to train different recognition modules. On the other hand, some NLP modules had been specifically built for the TTS system : a chunker (or shallow parser) and a postphonetizer. The chunker splits up the sentence into word groups in according to the grammar category of each word. The chunkers generally are probabilistic or full-rule-based and can tag these groups in terms of nominal, verbal or prepositional phrases...Our developed algorithm is based on local rules: one by one, each word are or not added to the current group in according to the one or two previous words. For our system, we don't need to tag these groups. We will consider these groups as support marks of the rhythm (as breath groups). This information is required for the syllabification and the postphonetizer to take in count, for instance, the liaison phenomena between two words. The postphonetizer has been built from rules to correct the phonetization (which it was made word by word). The liaison, elision, and denasalization phenomena are performed in this module.

All modules of the NLP are built, however some improvements and global evaluations remain to do.

Most modules required lexicons, dictionaries or training corpus. That's why, we have proposed to phonetize the "Morphalou 2" corpus from ATILF laboratory (Analyse et Traitement Informatique de la Langue Française). This project has been supported by a CNRTL grant during 3 months (Centre National de Ressources Textuelles et Lexicales). The corpus contains about 640000 forms (nouns, adjectives, conjugated verbs, ...). We have phonetized these forms with two automatic phonetizers ("LIA Phon" and our phonetizer) and we have corrected the forms for which we have obtained two different phonetizations. The corrections had been semi-automatic (automatic scripts or by hands). Some other checks and corrections have been made on the whole corpus. This work will allow us to re-train our phonetizer based on decision tree algorithm.

6.1.3.2. Selection and corpus building

The second part of TTS systems is the Non Uniform Unit (NUU) selection module. From NLP information, the module selects, in a recorded corpus, units which can be diphone, syllable or part of words. The selection is based on a Viterbi algorithm. The module has been built this year. The originality of our approach is that the selection is piloted by linguistic features without acoustic model of prosody (see [58]).

Moreover, we need to record the selection corpus. The corpus generally is 3-5 hours of speech recorded by one speaker. A greedy algorithm development for the corpus building is in progress : the goal is to extract N textual sentences from M textual sentences with $N << M$ and 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 by taking in priority the sentence containing the rarest units/features ([63]). This iterative algorithm is used until all features are covered. We have added a level priority for units which can occur only one times in a sentence. We works about 1.500.000 sentences extracted from LE MONDE corpus and we obtain, with a preliminary feature set, 4.000 sentences which cover almost all diphones represented 3 times at 5 different positions in the sentence.

The next step will be the recording of these N sentences by a speaker. Tools for the automatic transcription alignment will also be developed from team speech recognition tools.

6.1.3.3.

The release of this platform should be at the end of the next 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 pilot a talking head. The aim also is to obtain a open-source platform (Java/C) under GPL licence for most of modules and tools.

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

We have tested the modules of the algorithm devoted to the correction of a learner's prosody, in particular the alignment module, as well as the localization of the lexical stress. The tests showed that the success of the method largely depends upon phonetic alignment. Let us recall that phonetic alignment is necessary to compare the realization of the learner with that of a reference. Although this algorithm makes relatively few errors, these errors lead to erroneous corrections, which are hardly acceptable in a speech tutor software. The solutions we plan consists in using results from non-native speech recognition and knowledge in acoustical decoding (such as landmarks or selective cues).

The localization of lexical stress is reliable. Indeed, differences between the automatic localization of the model's accent and the place indicated by an English dictionary are found for only 5% of the realizations. These errors were due to alignment problems, and to the absence of a marked F0 peak (cue considered as the most important for isolated words). This absence can be explained by the dual aspect of lexical stress in isolated words, which can be considered as a phrase accent or a lexical stress. The solution we propose here consists either in analyzing the respective weight of prosodic cues in case of an absence of a marked F0 peak [37].

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

The evaluation of phonemic discrimination has been based on the test specially made by [66] 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 [79].

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

6.1.6. Development of robust data mining for speech data

Our data are tried to be treated by current statistical tools, like other biological data, after adaptation if necessary [17].

But their specificity (huge amount of observations, and numerous variables, quantitative and qualitative with numerous modalities) invit us to use data mining methods. A robust data mining method, based upon randomization tests, is currently under development in our team [32], [31].

6.2. Automatic Speech Recognition

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

Participants: Christophe Antoine, Ghazi Bouselmi, Armelle Brun, Christophe Cerisara, Emmanuel Didiot, Dominique Fohr, Jean-Paul Haton, Irina Illina, Pavel Kral, David Langlois, Viet-Bac Le, Julien Maire, Odile Mella, Marina Piat, Joseph Razik, Kamel Smaïli.

The team members have published an extensive review about the current state of the art on automatic speech recognition [65]. The major contributions of the team to this state of the art are summarized in the next section.

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 recently published an in-depth review and discussion about the existing approaches that can be used to compute missing data masks [15]. In 2006, we proposed an original algorithm that exploits frequency and temporal dependencies to generate better masks in the spectral domain. In 2007, we reused these models and further improved them by proposing a new solution to design more accurate marginalisation intervals for missing data recognition. Classical marginalisation techniques used in missing data recognition either considers that, for masked coefficients, the energy of the speech is completely unknown, and thus marginalizes the log-likelihood from minus to plus infinity, or assumes that the speech and noise energies are additive in the power spectral domain, and then marginalizes the spectral features from zero to the observation. Both these marginalization ranges are very coarse, and we proposed an original solution to greatly reduce them. The basic principle consists in exploiting the normal distribution properties to identify a reasonable interval within which at least 95 % of the observations fit. The remaining 5 % are considered as outliers and are neglected. The resulting marginalisation intervals are much smaller, which decreases the approximation error achieved when integrating the missing data log-likelihoods.

This approach has been validated on the HIWIRE database [33]. These research technologies have also been transfered into the European HIWIRE project, and are described in details in the Ph.D. thesis of Sebastien Demange [10].

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 evaluated several adaptation methods for non-native speech recognition. We have tested pronunciation modelling, MLLR and MAP non-native pronunciation adaptation and HMM models retraining on the HIWIRE foreign accented English speech database. The "phonetic confusion" scheme we have developed consists in associating to each spoken phone several sequences of confused phones. In our experiments, we have used different combinations of acoustic models representing the canonical and the foreign pronunciations: spoken and native models, models adapted to the non-native accent with MAP and MLLR. The joint use of pronunciation modelling and acoustic adaptation led to further improvements in recognition accuracy. The best combination of the above mentioned techniques resulted in a relative word error reduction ranging from 46% to 71% [25], [26].

This year, we proposed automated approaches for text-independent foreign accent classification. Results of foreign accent classification task could be used for adapting acoustic models, modifying lexicon, changing language model with regards to non-native speakers. In our study, we investigate statistical approaches which differ from the a priori knowledge they need: GMM, which do not required neither phonetic knowledge nor labelling, phone recognition (without a lexicon), sentence recognition (with a lexicon and a grammar) and extraction of discriminative sequences of phonemes. This work is done in the framework of the HIWIRE European project. We evaluated the proposed approaches on English speech corpus pronounced by French, Italian and Greek speakers. The best classification rate (83.3%) is achieved using sentence recognition or a phone-based recognition followed by language modelling approach [27], [36].

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 5.1.9).

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 ANTS versions 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).

For the real time system, we have trained specific acoustic models with less free parameters. Moreover, the speaker clustering and the adaptation module have been removed because of time constraints. Finally, the beam search was narrowed.

6.2.2.2. Speech/music/advertisement segmentation

In the framework of the CIFRE PhD. of Emanuel Didiot with the TNS company, we have been continuing the implementation of an automatic system for keywords detection in broadcast news. We chose an approach based on a large vocabulary recognition system.

To avoid false keyword detection in audio segments containing only music, jingles or songs, we addressed the problem of speech/music/advertisement segmentation.

For the speech/music segmentation, we continued with a new parameterization based on wavelets. We studied different decompositions of the audio signal based on wavelets (Daubechie, Coiflets, symlets) which allow a better analysis of non stationary signals like speech or music. We computed different energy types in each frequency band. Our first results on an audio broadcast corpus gave significant improvement compared to classical MFCC features for music/non music segmention. The thesis of Emmanuel Didiot has been defended on November 13, 2007 [11].

6.2.2.3. Confidence measure

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 have 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 have 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 and in particular during the first decoding recognition process. In 2007, we studied several variants of these measures. Those using reverse bigram and trigram probalilities achieve better results using the Equal Error Rate (EER) criterion.

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.

These new measures were evaluated on an 1-hour real broadcast news corpus (ESTER) using EER)criterion in an automatic transcription task. They were also compared to a state-of-the-art measure, based on posterior probability but which requires the recognition of the whole sentence [48]. We achieved performance very close to this reference measure with our local measures and a delay of less than one second (23.2% EER vs. 22%). These results were confirmed on an 1-hour real broadcast news test corpus.

We then validated our confidence measures in three other applications. In 2006, we used them in an on-the-fly keyword spotting task in order to reduce the false-acceptation rate. In 2007,we integrated one of our frame-synchronous measures in the decoding process of the recognition engine and achieved to decrease the word error rate of the original system of around 6% in relative. Finally, we have shown that using one of our local confidence measure can increase the comprehension of hard of hearing children by highlighting words of low confidence in the transcription provided by a speech recognizer [13].

6.2.3. Ubiquitous speech recognition

Ubiquitous speech recognition deals with the application and adaptation of speech recognition techniques in the context of Ambient Intelligence platforms. In Ambient Intelligence, the main innovation concerning speech interactions is the concept of implicit speech interactions: traditional Human-Computer dialogs assume that the user is directly interacting with the system. Such speech interactions are explicit. On the contrary, when the user intention is not to communicate with the system, every sentence he/she says can be used as an implicit speech interaction by the system. This can happen for example when the user is talking to someone, in a meeting, in a classroom, or simply when he is listening to the radio. Ubiquitous speech recognition may be considered as a new application domain rather than a new fundamental research area.

Concretely, we are addressing this challenging topic in the context of three activities: The first one concerns an extensive study of the state-of-the-art about user interactions and Ambient Intelligence in the national OFTA group, which is reported in [16]. The second one relates to our contribution in the European project Amigo, described in section 7.4.2. The third one deals with our research work on automatic recognition of dialog acts. Indeed, one of the main feature of implicit speech interaction concerns the computation of non-lexical information from the speech stream, such as emotions or 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 this work is to automatically identify dialog acts from the user's speech signal. The work on automatic recognition of dialog acts has started in the end of 2003. Since the beginning of 2007, we have mainly focused our work on three aspects. The first one concerns the use of prosody: although prosody has already been used earlier, we have reconsidered this year its impact on dialogue acts, thanks to the new corpora available and to the better understanding we have gained over the last years about dialogue acts. This new analysis is reported in [39]. The second focus concerns the development and evaluation of two confidence measures for semi-automatic labeling of corpora into dialog acts, which resulted in a new conference paper [38]. The third aspect concerns the evaluation and comparison of two algorithms (multiscale position and non-linear merging) previously designed in 2006, and the proposition of a new technique based on bayesian modeling, all three methods aiming at integrating some global sentence structure information into the dialog act recognition process. This work has been reported in a journal paper [18]. More generally, all the work that has been realized about dialog act recognition is reported in Pavel Kral's Ph.D. thesis [12].

6.2.4. Language Modeling for Automatic Speech Recognition

We extended our work about crossing n-gram models (see section 3.3.2) by using trigram models and by integrating the models into a speech recognition system.

Experimental investigation of crossing left contexts of baseline and distant n-grams showed the feasability of this idea and its contribution in the improvement of perplexity. It outperforms the standard bigram and trigram models by respectively 14% and 5.6%. Its integration in a real speech recognition system achieved a slight improvement of the word error rate on broadcast news corpus [30].

6.3. Language Modeling for Speech-to-Speech Translation

This year, we intensified our efforts on developing our Speech-to-Speech translation system:

- We developed the notion of Inter-Lingual triggers. An Inter-Lingual trigger describes the fact that a linguistic element in a Source sentence triggers the occurrence of another linguistic element in the Target sentence.
- We exploited this formalism in order to build up a bilingual dictionary for machine translation [41]. We compared our dictionary to ELRA and a free-downloaded dictionaries. The experiments showed that the obtained dictionary is well constructed and is suitable for machine translation.
- We used our dictionary to estimate the translation parameters [43]. We introduced this translation model into the Pharaoh decoder [68]. We evaluated the quality of translation with the BLEU measure [76]. We compared the results with the ones given by GIZA++ [74] in the same conditions. Depending on corpora, our performances are closed to or better than GIZA++, Model 3.
- We developed a decoder dedicated to the using of Inter-Lingual triggers. This decoder is under experiment.

We conducted our experiments on two parallel corpora:

- the European Parliament corpus [69], which is made up of transcriptions of parliament sessions.
- movies subtitles: we automatically built this corpus by using an alignment method based on Viterbi [42] (see section 3.4).

6.4. Recommendation Systems

For more than a year, the way to use of state of the art Statistical Language Models (SLM) for usage-based recommender systems has been investigated. A comparative study has first been carried out, which led to demonstrate that both domains are highly similar. The integration of SLM-like models in usage-based recommender systems can thus be investigated further. In this work, the concept of Statistical Grammar of Usage (SGU) has been introduced. This new concept exploits sequences of consultations of resources and highly correlated resources couples (see [29], [28] and [54]).

This research which started one year ago, has been conducted by a member of the team who will integrate a new research team in early 2008.

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

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.

7.3. National Contracts

7.3.1. STORECO project

This project is funded by the RIAM, Réseau pour la Recherche et l'Innovation en Audiovisuel et Multimédia (network for research and innovation in audiovisual and multimedia). The aim of this project is to automate the making of close captions for TV programs. To do that, we will use algorithms developed for ASR.

We are involved in three main tasks:

- detection of speech segments (speech/music segmentation),
- automatic alignment between the text scripts and the audio files,
- detection of speaker turns, i.e., each time a speaker change occurs.

This year, to detect speaker turns, we have proposed a new speaker diarization method based on the Normalized Cross Likelihood Ratio method [44]. Moreover, we implemented this method into the system SELORIA : a speaker diarization tool.

7.3.2. LABIAO project

The LABIAO project started in January 2005. It is funded by the RIAM and is lead by EDF. The aim is to provide hard of hearing people with an artificial talking head, piloted by ASR, which can be lip read and, optionally, can produce cued speech (including a disambiguating hand). Our contribution has been threefold.

First, we have developed a new version of the Viterbi algorithm. It runs in real time and performs continuous phoneme recognition with a maximum delay of 1 second. Quality has also been improved. By handling triphones and syllables and by extracting features from the speech signal which are closer to human perception, the recognition rate has been increased by a 12 % factor.

Second, a realistic model of a talking head has been produced in collaboration with the Magrit Project. From a stereo corpus of a female speaker, a dense and realistic 3D mesh has been fitted using linear morphing and radial basis function deformations. Two coarticulation algorithms have been developed and evaluated: the first is derived from that of Cohen and Massaro, the second is a concatenative algorithm that represents transitions in the form of sigmoid functions.

Third, evaluation of the combination of speech recognition and talking head with the help of 35 young deaf people is under progress. In collaboration with the French Ministry of Education and the School of Speech Therapy of Nancy, different test protocols have been implemented. The aim is to be able to understand by the end of the project, in July 2007, the conditions in which the talking head is relevant and can be used, or, on the opposite, should not be used.

7.3.3. ST&TAP project

In the framework of *Technologies pour le handicap* (technologies for handicap) funded by the French research department, we are involved in the ST&TAP project. The objective of this project is to provide, nearly in real-time, close captions of TV broadcast news for deaf people. We investigate approaches coming from ASR that have the potential to improve the generation of close captions. Therefore, two tasks could be considered:

- when the newscaster reads the teleprompter, the software must perform an alignment between the text of the teleprompter and the audio signal to obtain the beginning and the end of each uttered word;
- when the newscaster improvises or during an interview, an ASR system will operate and the result will be manually corrected.

Last year, we developed an algorithm based on DTW (Dynamic Time Warping) to perform alignment between the recognized words and the teleprompter text. This year we adapted this algorithm to take into account the on-the-fly output of our large vocabulary speech recognizer.

7.3.4. NEOLOGOS project

The NEOLOGOS project results from a collaboration in the speech recognition field between French laboratories (IRISA, ENSSAT, LORIA) and industrial companies (TELISMA, ELDA, FRANCE TELECOM) and is funded by the French research ministry (CNRS-Technolangue).

The aim of NEOLOGOS is to create new kinds of speech databases. The first one is an extensive telephone database of children's voices, called PAIDAILOGOS. For this database, one thousand of different children will be recorded.

The second is an extensive telephone database of adult voices, called IDIOLOGOS.

The starting point of this work is to consider that the variability of speech can be decomposed along two axes, one of speaker-dependent variability and one of purely phonetic variability. The classical speech databases seek to provide a sufficient sampling of both variabilities by collecting few data over many random speakers (typically, several thousands). Conversely, Neologos proposes to optimize explicitly the coverage in terms of speaker variability, prior to extending the phonetic coverage by collecting a lot of data over a reduced number of reference speakers.

In this framework, the reference speakers should come out of a selection process which guarantees that their recorded voices are non-redundant but keep a balanced coverage of the voice space.

The extraction of the reference speakers has been interpreted as a *clustering task*, which consists in partitioning the voice space in homogeneous subspaces that can be abstracted by a single reference speaker. First, the academic partners and FRANCE TELECOM formulated this problem in a general framework which remains compatible with a variety of speech/speaker modeling methods. Then, IRISA, FRANCE TELECOM and LORIA designed each a specific inter-speaker dissimilarity measure. The obtained lists of reference speakers were compared and jointly optimized. The corpus IDIOLOGOS, composed of 450 telephone speech sentences uttered by each of the 200 reference speakers, is now available at ELRA. The global methodology designed to extract the reference speakers is described in [14].

7.4. International Contracts

7.4.1. HIWIRE

The HIWIRE (Human Input That Works In Real Environments) Project is funded by the European Commission in the framework of the 6th PCRD. The HIWIRE project aims at making significant improvements to the robustness, naturalness, and flexibility of vocal interaction between humans and machines.

The overall objective of the HIWIRE project is to set the basis for much more dependable speech recognition in mobile, open and noisy environments, and needs technical breakthroughs. The achievements of the project will be validated through:

- Assessment of the potential of contribution of vocal interaction to safety and efficiency in future commercial cockpits.
- Usability evaluation of enhanced dialogue in an open environment on a mobile device.

This main objective at a strategic level is split into three working objectives:

1. To make significant improvements to the robustness of speech recognition in noisy environments.
2. To make significant improvements to the robustness of speech recognition to different user's voices and interaction abilities.
3. To evaluate the potential impact of more robust speech recognition in real-world applications.

Two kinds of activities are planed: long-term research and research for the fixed and mobile plateforms.

The partners are: Thales Avionics (F), Thales Research (F), Loquendo (I), Technical University of Crete TSI-TUC (G), University of Granada GSTC-UGR (SP), National Technical University of Athens ICSS-NTUA (G), Center for Scientific and Technological Research ITC-IRST (I) and LORIA (F).

During this year we worked on the following subjects:

- Missing data: we propose original approaches to deal with non-stationary noise (see section 6.2.1.1).
- Non-native speech recognition: we modify lexicon to take into account pronounciation variation due to non-native speakers (see section 6.2.1.2).

The project has organized a special session during INTERSPEECH 2007 in Anvers.

For Loria, 4 international publications [25], [27], [36], [33] have been published during year 2007.

In order to develop and test new approaches for non-native speakers, we have recorded 31 French speakers. Each speaker uttered 100 sentences corresponding to command language for aircraft pilots. The recording software has been developed by LORIA: it allows recording and listening lists of sentences.

During the final review (july 2007), reviewers judge that the project was "good to excellent".

7.4.2. Amigo

Amigo is an Integrated Project funded by the European Commission, whose main topic is "Ambient intelligence for the networked home environment". Its reference number is IST 004182; it is leaded by Philips Research Eindhoven and includes Philips Design - Philips Consumer Electronics (the Netherlands), Fagor (Spain), France Telecom (France), Fraunhofer IMS (Germany), Fraunhofer IPSI (Germany), Ikerlan (Spain), INRIA (France), Italdesign Giugiaro (Italy), Knowledge (Greece), Microsoft (Germany), Telin (the Netherlands), ICCS (Greece), Telefónica I+D (Spain), University of Paderborn (Germany) and VTT (Finland).

In this project, we are collaborating with the Langue & Dialogue team in Nancy to continue 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.

During the reporting period, we mainly addressed two challenges: the first one deals with integrating our contribution within the platform developed by the other partners, and the second one is related to the handling of implicit speech interactions in the Amigo ambient intelligent framework. To address the latter issue, we first proposed a generic architecture for implicit interactions that facilitates the work of application developpers who need to handle such kinds of interactions. Concretely, this architecture makes the link between the context management service, the multimodal fusion module and the calling application, and presents a simplified interface to the application developer. This architecture is described in the project deliverables. More recently, as a consequence to the evolution of the other Amigo services, we deeply modified this architecture in order to integrate both implicit and explicit speech interactions within the same dialog manager, and to factorize the subscription mechanism within the context management service. We further developed such an implicit speech interaction service that is based on a real-time keyword spotting Amigo web service. We plan to use this service to infer the user activity within a smart agenda application, in collaboration with Fraunhofer IPSI.

7.4.3. 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 worked about 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.

7.4.4. 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. Being able to recover this information automatically would be a major break-through in speech research and technology, as a vocal-tract representation of a speech signal would be both beneficial from a theoretical point of view and practically useful in many speech processing applications (language learning, automatic speech processing, speech coding, speech therapy, film industry...). The design of audiovisual-to-articulatory inversion involves two kinds of interdependent tasks. The first is the development of inversion methods that successfully answer the main acknowledged difficulties (non-unicity of inverse solution, lack of phonetic relevancy of inverse solutions, impossibility of using standard spectral data), and the second is the construction of an articulatory database that comprises dynamic images of the vocal tract together with the speech signal uttered, and that for several male and female speakers. 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 synchronization of the different imaging modalities [23], [22].The present 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 are starting the evaluation of inversion by using these articulatory data and other data acquired through an articulograph. In addition we are working about the inversion of fricative sounds and we also investigated the automatic processing of ultrasound and X-ray data with a view to track articulators.

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 and the incorporation of constraints.

7.4.5. CMCU - Tunis University

This cooperation involves the LSTS (Laboratoire des systèmes et Traitement du Signal) of Tunis University headed by Prof. Noureddine Ellouze and Kaïs 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.

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.

- Member of editorial boards :
  - Speech Communication (J.P. Haton)
  - Computer Speech and Language (J.P. Haton)
  - EURASIP Journal on audio, Speech, and Music Processing
- Member of scientific commitee of conference :
  - ICSLP (J.P. Haton)
  - LREC (Y. Laprie)
  - JEP (Y.Laprie)
  - MCCSIS'07 (S.Ouni, C. Cerisara)
- Member of evaluation commission of INRIA (I.Illina)
- 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)
- In charge of the "Assistant intelligent" project of the PRST "Intelligence Logicielle" (Y. Laprie)
- Member of the lorrain network on specific language and Learning disabilities and in charge of the speech and language therapy expertise in the Meurthe-et-Moselle House of Handicap (MDPH) (A. Kipffer-Piquard)
- The members of the team have been invited as lecturer :
  - at TAIMA (Traitement et Analyse de l'Information : Méthodes et Applications[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)
- Demonstration :
  - "journées de la science" : presentation of the ANTS system (see section 5.1.9)
  - 40th anniversary of INRIA (December 10 and 11, Lille) : presentation of a prototype of the talking head (see 6.1.2)

8.2. Distinctions

- Jean-Paul Haton is Professor at IUF (Institut Universitaire de France).

8.3. Invited lectures

- Frederic Apoux, LPP, Université Paris 5,
- Paavo H.T. Leppänen, Finnish Center of Excellence for Learning and Motivation, Psych. Dept., University of Jyväskylä, Finland,
- Matthieu Chabanas, ICP Grenoble, Université Paris 5,
- Liliane Sprenger-Charolles and Willy Serniclaes, LPP CNRS,
- Korin Richmond, CSTR, Edinburgh University, Scotland.

8.4. Higher education

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

- 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).

8.5. Participation to workshops and PhD thesis committees:

- Members of Phd thesis committees C. Cerisara, I. Illina, D. Fohr, J.-P. Haton, M.-C. Haton, Y. Laprie, K. Smaïli;
- 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] 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/.

[9] 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

Doctoral dissertations and Habilitation theses

[10] S. DEMANGE. *Contributions à la reconnaissance automatique de la parole avec données manquantes*, Ph. D. Thesis, Université Henri Poincaré - Nancy I, 11 2007, http://tel.archives-ouvertes.fr/tel-00187953/en/.

[11] E. DIDIOT. *Segmentation parole/musique pour la transcription automatique de parole continue*, Ph. D. Thesis, Université Henri Poincaré - Nancy I, 11 2007, http://tel.archives-ouvertes.fr/tel-00187941/en/.

[12] P. KRAL. *Reconnaissance automatique des actes de dialogue*, Ph. D. Thesis, Université Henri Poincaré - Nancy I, 11 2007, http://tel.archives-ouvertes.fr/tel-00188197/en/.

[13] J. RAZIK. *Mesure de confiance trame-synchrones et locales en reconnaissance automatique de la parole*, Ph. D. Thesis, Université Henri Poincaré - Nancy I, 10 2007, http://tel.archives-ouvertes.fr/tel-00185747/en/.

Articles in refereed journals and book chapters

[14] F. BIMBOT, O. BOËFFARD, D. CHARLET, D. FOHR, S. KRSTULOVIC, O. MELLA. *Selecting Representative Speakers for a Speech Database on the Basis of Heterogeneous Similarity Criteria*, in "Speaker Classification II Lecture Notes in Computer Science", C. MÜLLER (editor), Lecture Notes in Computer Science, vol. 4441, Springer Berlin / Heidelberg, 2007, p. 276-292, http://hal.inria.fr/inria-00187732/en/.

[15] 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, 2007, p. 443-457, http://hal.inria.fr/inria-00160554/en/.

[16] C. CERISARA, Y. HARADJI. *Nouvelles formes d'interaction homme-machine pour l'informatique diffuse*, in "Informatique diffuse ARAGO", M. DUPUIS (editor), ARAGO, vol. 31, OFTA, 2007, http://hal.inria.fr/inria-00160547/en/.

[17] Y. CLÉMENT, C. JOUBERT, C. KOPP, E. LEPICARD, P. VENAULT, R. MISSLIN, M. CADOT, G. CHAPOUTHIER. *Anxiety in Mice: A Principal Component Analysis Study*, in "Neural plasticity", 2007, http://hal.inria.fr/inria-00186107/en/.

[18] P. KRAL, C. CERISARA, J. KLECKOVA. *Lexical Structure for Dialogue Act Recognition*, in "Journal of Multimedia", vol. 2, 2007, p. 1-8, http://hal.inria.fr/inria-00184475/en/.

[19] S. OUNI, M. COHEN, H. ISHAK, D. MASSARO. *Visual Contribution to Speech Perception: Measuring the Intelligibility of Animated Talking Heads*, in "EURASIP Journal on Audio, Speech, and Music Processing", vol. 2007, 2007, ID 47891, http://hal.archives-ouvertes.fr/hal-00184425/en/.

[20] B. POTARD, Y. LAPRIE. *Inversion acoustique-articulatoire en utilisant des contraintes phonétiques*, in "Perturbations et réajustements : langue et langage", B. VAXELAIRE, R. SOCK, G. KLEIBER, F. MARSAC (editors), Textes issus des Journées fédératrices "Perturbations et réajustements, langue et langage" organisées à Haguenau en décembre 2004 par le Réseau des sciences cognitives du Grand Est, Cogniest, l'E.A. 1399 Linguistique, langues et paroles -LilPa..., Publications de l'Université Marc Bloch (Strasbourg), 2007, http://hal.inria.fr/inria-00180716/en/.

[21] V. ROBERT, A. BONNEAU, B. WROBEL-DAUTCOURT, Y. LAPRIE. *Prédiction phonétique de la coarticulation labiale*, in "Perturbations et réajustements : langue et langage", B. VAXELAIRE, R. SOCK, G. KLEIBER, F. MARSAC (editors), Textes issus des Journées fédératrices "Perturbations et réajustements, langue et langage" organisées à Haguenau en décembre 2004 par le Réseau des sciences cognitives du Grand Est, Cogniest, Publications de l'Université Marc Bloch (Strasbourg), 2007, p. 155-167, http://hal.inria.fr/inria-00180714/en/.

Publications in Conferences and Workshops

[22] M. ARON, N. FERVEUR, E. KERRIEN, M.-O. BERGER, Y. LAPRIE. *Acquisition and synchronization of multimodal articulatory data*, in "8th Annual Conference of the International Speech Communication Association - Interspeech'07, Antwerpen Belgique", 2007, p. 1398-1401, http://hal.inria.fr/inria-00165869/en/.

[23] M. ARON, E. KERRIEN, M.-O. BERGER, Y. LAPRIE. *Fusion de capteurs électromagnétiques et d'échographies pour le suivi de la langue*, in "Onzième congrès francophone des jeunes chercheurs en vision par ordinateur - ORASIS'07, Obernai France", 06 2007, http://hal.inria.fr/inria-00142462/en/.

[24] L. BOUGRAIN, J. FELDMAR, E. SIDHOUM. *Modélisation connexionniste du traitement de l'accès lexical et mise en relation avec des données électro-encéphalographiques*, in "ARCo07, Nancy France", 2007, http://hal.inria.fr/inria-00179885/en/.

[25] G. BOUSELMI, D. FOHR, I. ILLINA. *Combined Acoustic and Pronunciation Modelling for Non-Native Speech Recognition*, in "InterSpeech 2007, Antwerp Belgique", ISCA, Universiteit antwerpen, Radboud University Nijmegen and Katholieke Universiteit Leuven, 2007, http://hal.inria.fr/inria-00184560/en/.

[26] G. BOUSELMI, D. FOHR, I. ILLINA, J.-P. HATON. *Amélioration des Performances des Systèmes Automatiques de Reconnaissance de la Parole pour la Parole Non Native*, in "Traitement et Analyse de l'Information : Méthodes et Applications - TAIMA'07, Hammamet Tunisie", Jean-Paul Haton and Faouzi Ghorbel, 2007, http://hal.inria.fr/inria-00184565/en/.

[27] G. BOUSELMI, D. FOHR, I. ILLINA, J.-P. HATON. *Discriminative Phoneme Sequences Extraction for Non-Native Speaker's Origin Classification*, in "International Symposium on Signal Processing and its Applications - ISSPA 2007, Sharjah Arabie saoudite", IEEE, The American University, The University of Sharjah, Etisalat University College and Eurasip, 2007, http://hal.inria.fr/inria-00184553/en/.

[28] A. BOYER, A. BRUN. *Natural language processing for usage based indexing of web resources*, in "29th European Conference on Information Retrieval - ECIR'07 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007, Proceedings Lecture Notes in Computer Science, Rome Italie", Lecture Notes in Computer Science, The original publication is available at www.springerlink.com ISBN: 978-3-540-71494-1; ISSN 0302-9743 (Print) 1611-3349 (Online), vol. 4425, Springer Berlin / Heidelberg, Fondazione Ugo Bordoni; BCS-IRSG; ACM SIGIR, 2007, p. 517-524, http://hal.inria.fr/inria-00172231/en/.

[29] A. BRUN, A. BOYER. *Usage based indexing of web resources with natural language processing*, in "3rd International Conference on Web Information Systems and Technologies - Webist 07, Barcelone Espagne", INSTICC - Institute for Systems and Technologies of Information, Control and Communication ; Open University of Catalonia, 2007, http://hal.inria.fr/inria-00172234/en/.

[30] A. BRUN, D. LANGLOIS, K. SMAÏLI. *Improving language models by using distant information*, in "International Symposium on Signal Processing and its Applications - ISSPA 2007, Sharjah Émirats arabes unis", 2007, http://hal.inria.fr/inria-00187084/en/.

[31] M. CADOT, P. CUXAC, A. LELU. *Random simulations of a datatable for efficiently mining reliable and non-redundant itemsets*, in "12th International Conference on Applied Stochastic Models and Data Analysis - ASMDA 2007, Chania, Crête Grèce", 2007, http://hal.inria.fr/inria-00186100/en/.

[32] M. CADOT, A. LELU. *Simuler et épurer pour extraire les motifs sûrs et non redondants*, in "7èmes Journées Francophones "Extraction et Gestion des Connaissance" - EGC 2007 - Troisième Atelier Qualité des Données et des Connaissances - QDC, Namur Belgique", Stéphane Lallich, Philippe Lenca et Fabrice Guillet, 2007, p. 15-24, http://hal.inria.fr/inria-00186096/en/.

[33] S. DEMANGE, C. CERISARA, J.-P. HATON. *Accurate marginalization range for missing data recognition*, in "INTERSPEECH 2007, Antwerp Belgique", 2007, http://hal.inria.fr/inria-00184423/en/.

[34] D. DURED, S. FRANC, J. FELDMAR. *L'adaptation des doses d'insuline chez les sujets diabétiques de type 2.*, in "Diabétologie et Endocrinologie, Marseille France", 2007, http://hal.inria.fr/inria-00186669/en/.

[35] J. FELDMAR, J. SAGOT, P. SUIGNARD. *L'étude Labiao, une aide à l'intégration des étudiants sourds, un partenariat scientifique et pédagogique*, in "Colloque inaugural de l'INS HEA, Surresnes France", 2007, http://hal.inria.fr/inria-00186676/en/.

[36] D. FOHR, I. ILLINA. *Text-Independent Foreign Accent Classification Using Statistical Methods*, in "International Conference on Signal Processing and Communications, Dubai Émirats arabes unis", IEEE, 11 2007, 4, http://hal.archives-ouvertes.fr/hal-00163745/en/.

[37] G. HENRY, A. BONNEAU, V. COLOTTE. *Tools devoted to the acquisition of the prosody of a foreign language*, in "International Congress of Phonetic Sciences - ICPhS 2007, Saarbrücken Allemagne", 2007, p. 1593-1596, http://hal.inria.fr/inria-00184530/en/.

[38] P. KRAL, C. CERISARA, J. KLECKOVA. *Confidence measures for semi-automatic labeling of dialog acts*, in "IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP 2007, Honolulu États-Unis d'Amérique", pp. IV-153 - IV-156, vol. 4, 2007, http://hal.inria.fr/inria-00184469/en/.

[39] P. KRAL, C. CERISARA, J. KLECKOVA. *Importance of Prosody for Dialogue Acts Recognition*, in "XIIth International Conference "Speech and Computer" - SPECOM'07, Moscou Russie", 2007, http://hal.inria.fr/inria-00184473/en/.

[40] Y. LAPRIE, A. BONNEAU. *Construction of perception stimuli with copy synthesis*, in "16th International Congress of Phonetic Sciences - ICPhS 2007, Saarbrücken Allemagne", Universität des Saarlandes, 2007, http://hal.inria.fr/inria-00180226/en/.

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

[42] 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/.

[43] 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/.

[44] V.-B. LE, O. MELLA, D. FOHR. *Speaker Diarization using Normalized Cross Likelihood Ratio*, in "INTERSPEECH 2007, Antwerp Belgique", ISCA, 08 2007, 4, http://hal.archives-ouvertes.fr/hal-00163855/en/.

[45] S. OUNI, K. OUNI. *Arabic Pharyngeals in Visual Speech*, in "International Conference on Auditory-Visual Speech Processing 2007 (AVSP), Hilvarenbeek Pays-Bas", vol. 1, University of Van Tilburg, 08 2007, p. 212-215, http://hal.archives-ouvertes.fr/hal-00174344/en/.

[46] S. OUNI, K. OUNI. *Aspects of Visual Speech in Arabic*, in "Interspeech 2007, Antwerp Belgique", ISCA, 08 2007, p. WeB.P1a-7, http://hal.archives-ouvertes.fr/hal-00174350/en/.

[47] B. POTARD, Y. LAPRIE. *Compact representations of the articulatory-to-acoustic mapping*, in "INTERSPEECH 2007, Antwerp Belgique", 2007, p. 2481-2483, http://hal.inria.fr/inria-00180230/en/.

[48] J. RAZIK, O. MELLA, D. FOHR, J.-P. HATON. *Frame-Synchronous And Local Confidence Measures For On-The-Fly Keyword Spotting*, in "International Symposium on Signal Processing and its Applications - ISSPA 2007, Sharjah/Emirats Arabes Unis", 02 2007, http://hal.inria.fr/inria-00134135/en/.

[49] V. ROBERT, Y. LAPRIE, A. BONNEAU. *A phonetic concatenative approach of labial coarticulation*, in "INTERSPEECH 2007, Antwerp Belgique", ISCA, 2007, p. 1402-1405, http://hal.inria.fr/inria-00184252/en/.

Internal Reports

[50] L. MOUROT, M. ROVEL, J. FELDMAR. *Evaluation of a talking head for helping HOH people in the classroom*, Research Report, 2007, http://hal.inria.fr/inria-00185030/en/.

References in notes

[51] M. ABBAS, K. SMAÏLI. *Comparison of Topic Identification methods for Arabic Language*, in "International Conference on Recent Advances in Natural Language Processing - RANLP 2005, Borovets, Bulgaria", 2005, p. 14-17.

[52] 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.

[53] 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.

[54] A. BOYER, A. BRUN. *Towards a statistical grammar of usage for document retrieval in digital libraries*, in "International Symposium on Signal Processing and its Applications (ISSPA'07), Shirjah, Associates Emirates", 02 2007, http://hal.inria.fr/inria-00119476/en/.

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

[56] A. BRUN, D. LANGLOIS, K. SMAÏLI. *Exploration et utilisation d'informations distantes dans les modèles de langage statistiques*, in "13ème Conférence sur le Traitement Automatique des Langues Naturelles - TALN'2006, Leuven/Belgique", 04 2006, p. 425-434.

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

[58] 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/.

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

[60] M. DEVIREN, K. DAOUDI, K. SMAÏLI. *Rethinking Language Models within the Framework of Dynamic Bayesian Networks*, in "18th Conference of the Canadian Society for Computational Studies of Intelligence - Canadian AI 2005, Victoria, Canada", B. KÉGL, G. LAPALME (editors), Lecture Notes in Computer Science, vol. 3501, Springer, 2005, p. 432-437.

[61] ETSI ES 202 050 V1.1.1. *Distributed speech recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms*, 2002.

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

[63] H. FRANÇOIS, O. BOËFFARD. *The Greedy Algorithm and its application to the Construction of a Continuous Speech Database*, in "proceedings of Language Resources And Evaluation Conference LREC 2002", 2002, p. 1420-1426.

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

[65] J.-P. HATON, C. CERISARA, D. FOHR, Y. LAPRIE, K. SMAÏLI. *Reconnaissance Automatique de la Parole Du signal à son interprétation*, UniverSciences (Paris) - ISSN 1635-625X, I.: Computing Methodologies/I.2: ARTIFICIAL INTELLIGENCE, I.: Computing Methodologies/I.5: PATTERN RECOGNITION, DUNOD, 2006, http://hal.inria.fr/inria-00105908/en/.

[66] 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/.

[67] 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/.

[68] 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.

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

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

[71] D. LANGLOIS, K. SMAÏLI, J.-P. HATON. *Efficient linear combination for distant n-gram models*, in "8th European Conference on Speech Communication and Technology - Eurospeech'03, Genève, Suisse", vol. 1, Sep 2003, p. 409-412.

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

[73] 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.

[74] 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.

[75] 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/.

[76] K. PAPINENI, S. ROUKOS, T. WARD, W.-J. ZHU. *Bleu: a Method for Automatic Evaluation of Machine Translation*, in "Proceedings of the 40th Annual of the Association for Computational linguistics, Philadelphia, USA", 2001, p. 311-318.

[77] K. SMAÏLI, C. LAVECCHIA, J.-P. HATON. *Linguistic features modeling based on Partial New Cache*, in "International Conference on Language Resources and Evaluation - LREC 2006", 05 2006.

[78] 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/.

[79] 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".

[80] 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.