INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team Parole

Analysis, Perception and Speech Recognition

Lorraine

THEME COG

Table of contents

1. Team

PAROLE is common 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]

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 since 2005, July 7th]
Vincent Colotte [Assistant Professor, Henri Poincaré University]
Joseph di Martino [Assistant Professor, Henri Poincaré University]
Jean-Paul Haton [Professor, Henri Poincaré University, Institut Universitaire de France]
Marie-Christine Haton [Professor, Henri Poincaré University]
Irina Illina [Assistant Professor, I.U.T Charlemagne, Nancy 2 University working for INRIA since 2004, September]
David Langlois [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 since 2004, September 1st]
Kamel Smaïli [Professor, Nancy 2 University]

Phd Students

Ghazi Bouselmi [TA, thesis to be defended in 2007]
Matthieu Camus [INRIA Grant, thesis to be defended in 2008]
Sébastien Demange [TA, thesis to be defended in 2007]
Emmanuel Didiot [CIFRE grant, thesis to be defended in 2008]
Pavel Král [Czech coPhD, thesis to be defended in 2006]
Blaise Potard [MENRT grant, thesis to be defended in 2006]
Joseph Razik [INRIA grant, thesis to be defended in 2006]
Vincent Robert [High school teacher, thesis to be defended in 2007]
Caroline Lavecchia [EADS foundation grant, thesis to be defended in 2008]

Project technical staff

Christophe Antoine [INRIA, since 2005, January 15th]
Alexandre Lafosse [CNRS, since 2005, October 1st]
Julien Maire [CNRS, since 2005, September 1st]

Post Doctoral fellow

Vincent Barreaud [ATER, ESIAL, since 2005, September]
Salma Jamoussi [ATER, Nancy 2 University, until 2005, August]
Jean-Baptiste Maj [ATER, I.U.T. St Dié, Henri Poincaré University, since 2005, September]

2. Overall Objectives

2.1. Overall Objectives

PAROLE is a common 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, or 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.

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 or future applications: vocal rehabilitation, improvement of hearing aids, 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 or future applications: automatic speech recognition, automatic translation, text-to-speech alignment, audio indexing.

Our pluridisciplinary scientific culture combines works in phonetics and in pattern recognition as well. This pluridisciplinarity turns out to be a decisive asset to address new research topics, particularly language learning or 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 that 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 Syncmagic and Thales Aviation. The latter gave rise to a european project in the field of non-native speech recognition in a noisy environment. Moreover, we are involved in the 6th PCRD projects MUSCLE, AMIGO, HIWIRE and more recently ASPI as the coordinator team, and in a regional project with teachers in 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*, *stochastic models*, *telecommunications*.

Taken as a whole research in speech 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 good quality artificial speech signals, phoneticians 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 influenced between 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 aroused a second approach that consists in model observations of speech production and perception. Efforts focused onto the design of numerical models (first simple vectors of spectral shapes and now stochastic or neural models) of word or phoneme acoustical realizations, and onto the development of statistical language models.

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

3.2. Speech Analysis

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

3.2.1. Acoustic cues

The notion of strong and weak cues has been introduced to palliate a weakness of ASR (automatic speech recognition) systems: the lack of certitude. Indeed, due to the variability of speech signals, acoustical regions representing different sounds overlap one with another. Nevertheless, we know, from previous perceptual experiments [36], 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 strong cues, devoted to the reliable recognition of stop place of articulation. Strong cues, as we call them, identify or eliminate a feature of a given sound with certainty (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 strong 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 strong cue must not be merged into that 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 one case out of four.

Strong 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.

3.2.1.1. Automatic detection of “well realized” sounds

The detection of strong 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 certitude are probably well realized and well pronounced sounds. 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. On one hand, it is possible to find well realized features with a speech recognition system based upon phonetic knowledge, through the use of "strong cues". But this method cannot be entirely automatic, especially because of segmentation problems. Stochastic methods, such as the 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 HMM to modelize those well identified sounds in the following way. First, on a training corpus, the system modelizes 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 of this same strategy, 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 very few 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. Our project concerning the design and development of computer-assisted learning of prosody is presented in section 8.1.1 (national projects).

3.2.2.1. Speech signal transformation

In order to improve oral comprehension, we use a speech signal transformation method called PSOLA (Pitch Synchronous Overlap and Add). This method 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. This method is well known for its easy implementation and the quality of the slowed down signals. However, temporal discrepancies can appear in the region of the synthesis marks and noise can be generated between harmonics. In order to reduce the loss of quality, the method was improved in the two following ways. First, a pruning algorithm has been introduced to seek analysis marks (for pitch synchronization). It increases 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 high quality speech signal [40].

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 (see section 8.1.1), we start 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, via the 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 instigated 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.

3.2.3. Acoustic-to-articulatory inversion

Acoustic-to-articulatory inversion aims at recovering the articulatory dynamics from the speech signal that may be supplemented by images of the speaker face. Potential applications concern low bit rate speech coding,

automatic speech recognition, assessing speech production disorders, articulatory investigations of phonetics, talking heads and articulatory feedback for language acquisition or learning.

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

Solving acoustic equations. In order to solve 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 ones through the acoustical-electrical analogy, **(ii)** spatio-temporal, 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 [48].

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

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

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

3.2.4. Strategies of labial coarticulation

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

In the long term, we want 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 [43] 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 performances 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 were developed, that of Abry and Lallouache [35] advocates for the theory of expansion movements: the movement tends to be anticipated when no phonological constraint is imposed on labiality. Cohen and Massaro [39] proposed dominance functions that require a substantial numerical training.

It should be noted that most of these models derive from the observations of a few number of speakers only. We want to develop 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 can not 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 the speech signal. The second one is called “language models”, 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, missing data recognition requires 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.

Finally, we have developed a new robust front-end, which is based on wavelet-decomposition of the speech signal. This front-end generalizes the Frequency filtered coefficients.

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, they 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 make research on BNs, but we rather exploit them to reason about the important statistical properties of robust speech recognition.

3.3.1.3. Robustness and invariance

The core of our research activities about speech recognition aims at improving the robustness of automatic speech 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 automatic speech recognition systems is certainly the most problematic one that still prevents the wide deployment of speech recognizers nowadays. In the past, we had developed a large range of techniques to address this difficult topic, including robust acoustic models (such as stochastic trajectory or 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 the channel mismatches (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 acts 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 good classifier. Various classifiers are commonly used: k-Nearest-Neighbors, Hidden Markov Models, Gaussian Mixture Models, Artificial Neural Networks.

As to detect speaker turns, the main approach consists of splitting the audio signal into 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

From an acoustical point of view, we can consider today that several systems achieve good performances. Nevertheless, some problems due to the complexity of natural language remain without any satisfactory solution. Our group, as others through the world, make more and more efforts in order to design efficient language models. All the language models we propose are based on information theory and statistics. Moreover, some of them also use linguistic knowledge to guide the statistical one. The combination of different language models allows better performances to be obtained but we pursue the development of new methods in order to make models more efficient. To do so, we work through several directions:

- Language model adaptation using topic identification. The objective of this research area is first to find out the topic of the uttered sentences, second, to adapt the baseline language model using the

one which corresponds to the retrieved topic. Research is in both identification and adaptation [38], [14].

- Selecting the best language model in accordance with 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 or translation system is to select dynamically the best language model depending on a history as in the SHP principle [47]. We pursue in this research direction in order to obtain an efficient selection of language models.
- One direction of research is to consider a word not only as an orthographic form but as a complex unit which contains a class tag, a gender and number features, semantic tag... This makes a statistical language model more realistic and more in agreement with linguistic theory. Some tracks were explored and confirmed the feasibility of this approach [51], [22].
- One of the most promising research directions for the next ten years in our group is speech to speech translation. In fact, this activity concerns not only speech recognition problems, but also machine translation, language model adaptation, speech understanding and decoding problems.

4. Application Domains

4.1. Application Domains

Our research works are applied in a variety of fields from automatic speech recognition to paramedical domains. Speech analysis methods will contribute to the development of new technologies for language learning (for hearing-impaired persons or for the teaching of foreign languages) as well as for hearing aids. We developed in the past a set of teaching tools based on the speech analysis and recognition algorithms of the group (cf. the ISAEUS [44] 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 IVOMOB project of RNRT for the use of speech recognition in a car), interactive vocal servers, telephone and domestic applications, etc. Most of these applications will necessitate to integrate the type of speech understanding process our group is presently studying. 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, or key word spotting.

5. Software

5.1. Software

5.1.1. PhonoLor

PhonoLor is a phonetizer enabling word or sentence translations into a sequence of phonemes. This software exploits phonetization rules learnt from a corpus of examples.

5.1.2. Snorri and WinSnoori

Snorri is a speech analysis software we have been developing since 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) because 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 speech files phonetically or orthographically, to extract fundamental frequency, to pilot the Klatt synthesizer and to utilize PSOLA resynthesis.

Snorri was used as a software resource for several works in our team (formant tracking, stop identification, perceptive studies, ...). Given the interest it represents for speech analysis we distributed it to about fifteen French-speaking teams. Initially developed under Unix and Motif it was ported under Windows and we sell it under the name WinSnoori through Babel Technologies (startup located in Mons in Belgium and distributing text-to-speech and automatic speech recognition software).

This year we added one module to compute Mel cepstrally smoothed spectra. Mel cepstra are widely used in speech recognition but there is very little knowledge about the exact spectral effect they produce, and consequently the nature of the spectral information stored into acoustic HMMs. This module enables the visualization of spectra obtained with Mel cepstra. For that purpose we applied an inverse DCT (DCT-III since the discrete cosine transform used to compute Mel cepstral coefficient is the DCT-II) after the calculation of Mel cepstral coefficients to produce a smoothed spectrum. These spectra are displayed in the linear frequency scale (Hertz) to enable the comparison with other spectral techniques. Mel-scale filters or Bark critical band filters can be used to compute the energy data before the application of the DCT used to compute Mel cepstral coefficients. In addition to the computation of running spectral slices, it is possible to display spectrograms obtained after the application of the filter banks with Bark critical bands or Mel filters.

We also worked about copy synthesis for the Klatt synthesizer, i.e. the possibility of generating parameters for the formant synthesizer that allow the original speech signal to be copied faithfully. Copy synthesis is an invaluable tool to generate speech stimuli usable in speech perception studies, which sound like natural speech. One of the most difficult problems is to adjust the amplitudes of formants. This often requires the detection of the closed part of each pitch period which is a very difficult problem. We thus proposed to evaluate the amplitude of formants by localizing the most intense part in each pitch period because it is likely the most informative in terms of formant amplitudes. Pitch periods are detected through our pitch marking algorithm we previously developed. The amplitudes are evaluated by means of a short term Fourier transform applied in each pitch period where the energy of the signal is maximal over a small temporal window. Results show that this new strategy to compute formant amplitudes enable sharp energy transitions of stop consonants to be captured efficiently. This new facility implemented in WinSnoori supplements tools already available to edit parameters of the Klatt synthesizer.

5.1.3. Labelling corpora

We developed a labelling tool which allows corpus syntactic ambiguities to be solved. To each word, its syntactic class 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 an error labelling of about 1%.

5.1.4. Automatic lexical clustering

In order to adapt language models in speech recognition applications, a new toolkit has been developed 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 is the one minimizing the perplexity of the corresponding language model. Several options are available: the user can, for example, 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 the 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 obtaining 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 a 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 a PC-Linux or PC-Windows.

5.1.8. 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 is performed 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-trellis. In the second pass a stack decoding algorithm using a trigram language model gives the N-best recognition sentences.

5.1.9. 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 are 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. For example, this toolkit shall be used to achieve Bayesian denoising of speech signals (like in the SPLICE algorithm), or uncertainty decoding. We will not support dynamic Bayesian networks, but we rather plan to interface BNTK with the standard HMM toolkit (HTK). Hence, BNTK will be used to model or denoise every speech frame, while HTK might handle the dynamic (stochastic) properties of the speech signal.

The main features that distinguish this toolkit from the other Graphical Model toolkits available are:

- BNTK is open-source, at the contrary of several other toolkits (e.g. Murphy's toolkit requires Matlab);
- BNTK supports both continuous (linear Gaussian) and discrete variables;
- BNTK realizes exact inference (strong junction tree) with most graphical networks;
- BNTK assumes few constraints on the network topology (e.g. RISO supports only polytrees);
- BNTK supports multidimensional variables, and can read/write observations through dedicated C++ classes. It can thus be used with real databases, as the observations do not need to be typed by hand;
- BNTK does not have a GUI, but the inference and training engines can be accessed via C++ interfaces, making it easy to interface the toolkit with another program.

This toolkit is currently under development, and a first version, under the LGPL license, should be distributed very soon (at the very beginning of 2006) on the GForge INRIA Web Site.

5.1.10. 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 and OZONE platforms, exploits HTK acoustic models. Hence, we have developed our own set of additional recognition tools that support the HTK format, and that can interface with HTK. These tools are described next:

- The HMM parallel training tool distributes the training process over the 30 computers of the PAROLE PC cluster (see section 8.2.1). It has a graphical user interface in JAVA.
- The missing data library is composed of several utilities to experiment missing data recognition algorithms. The first program is the HTK missing data patch, which alters the way HTK computes the observation likelihood, by calling an external library that marginalizes out the masked coefficients. Three types of marginalization have been implemented: full, bounded and SNR-based. Another program computes oracle masks, using two different approaches. A third utility estimates the masks from mask models trained in different noisy conditions (see section 6.2.1).
- The HMMModelConv toolkit is an extensible software that converts acoustic models between different formats: HTK, ESPERE and Sphinx3.
- GMMlib is a JAVA library that manipulates Gaussian Mixture Models (GMM), together with the most important data formats used by HTK (MFCC files, model files, label files). It is highly modular, and is validated by JUnit regression test suites. Its high flexibility makes it a valuable tool to experiment in different research areas: for example, we are extensively using it in missing data recognition (it supports marginalization and it can manipulate and display masks) and in Bayesian denoising (to train the joint clean and noisy speech GMMs, to marginalize them, map one of them onto the other, and denoise speech). We plan to shortly distribute it under the LGPL license on the INRIA GForge web site.

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, Jean-Baptiste Maj, Joseph di Martino, Slim Ouni, Blaise Potard, Matthieu Camus.

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 [13].

This year we continued the work about the incorporation of phonetic constraints in two directions. The first is the comparison of inverse solutions with and without constraints on a human speaker. As there are very few real data we used the X-ray data used by Maeda to build his articulatory model. These data present another strong advantage since there should be no model mismatch between the analyzing model and the underlying human production model. It should be noted that despite this favorable situation the analyzing model is unable to produce exactly the same formants as those measured in the original vowels. Experiments show that the incorporation of constraints allows relevant articulatory trajectories to be recovered [30]. The second direction of research is the exploitation of 3D data of the speaker's face to reduce the under-determination of the inversion. Indeed, three out of the seven parameters of the articulatory model correspond to visible articulators (lower jaw and lips). On the other hand, there is strong evidence that human speakers/listeners exploit the multimodality of speech, and more particularly the articulatory cues: the view of visible articulators improves speech intelligibility. We thus carried out a pilot study to investigate how visual data could be used in the inversion process. Visual data were obtained by stereo-vision and enable the 3D recovery of jaw and lip movements. These data were processed to fit the nature of parameters of Maeda's articulatory model. Contrary to the inversion from formant data only, the mismatch between the analyzing and the analyzed model (the human subject) may strongly influence the inversion results. Indeed, the incorporation of the three visible parameters in addition to the frequencies of the first three formants may give rise to erroneous compensatory articulatory effects concentrated on the free parameters. In order to prevent the method from wrongly over-determine the inversion we used visible articulatory parameters to select hypercubes and not precise articulatory points within cubes. Combined with phonetic constraints these visual constraints enable the exploration space to be substantially reduced [34].

These 3D data were also used to show evidence of the existing of inter-speakers variability during speech production. This means that we cannot retrieve the vocal tract dynamics as produced by a given speaker using the inversion, if the speaker choices (context, personal preferences, habits or social conventions) are not taken into consideration [13].

6.1.2. Talking Head

As part of the France-Berkeley cooperation with PSL (Perceptual Science Laboratory - University of California, Santa Cruz), we were involved in designing the French articulations of a talking head. The talking head Baldi is a multilingual system. In addition to a text-to-speech for French, the articulatory definition of all the phonemes was required. During this year, we worked on defining each phoneme visually (definition of the position of the different articulators of the vocal tract, inner and outer parts). We used Several X-ray tracings

of French phonemes for this purpose, in addition to making sure that these definitions were consistent with phonetics. Several coarticulation aspects were incorporated based on phonetic knowledge.

Within the framework of a cooperation with the Magrite team (computer vision and augmented reality) we have been working on the design of labial coarticulation prediction algorithms. The contribution of Magrite team consists of the development of a 3D acquisition infrastructure. This enables the acquisition of a large amount of 3D data with a low-cost system more flexible than existing motion capture systems (using infrared cameras and glued markers). This system only uses two standard cameras, a PC and painted markers that do not change speech articulation and provides a sufficiently fast acquisition rate to enable an efficient temporal tracking of 3D points. Ten French native speakers (5 female and 5 male speakers) were recorded. Our corpus was made up of 4 isolated vowels (/i, y, a, o/), 6 consonants (/p, t, d, s, Elzesh, f/) followed by schwa, 8 CV, 20 VCV, 18 VCCV and 2 phonetically balanced sentences. In order to limit the preparation time only 15 markers were painted on the speaker's face. Unlike most of the previous studies we also include consonants with a primary or secondary labial articulation (/p, Elzesh/) because we are interested in the general process of labial coarticulation. We exploited this corpus to evaluate the interspeaker variability and determine the most invariant coarticulation strategies [32], [33].

A second corpus with a much higher number of markers (190) was recorded for a female subject. Here, the objective is to design a labial coarticulation prediction algorithm intended to pilot the artificial 3D face of this speaker. This corpus will be used within the framework of the LABIAO project.

6.1.3. Text-to-Speech synthesis

In the context of a Text-to-Speech synthesis system, a new synthesizer based on Non-Uniform Units (NUU) selection has been set up during the postdoctoral position of Vincent Colotte at the MULTITEL research center (in Belgium). It aims at compensating drawbacks of current NUU-based synthesis systems: the intrinsic weakness of their prosodic model, restricted to some acoustic and symbolic parameters, that does not allow sufficient and natural enough prosodic variations for synthesized sentences. The system is called LiONS, which stands for *LInguistically-Oriented Non-uniform units Selector*. It is original in two ways. First, it is freed from any prosodic model, whatever acoustic or symbolic: speech units are selected only using linguistic features, taken among the linguistic analysis of the text (given in input). Second, linguistic features used for selecting speech units are automatically weighted thanks to an original, entropy-related method. We made an evaluation of the system quality with the collaboration of MULTITEL [20].

The general evaluation is definitely positive. Concatenation, melody and listening comfort are felt as normal, while the intelligibility of the speech is very highlighted. Results also show the quality lack of the original female voice, always less appreciated than one could expect from a human voice. We may hope better results as soon as we will have a better database with respect to the speaker's voice quality.

This year, we also worked on the automatic weighting. In particular, the work concerned the acoustical clustering. Indeed, the more the clustering is reliable, the more the weighting should be relevant. The preliminary goal is to check the clustering to evaluate the relevance of the Kullback-Liebler distance for this task. We are using the Kohonen neural network to compare the results with the K-Means method which has been used in the system.

6.1.4. Automatic detection of well-realized speech sounds

This detection is based upon an iterative learning procedure intended to define acoustic models from systematically well detected realizations. An evaluation stage, using the HMM models created in the learning phase, has been performed to assess the rate of trigger action for the systematically well detected speech sounds as well as the rate of false alarms. We were mainly interested in the identification rate of the unvoiced speech sounds, in order to develop a speech enhancement technique. The global rate of correct detection of the elitist approach, using both models of systematically well-detected and other speech sounds, was 79.2% on average over the unvoiced speech sounds. The speech recognizer identifies 55% of the speech sounds as well-realized, with a very small rate of false alarms (0.82% on average) [27]. Hence, the elitist approach can extract automatically speech sounds, assumed to be well-realized, with high confidence. With the aim of developing

a speech enhancement technique, this is an important result. Next work will consist in verifying whether the sounds identified by the models for "well-detected" sounds have well marked acoustic cues.

For this purpose, we'll use data mining methods which are more suited to our problem than the statistical ones [11]. A robust data mining method, based upon randomization tests, is currently under development in our team [18].

6.2. Automatic Speech Recognition

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

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

The most important works on automatic speech recognition that have been recently achieved are presented in the following.

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. Bayesian denoising

We investigated an original speech denoising approach, in collaboration with K. Daoudi (IRIT-CNRS). We proposed an approach that maps noisy speech Gaussian Mixture Models (GMM) onto clean speech GMMs. This approach is a generalization of the SPLICE algorithm, recently proposed by Microsoft researchers. We further proposed a Maximum A Posteriori (MAP) adaptation procedure to adapt the method to unknown noisy test conditions. This first system has been advantageously compared to SPLICE in [21]. We next tested it on the standard Aurora2 database, and we proposed to jointly train the clean and noisy distributions to improve the correspondence between both GMMs. Then, we developed a dedicated adaptation procedure that guarantees this correspondence to be preserved even for unseen testing environments. This work has been submitted for publication at ICASSP'2006.

6.2.1.2. 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.

The progress we have achieved in this area during the past year can be summarized into the following points:

- We have realized an extensive literature review about the research areas that can bring relevant information to estimate the missing data masks, such as Computational Auditory Scene Analysis or Useful Speech. We have submitted this work for publication as a journal paper: it is currently under reviewing.
- We have built a baseline marginalization system that achieves comparable performances with state-of-the-art algorithms with oracle masks. To build this system, we have tested different alternatives, such as Frequency Filtered coefficients with full marginalization. The best results (in this oracle configuration) have been obtained with SNR-based marginalization and cubic-root compressed Mel frequency energies. SNR-based marginalization assumes that masks are computed from an estimate of the local SNR, and computes a marginalization range that is derived from this SNR information. The advantage of this approach is that the precision of marginalization directly reflects the accuracy and confidence of SNR estimation.

- We have proposed an original mask modeling technique that trains noise-dependent Bayesian mask models. For testing, the probability that the unknown environment is close to each of the training conditions is computed, and bayesian integration gives the most probable masks. The preliminary results obtained with this approach are encouraging, but a lot of work has still to be done, for example to handle large sets of known conditions.
- We have studied original mask definitions, which depend both on the noisy speech signal and on the speech decoder. These new masks actually approximate the final objective criterion, i.e. the recognition accuracy. Preliminary results are quite good, but the challenge is now to automatically infer such masks from the information available at test time.

6.2.1.3. Non native speakers

The performance of automatic speech recognition (ASR) systems drastically drops when they are confronted 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).

The main motivation of our work 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. On the non native database HIWIRE, this new method gave a significant improvement of recognition rates [16].

6.2.2. Core recognition platform

6.2.2.1. Broadcast News Transcription

In the framework of the *Technolangue* project ESTER, we developed a complete system, named ANTS, for French broadcast news transcription (see section 5.1.8).

In order to adapt acoustic models to the speaker, we added two new modules: one for speaker turn detection and speaker clustering and another one for MLLR-MAP adaptation [17]. The clustering process is based on the Bayesian Information Criterion (BIC).

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

For the real time system, we trained specific acoustic models with less free parameters. Moreover, the speaker clustering and the adaptation module were removed due to 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 continued to implement 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 focused on 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.

During broadcast programs, a lot of advertisement spots are emitted and could yield to false keywords detection. In order to remove such advertisements from the audio stream, we developed an approach based on finger print.

The objective of audio fingerprinting is an efficient mechanism to establish the perceptual equality of two audio files: not by comparing the (typically large) files themselves, but by comparing the associated fingerprints (small by design). The advantage of using fingerprints are: reduced memory/storage requirements as fingerprints are relatively small; efficient comparison as perceptual irrelevancies have already been removed from fingerprints; efficient searching as the dataset to be searched is smaller. The most important perceptual audio features live in the frequency domain. Therefore a spectral representation is computed by performing a Fourier transform on every frame. In order to extract a 32-bit sub-fingerprint value for every frame, 33 non-overlapping frequency bands are selected. These bands lie in the range from 300Hz to 2000Hz. This method gave very good accuracy for advertisement detection on French broadcast news, but short jingles segmentation needs more study to avoid false alarms.

6.2.2.3. Confidence measure

The engines used in large vocabulary speech recognition are mostly based on a probabilistic approach, and even with a huge dictionary (60000 words), the number of words known by the system is limited. Then, the results of the engines may bring out some errors due to false recognition and unknown words. That is why, having a criterion like a confidence measure can help the system to determine whether a word should be kept or not.

More and more applications need fast estimation of any measures in order to stay real-time. We propose some simple and fast measures, locally computed, that can be directly used within the first decoding recognition process.

We designed some measures based on a local view around the considered word. These measures are computed from acoustic likelihood of the words and bigram language models probabilities. They use the internal wordgraph built by the speech recognition engine within the first decoding pass and the n-best list provided by the full recognition process.

These new measures were evaluated on a 1-hour broadcast news corpus [31].

6.2.3. Ubiquitous speech recognition

We have just initiated in 2004 a new research area related to Ambient Intelligence: ubiquitous speech processing. 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 says can be used as an implicit speech interaction by the system. This can happen for example when the user is talking with someone else, in a meeting, in a classroom, or simply when he is listening to the radio. Note that in the PAROLE team, we are already working on these areas, via the ESTER and LABIAO projects for example. However, in Ambient Intelligence, implicit speech interactions are considered with the objective of assisting the user in his usual tasks.

We have mainly studied for now the challenges related to this application domain. We have thus identified the following issues and tasks:

- Processing speed: when the user is talking with someone else, he does not adapt his speaking style and rate to the system's limits, like in explicit speech interactions. Instead, speech is much likely to be very fast, unstructured and spontaneous.
- Speech semantic: the sense of sentences has to be interpreted to be helpful for the user. It is not possible nowadays to process and interpret efficiently such conversational speech in real-time. Hence, alternatives to the classical speech recognition-interpretation processing scheme have to be developed. They can be based on automatic topic recognition for example.
- Context: implicit speech inputs are only one aspect of the user context. The other aspects relate to the user profile, his environment, his activity, etc. Implicit speech interactions have to be integrated and exploit this information.

- Explicit vs. implicit detection: although implicit speech interactions are much more difficult to recognize and interpret than explicit speech interactions, the user's expectations are also much lower: even though the system misses a potential implicit speech interaction, it will not necessarily be even noticed by the user (Ambient Intelligence aims at assisting the user without disturbing him from his current task, i.e. in an unobtrusive way). This also implies that, when the user is at some time asking something to the system, this explicit request shall be recognized and treated at once: the role of the implicit/explicit detector is to detect such explicit interactions. It can be based on dialog act recognition for example.

Our main objective is to address parts of these challenges in the next few years, by first adapting the technologies that we have already developed in other projects, such as ESTER and LABIAO. The following two current researches are directly related to ubiquitous speech processing.

6.2.3.1. Dialog act automatic recognition

Dialog acts represent the role of successive sentences, or sequences of words, in the course of a dialog. We have focused our work on a very limited set of dialog acts: yes/no questions, open questions and declarative sentences. The objective of this work is to automatically identify the dialog acts. This information can be used for example to discriminate implicit from explicit speech interactions. The work realized during the past year concerns (i) the comparison of different combination schemes (naive Bayes, order statistics and non-linear combination) to merge both the prosodic and lexical information to recognize the dialog acts in Czech and in French, and (ii) the proposal of a new algorithm to further include the global sentence structure as an additional information for dialog act recognition. The proposed approach is based on a multiscale description of the words position in the sentence. Experimental results confirm that merging these three kinds of information (prosody, lexical and position) improve the classification accuracy.

6.2.3.2. OZONE platform

The OZONE real-time speech recognizer is based on the Julius decoder engine. It uses BNF grammars derived from the LTAG application grammars developed by the Langue&Dialogue team. It runs on a touch-screen laptop and is implemented as a WSAMI-compliant Web Service, which is the format developed at INRIA Rocquencourt ARLES. It interacts with the multimodal fusion module in the MMIL description language, and exploits the SOAP messaging protocol. It is our first prototype of a speech recognizer dedicated to Ambient Intelligent platforms. It also demonstrates the integration of our research activities within the dialogue system developed in the Langue&Dialogue team, and the web services software environment proposed by INRIA Rocquencourt.

6.2.4. Dynamic Bayesian networks (DBNs)

We propose novel modeling approaches for acoustic and linguistic modeling within the Bayesian networks formalism. Bayesian networks are a subset of probabilistic graphical models that include the most widely used probability models in speech recognition. These models are encoded with a graph structure that defines the probabilistic relations between its variables and a set of associated conditional probabilities. One of the main advantages of this representation is the graphical abstraction that provides a visual understanding of the modeled process. Moreover as a combination of probability theory and graph theory, this formalism covers several advantages from both domains. Therefore rethinking the modeling problems in this formalism provides new perspectives that were not considered previously.

6.2.4.1. Acoustic Modeling

State-of-the-art automatic speech recognition systems are based on probabilistic modeling of the speech signal using Hidden Markov Models (HMM). We reformulate the acoustic modeling problem in speech recognition within the probabilistic graphical models (PGM) formalism. *Dynamic Bayesian networks* (DBN) are a subset of PGM which include HMM as a special case. One of the principle weakness of HMMs is the independence assumptions on the observed and hidden processes of speech. We propose to use the DBN setting to extract the proper dependence structure for speech modeling rather than limiting ourselves with

HMMs. The proposed approach is based on structure learning paradigm in DBN framework. This approach has the advantage to guaranty that the resulting model represents speech with higher fidelity than HMM [42]. Recently, we proposed a new noise robust modeling technique in this framework that takes into account the variation of the acoustic environment.

6.3. Language Models

Language modeling is one of the important activities of our team. These last years, the PAROLE team worked on several projects: large vocabulary dictation machine, news transcription [45], automatic categorization of mails, dialog systems [50], vocal services [46], automatic speech-to-speech translation. All these applications include a linguistic module that deals with very different linguistic characteristics. In spite of all the improvements we obtained, the results are not entirely satisfactory. This is due to the high complexity of natural language. To cope with these limits, our group proposes several different and complementary solutions.

We are highly interested in language model adaptation to improve speech recognition quality. In our case, language models are adapted to the topic of the utterance. Topic identification consists in assigning a label to an utterance, among a set of predefined topics. During speech recognition, given the set of words recognized, the topic is identified and the corresponding language model is used for the next words to be recognized [38]. Work will be oriented now towards other topic identification methods [14] and efforts have to be done to improve methods based on dynamic adaptation of the probabilities of the specific models

We proposed a new model (Statistical Feature Language Model) which can take into account a maximum of word features as gender and number. For that, a word is considered as a feature vector in which the word itself, its syntactic class, its gender, ...are integrated. In other words, this approach considers a word as a complex object which is related to other complex objects. First experiments show an improvement in terms of perplexity and Shannon's game [51]. We used the Selected History Principle in order to select, for each history, the best subset of features. In a first step, we automatically tagged a corpus with gender and number features. Then, we evaluated all the subsets of features for all the histories, and applied the selection principle. The conclusion is that the tagging quality has to be improved (ambiguities are frequent in French) and we need to find a better way to measure the amount of information provided by each feature.

Speech understanding is another research activity in which we are involved. The automatic speech understanding problem could be considered as an association problem between two different languages. The request expressed in natural language is transformed in terms of concepts. A concept represents a given meaning and is defined by a set of words sharing the same semantic properties. We proposed to use a naive bayesian classifier to automatically extract the underlined concepts. We also propose a new approach for the vector representation of words. This step allows to validate our speech understanding approach. In fact, a test corpus automatically rewritten in terms of concepts has been transformed on SQL requests and achieved a result of 92.5% of well formed SQL requests. The understanding process has been integrated in our speech recognition system ESPERE. The first results are very encouraging and show the robustness of the proposed method [23].

Work is pursued to integrate the concept of impossible events in a speech recognition system. This new and original idea tries to cut off all the impossible linguistic events from the classical language models. The challenge consists in finding out automatically these events [47].

Another research activity in which we are concerned by is the development of a new framework for combining language models. This framework is based on a dynamic bayesian network (c.f. § 6.2.4). In this respect we use the DBNs framework in order to achieve a better exploitation of each linguistic unit considered in modeling. We develop a unifying approach that processes each of these units in a unique model and construct new data-driven language models with improved performances. The principle of this approach is to construct 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 is reported in [41], [42]. We decided to pursue this work in the framework of a PhD thesis.

The speech technology is ready to support new services. One of the most attractive applications is without doubt the speech-to-speech translation. A new PHD thesis started: we implemented the model 1 proposed by

Brown [37] to train the translation model. An original method has been also implemented to train the language model. This method based on statistic approach permits to take into account the agreement in terms of gender and number. It improves the performance of a classic trigram model. Finally, in order to find the best target sentence given the source sentence, we developed a decoder based on the Viterbi algorithm. To evaluate the machine translation quality, we choose to adopt the BLEU method proposed by Papineni [49]. It is a quick, inexpensive and language independent method which substitutes for human judges when there is need for quick or frequent evaluations.

7. Contracts and Grants with Industry

7.1. National Contracts

7.1.1. TNS project

TNS is the French company which monitors all types of media: written press, radio, television, news agencies, Internet. It collects, selects, analyses, organizes and transmits information. It aims to automatically detect key information useful for its customers. In this framework, we have started the CIFRE thesis of Emmanuel Didiot. Last year, a prototype has been realized and currently, we improve acoustic models and segmentation modules (see section 6.2.2.2).

7.1.2. STORECO project

This project is founded 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 caption for TV programs. To do that, we will use algorithms developed for the automatic speech recognition.

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, we have implemented software to perform clustering and training of acoustic models for segmentation.

7.1.3. LABIAO project

This project started in January 2005 is founded 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 provide hard of hearing people with an artificial talking head adapted to cued speech, i.e. incorporating a hand, piloted through automatic speech recognition. Our contribution is twofold. First we are developing a continuous speech recognition system at the phonetic level. Indeed, it seems better to provide subjects with phonetic symbols (coded in cued speech) even with errors rather than with words. The weakness of the second solution lies in the presence of unknown words that compromise the recognition results. For this purpose we developed a recognition system with only a very small delay (less than one second) with respect to the original speech. This required the Viterbi algorithm to be revisited to reduce the scope of exploration in time. Second, we are developing an algorithm that predicts labial coarticulation phenomena in order to generate relevant face deformations. We already recorded a corpus of 3D face data of a female speaker. These data will be used to train the prediction algorithm and the evaluation will be carried out in 2006. The PhD of Matthieu Camus (started on March 15th) is dedicated to the face animation. Alexandre Lafosse is working on the speech recognition framework (adaptation to speaker, evaluation, training...).

7.1.4. 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 speech recognition 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 automatic speech recognition will be performed and the result will be manually corrected.

7.1.5. 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 founded 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 that database, one thousand of different children will be recorded.

The second is an extensive telephone database of grown up 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. Thus, the collection of the Neologos corpus is a three stage process:

1. the BOOTSTRAP database is collected by recording a first set of 1,000 different speakers over the fixed telephone network. The recorded utterances are a set of 45 phonetically balanced sentences, identical for all the speakers and recorded in one call. Such sentences are optimized to facilitate the comparison of the speaker characteristics;
2. a subset of 200 reference speakers is selected through a clustering of the voice characteristics of the 1,000 bootstrap speakers;
3. the final database of 200 reference speakers, called IDIOLOGOS, is collected. The reference speakers are requested to pronounce a large corpus of 450 specific sentences, identical for all the speakers.

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 each of IRISA, FRANCE TELECOM and LORIA designed a specific inter-speaker dissimilarity measure. The obtained lists of reference speakers were compared and jointly optimized [19].

7.2. International Contracts

7.2.1. 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.

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:

- Bayesian networks: we begin to design a toolkit for training Bayesian Network models (see section 5.1.9).
- Missing data: we propose original approaches to deal with non stationary noise (see section 6.2.1.2).
- Non native speech recognition: we combine acoustic models from the spoken language with acoustic models from the mother language (see section 6.2.1.3).

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.

7.2.2. 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 Langue & Dialogue 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 past year, we mainly participated to the design of the Intelligent User Services layer architecture. We thus proposed a taxonomy of user interaction modalities that gives as much genericity as possible concerning the interaction devices that are supported by the platform: indeed, in Ambient Intelligence, the number of possible interaction devices is very large and is continuously increasing. We further proposed architectural solutions to handle both implicit and explicit user interactions. In the next step, we plan to focus our efforts on implicit speech inputs.

7.2.3. 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 the 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 on the one hand and machine learning. Our contribution will be on the development of acoustic-to-articulatory inversion and the improvement of the robustness of automatic speech recognition through the use of Bayesian networks.

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

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

7.2.4. France-Berkeley cooperation with Perception Science Laboratory at UCSC

This project involves the accurate generation of relevant lip deformations and jaw movements of artificial talking heads as the latter perform significantly poorly than real speakers. This issue is particularly crucial to improve lip reading by deaf people and to learn the articulation of phonemes that do not exist in the native language in the case of language learning. The expected contributions of this project are to improve the modeling of labial coarticulation of Baldi (a talking head developed by Dominic Massaro and Michael Cohen at the Perceptual Sciences Laboratory PSL, university of California at Santa Cruz) in English and French and to evaluate the benefit of using this talking head for native speakers learning English as a foreign language, and for hard of hearing or deaf people learning and/or performing lip reading.

We will exploit the data being acquired by using the tracking system designed by the Magrite team. Within the context of language learning the work will consist of investigating how Baldi can be used to make the learner more sensitive to acoustical and articulatory features of both French and English sounds. This work will exploit standard phonetic knowledge of French and English pronunciation together with the available articulatory data.

The collaboration will mainly rely on sharing coarticulation data acquired by the other team, organizing complementary research efforts and evaluating the use of a talking head for language learning and lip reading.

8. Other Grants and Activities

8.1. Regional Actions

8.1.1. Assistance to language learning. Action from the "Plan État Région" project

The aim of the project is to design a computer-assisted learning system of English prosody for French students. The development of this system has been achieved in the framework of a project supported by our region and gathering scientists from different domains (phonetics, automatic speech processing, ergonomy and language learning).

The system exploits signal visualization and transformation techniques that are intended to be used by teachers of foreign languages in their courses. In order to develop a method for the automatic alignment of text to speech available for non-native speech, we have recorded sentences extracted from the Timit corpus and uttered by young French speakers of a high college and two universities of Nancy. About 2000 sentences have been collected.

Besides signal processing and automatic speech recognition tools, our system includes a course on prosody designed for teachers and will contain a database of characteristic sentences.

A set of progressive exercises have been designed by teachers of English as a foreign language. These exercises exploit our speech tools (modifications of prosodic cues, filtering of speech signals) and the facilities of Snorri. Their aim is to make learners aware of prosody in general (lexical stress, rhythm and intonation) and French and English prosodies in particular by listening, visualisating and exaggerating errors and targets to be reached from their own productions.

A database of sentences uttered by two native speakers, has been recorded this year. It will serve as an illustration of English prosody and as a base of comparison to judge learners' production. The corpus is made up of small texts, transparent words, and sentences varying in the location of the focus accent. A project devoted to the automatic detection of the lexical accent has been initiated (see section 3.2.2).

8.1.2. Improvement of a talking head for cued speech

This project is about the improvement of the labial coarticulation. The first part of the work focused on the design of a tracking system of 3D markers painted on the face of one subject. This system has been developed by the Magrite team and it enables the tracking of 150 markers from stereo images acquired at the rate of 120 fps. We are now developing a prediction model for labial coarticulation.

The second part of this work is to drive the talking head by using automatic speech recognition. This year efforts were about the speech/non-speech detection. This project is a cooperation with the association DATHA and involves a development action of INRIA that aims at adding talking head functionalities to the Graphite software developed by Bruno Lévy (Alice project). This project shares the same objectives as the LABIAO project mentioned above.

8.2. National Actions

8.2.1. ESTER Project

As, in USA, NIST organizes every year an annual evaluation of the systems performing an automatic transcription of radio and television broadcast news, the French association AFCP (Association Francophone de la Communication Parlée) has initiated such an evaluation for the French language, in collaboration with ELRA (European Language Resources Association) and DGA (Délégation Générale pour l'Armement). The ESTER (Évaluation des Systèmes de Transcriptions Enrichies des émissions Radiophoniques) project is supported by the French research ministry (CNRS-Technolangue-EVALDA) for two years. ESTER is composed of two evaluation phases. Phase one evaluates the segmentation systems, as speech/music segmentation, the speaker tracking systems and the orthographic transcription systems. We have decided to participate in the evaluation of the orthographic transcription task and of the acoustic segmentation task (as speech/music/noise). We have developed a fully automatic transcription system (Automatic News Transcription System: ANTS) containing

a segmentation module (speech/music, broad/narrow band, male/female) and a large vocabulary recognition engine (see section 5.1.8).

The first evaluation has been conducted in January 2005. The word error rate (WER) for the transcription task are:

Labs	WER (% : low values are better)
LIMSI	11.9
LIUM	23.6
LIA	26.7
LORIA	27.6
IRISA	35.4
CLIPS	40.7
ENST	45.4
IRIT	61.9

9. Dissemination

9.1. Animation of the scientific community

The members of Parole are involved in several committee programs and scientific review panels

- 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.
- A. Bonneau is an elected member of the Instil Board (Integration of speech technology in learning). She is in charge of the project “assistance to language learning” of the “Plan État Région” and member of Eurospeech scientist committee.
- J.P. Haton is a member of CSL and ICSLP program commitee, chairman of French Science and Technology Association.
- Y. Laprie is a member of (LREC, JEP) scientific committee. He is in charge of the “Assistant intelligent” project of the PRST “Intelligence Logicielle” and associate editor of the EURASIP Journal on Audio, Speech, and Music Processing.
- O. Mella, D. Fohr, I. Illina and C. Cerisara are involved in several European and national projects.
- K. Smaïli is a member of (Eurospeech, JEP) scientific committee.
- K. Smaïli has been invited as lecturer at TISR Conference.
- I. Illina is a member of the evaluation commission of l’INRIA.
- I. Illina is a member of AFCP board.

9.2. Distinctions

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

9.3. Invited lectures

- Dan Istrate, LIA (Laboratoire d'Informatique d'Avignon),
- Catherine Pelachaud, IUT de Montreuil, University of Paris 8,
- Shinji Maeda, ENST, Paris,
- Richard Beaufort, Multitel, Mons, Belgium,
- Sofia Ben Jebara, SUP'COM, Tunis, Tunisia.

9.4. Higher education

- A strong involvement of the team members in education and administration (University Henri Poincaré, University Nancy 2, INPL): Master of Computer Science, IUT, MIAGE;
- Head of teaching and research unit (UFR) STMIA (Sciences et Techniques Mathématiques, Informatique, Automatique) (M.-C. Haton),
- Head of MIAGe department (K. Smaïli),
- Head of Network Speciality of University Henri Poincaré Master of Computer Science (O. Mella).

9.5. Participation to workshops and PhD thesis committees:

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

10. 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] A. BONNEAU. *Identification of vocalic features from French stop bursts*, in "Journal of Phonetics", 2001.

[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] M.-C. HATON. *Issues in Using Models for Self Evaluation and Correction of Speech*, in "Computational Models of Speech Pattern Processing, Berlin", M. PONTING (editor). , Computer and Systems Sciences, Springer-Verlag, 1998.

[7] I. ILLINA, M. AFIFY, Y. GONG. *Environment Normalization Training and Environment Adaptation Using Mixture Stochastic Trajectory Model*, in "Speech Communication", vol. 24, 1998.

[8] J.-C. JUNQUA, J.-P. HATON. *Robustness in Automatic Speech Recognition*, Kluwer Academic, 1996.

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

[10] I. ZITOUNI, K. SMAÏLI, J.-P. HATON. *Statistical Language Modeling Based on Variable-Length Sequences*, in "Computer Speech and Language", vol. 17, n° 1, Jan 2003, p. 27-41.

Articles in refereed journals and book chapters

[11] M. CADOT, J.-B. MAJ, T. ZIADÉ. *Association Rules and Statistics*, in "Encyclopedia of Data Warehousing and Mining", J. WANG (editor). , vol. 1, Idea Group Inc., 2005, p. 74–77.

[12] S. OUNI, M. M. COHEN, D. W. MASSARO. *Training Baldi to be multilingual : A case study for an Arabic Badr*, in "Speech communication", vol. 45, n° 2, 2005, p. 115-137.

[13] 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, http://hal.ccsd.cnrs.fr/ccsd-00008682/en/.

Publications in Conferences and Workshops

[14] 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.

[15] V. BARREAUD, D. O'SHAUGHNESSY, J.-G. DAHAN. *Experiments on Speaker Profile Portability*, in "Proceedings of the 9th European Conference on Speech Communication and Technology - Interspeech - Eurospeech 2005, Lisbon, Portugal", 2005, p. 997–1000.

[16] G. BOUSELMI, D. FOHR, I. ILLINA, J. PAUL HATON. *Fully Automated Non-Native Speech Recognition Using Confusion-Based Acoustic Model Integration*, in "Proceedings of INTERSPEECH / EUROSPEECH 2005", ISCA, 2005, p. 1369-1372.

[17] A. BRUN, C. CERISARA, D. FOHR, I. ILLINA, D. LANGLOIS, O. MELLA. *ANTS le système de transcription automatique du LORIA*, in "Workshop Ester, France Avignon", 2005, http://hal.ccsd.cnrs.fr/ccsd-00013961/en/.

[18] M. CADOT. *A Simulation Technique for extracting Robust Association Rules*, in "3rd IASC world conference on Computational Statistics & Data Analysis, Limassol, Chypre", 2005.

[19] D. CHARLET, S. KRSTULOVIC, F. BIMBOT, O. BOEFFARD, D. FOHR, O. MELLA, F. KORKMAZSKY, D. MOSTEFA, K. CHOUKRI, A. VALLÉE. *Neologos: an optimized database for the development of new speech processing algorithms*, in "proceedings of EUROSPEECH/INTERSPEECH 2005", ISCA, 2005, p. 1549-1552.

[20] 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/.

[21] K. DAOUDI, C. CERISARA. *The MAP-SPACE denoising algorithm for noise robust speech recognition*, in "Proc. ASRU, Cancuun, Mexico", 2005.

[22] 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.

[23] S. JAMOUSSI, K. SMAÏLI, J.-P. HATON. *From speech to SQL queries : a speech understanding system*, in "AAAI Workshop on Spoken Language Understanding - SLU 2005, Pittsburgh, Pennsylvania", 2005.

[24] S. JAMOUSSI, K. SMAÏLI, J.-P. HATON. *Une approche neuronale pour la compréhension automatique de la parole*, in "Traitement et Analyse de l'Information : Méthodes et Applications - TAIMA'2005, Hammamet, Tunisie", 2005.

[25] P. KRAL, C. CERISARA, J. KLECKOVA. *Combination of classifiers for automatic recognition of dialog acts*, in "Proceedings of the 9th European Conference on Speech Communication and Technology - Interspeech - Eurospeech 2005 - Lisbon, Portugal", 2005, p. 825-828, http://hal.ccsd.cnrs.fr/ccsd-00013940/en/.

[26] P. KRAL, J. KLECKOVA, C. CERISARA. *Sentence modality recognition in French based on prosody*, in "VI. International Conference on Enformatika, Systems Sciences and Engineering - ESSE 2005", 2005, p. 185–188, http://hal.ccsd.cnrs.fr/ccsd-00013968/en/.

[27] J.-B. MAJ, A. BONNEAU, D. FOHR, Y. LAPRIE. *An elitist approach for extracting automatically well-realized speech sounds with high confidence*, in "Proceedings INTERSPEECH / EUROSPEECH 2005", ISCA, 2005, p. 2925-2928.

[28] S. OUNI, M. M. COHEN, D. W. MASSARO, H. ISHAK. *Visual Contribution to Speech Perception : Measuring the Intelligibility of Talking heads*, in "Auditory-Visual Speech Processing - AVSP'05, British Columbia, Canada", 2005.

[29] S. OUNI. *Can We Retrieve Vocal Tract Dynamics that Produced Speech? Toward a Speaker Articulatory Strategy Model*, in "Interspeech 2005 - Eurospeech, Lisbon, Portugal", 2005, p. 1037–1040.

[30] B. POTARD, Y. LAPRIE. *Using phonetic constraints in acoustic-to-aticulatory inversion*, in "proceedings of EUROSPEECH/INTERSPEECH 2005", ISCA, 2005, p. 3217-3220.

[31] J. RAZIK, O. MELLA, D. FOHR, J.-P. HATON. *Local Word Confidence Measure Using Word Graph and N-Best List*, in "proceeding of EUROSPEECH/INTERSPEECH 2005", 2005, p. 3369-3372, http://hal.ccsd.cnrs.fr/ccsd-00013775/en/.

[32] V. ROBERT, B. WROBEL-DAUTCOURT, Y. LAPRIE, A. BONNEAU. *Inter-speaker variability of labial coarticulation with the view of developing a formal coarticulation model for French*, in "Proceedings of

International Conference on Auditory-Visual Speech Processing (AVSP'05), Vancouver", July 2005, p. 65–70.

[33] V. ROBERT, B. WROBEL-DAUTCOURT, Y. LAPRIE, A. BONNEAU. *Strategies of labial coarticulation*, in "proceedings of EUROSPEECH/INTERSPEECH 2005", September 2005, p. 1021-1024.

[34] B. WROBEL-DAUTCOURT, M.-O. BERGER, B. POTARD, Y. LAPRIE, S. OUNI. *A low-cost stereovision based system for acquisition of visible articulatory data*, in "Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'05), Vancouver", July 2005, p. 145–150, http://hal.inria.fr/inria-00000432/en/.

Bibliography in notes

[35] 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° 4, 1995, p. 85–89.

[36] 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° 1, Jul 1996, p. 555-564.

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

[38] A. BRUN, K. SMAILI. *Fiabilité de la référence humaine dans la détection de thème*, in "Proceedings of the Traitement Automatique des Langues Naturelles (TALN) Conference, Fès, Maroc", 2004.

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

[40] 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.

[41] M. DEVIREN, K. DAOUDI, K. SMAÏLI. *Une nouvelle approche de modélisation du langage par des réseaux Bayésiens dynamiques*, in "XXVes Journées d'Etudes sur la Parole - JEP-TALN-RECITAL 2004 , Fès, Maroc", Apr 2004, http://www.loria.fr/publications/2004/A04-R-098/A04-R-098.ps.

[42] M. DEVIREN. *Systèmes de reconnaissance de la parole revisités : Réseaux Bayesiens dynamiques et nouveaux paradigmes (Revisiting speech recognition systems : dynamic Bayesian networks and new computational paradigms)*, Thèse d'université, Université Henri Poincaré, Oct 2004.

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

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

[45] I. ILLINA, D. FOHR, O. MELLA, C. CERISARA. *The Automatic News Transcription System : ANTS some Real Time experiments*, in "8th International Conference on Spoken Language Processing - ICSLP' 2004, Jeju, South Corea", October 2004.

[46] S. JAMOUSSI, K. SMAÏLI, D. FOHR, J.-P. HATON. *A complete understanding speech system based on semantic concepts*, in "4th International Conference on Language Resources and Evaluation - LREC'04, Lisbonne, Portugal", vol. 5, May 2004, p. 1615-1618.

[47] 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.

[48] 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.

[49] 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.

[50] L. ROMARY, A. TODIRASCU, D. LANGLOIS. *Experiments on Building Language Resources for Multi-Modal Dialogue Systems*, in "International Conference on Language Resources and Evaluation - LREC'2004, Lisbonne, Portugal", vol. 2, May 2004, p. 533-536.

[51] K. SMAÏLI, S. JAMOUSSI, D. LANGLOIS, J.-P. HATON. *Statistical Feature Language Model*, in "International Conference on Speech and Language Processing - ICSLP' 2004, Jeju, Corée du Sud", October 2004.