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 a common project to INRIA, CNRS and Henri Poincaré University through LORIA laboratory (UMR 7503).

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, U. Nancy 2]
Vincent Colotte [Assistant Professor, U. H. Poincaré]
Joseph di Martino [Assistant Professor, U. H. Poincaré]
Jean-Paul Haton [Professor, U. H. Poincaré, Institut Universitaire de France]
Marie-Christine Haton [Professor, U. H. Poincaré]
Irina Illina [Assistant Professor, I.U.T Charlemagne, U. Nancy 2, working for INRIA since september 2004]
David Langlois [Assistant Professor, IUFM]
Odile Mella [Assistant Professor, U. H. Poincaré, working for CNRS until september 2004]
Slim Ouni [Maître de conférences, I.U.T Charlemagne, U. Nancy 2, since the first of september, 2004]
Kamel Smaïli [Professor, U. Nancy 2]

Phd Students

Vincent Barreaud [TA]
Yassine Benayed [TA]
Emmanuel Didiot [CIFFRE grant]
Murat Deviren [TA]
Salma Jamoussi [TA]
Fabrice Lauri [CIFRE grant]
Vincent Robert [High school teacher]
Pavel Král [Czech coPhD]
Blaise Potard [MENRT grant]
Joseph Razik [INRIA grant]
Ghazi Bousselmi [TA]
Sébastien Demange [TA]

Project technical staff

Matthieu Camus [INRIA, project "Plan Etat Région"]

Post Doctoral fellow

Jean-Baptiste Maj [INRIA]

Research scientist partner INRIA

Filipp Korkmazsky [until september, 2004]

2. Overall Objectives

PAROLE is a common project to INRIA, CNRS and Henri Poincaré University through the LORIA laboratory (UMR 7503). The purpose of our project is to automatically process speech signals to understand their meaning, or to analyse and enhance their acoustic structure. It inscribes within the view of offering efficient vocal interfaces and necessitates works in analysis, perception and automatic recognition of speech.

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 themas 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 scientific culture is pluridisciplinary and 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 competence in automatic speech recognition and phonetics.

Our policy in terms of industrial partnership consists in favouring contracts that quite precisely fit our scientific objectives. We are involved in several cooperations with companies using automatic speech recognition, as DIALOCA. We have a cooperation with SyncMagic Procoma in the form of an RNRT project. We recently had a contract with Lipsync, Thales Aviation. We are conducting a study on non-native speech recognition in a noisy environment. Babel Technologies retails our speech analysis software WinSnoori as other companies. Moreover, we are involved in the 5th PCRD projects OZONE and MIAMM 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 approach:

- 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 (spectal analysis, stochastic acoustic or linguistic models).

The former research topic is motivated by the high specificity of speech among other acoustical signals: the speech production system is easily accessible (at least at first approach); acoustical equations are reasonably difficult from a mathematical point of view (with simplifications that are moderately restrictive); sentences built 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 production and perception of speech do not enable using an analysis by synthesis approach for automatic speech recognition. Automatic speech recognition thus aroused a second approach

[1] Hidden Markov Models

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 second borrows theoretical results on speech from the first, 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 must be increased.

Our activities are structured according to these two approaches:

- **Speech analysis.** Our works are about automatic extraction and perception of acoustic cues, acoustic-to-articulatory inversion and speech analysis. These themas 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 about stochastic models (HMM[2], bayesian networks and missing data models), multiband approach, adaptation of a recognition system to a new speaker or to the communication channel and on language models. These themas give rise to a number of ongoing or future applications: automatic speech recognition, automatic translation, text-to-speech alignment, audio indexing.

3.2. Speech Analysis

Participants: Anne Bonneau, Vincent Colotte, Dominique Fohr, Jean-Paul Haton, Marie-Christine Haton, Yves Laprie, Jean-Baptiste Maj, Joseph di Martino, Slim Ouni.

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 recongition and the improvement of the oral component of language learning.

3.2.1. Acoustic cues

We have introduced the notion of strong and weak cues 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 [43], 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 fulfil 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

[2]Hidden Markov Models

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. Another approach consists in enhancing only the sounds that are well realized, so as to provide listeners some islands of confidence during the acoustic decoding stage.

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 modelise those well identified sounds in the following way. First, on a training corpus, the system modelises 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 [36].

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 7 (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, we improved the method in the two following ways. Firstly, we introduced a pruning algorithm to seek analysis marks (for pitch synchronization). It increases the robustness of pitch marking for speech segments with strong formant variation. Secondly, 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 [45].

3.2.2.2. Perceptual experiments

In order to improve oral comprehension, we developed speech transformation tools which slow down the rate of speech and enhance some acoustical cues. To avoid the introduction of acoustical artefacts which may deteriorate sound identification, we elaborated a strategy based on the enhancement of voiceless consonants and fast spectral transitions. A first experiment showed that our transformations improve significantly the comprehension of french sentences for foreign students. We prepare this year a new experiment to validate our approach with two other objectives. We will test our modifications on isolated sounds in order to adjust our transformation and possibly discard those which could have too bad effect on the identification of some sounds. That is why we are building a new corpus of logatoms (VCV). The second goal is to show that our strategy improves continuous speech intelligibility. We are currently recording a short text about weather forecast. Perception experiments will start when corpora will be built and transformations applied.

3.2.3. Acoustic-to-articualtory 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 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 that built by Maeda [50].

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 in two families:

- methods that optimize a function generally combining speaker's articulatory effort and acoustical distance between natural and synthesized speech. They exploit number of 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 compute 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.

3.3. Automatic speech recognition

Participants: Dominique Fohr, Jean-Paul Haton, Irina Illina, Odile Mella, Kamel Smaïli, Christophe Antoine, Armelle Brun, Christophe Cerisara, David Langlois, Khalid Daoudi, Yassine Benayed, Murat Deviren, Fabrice Lauri, Vincent Barreaud, Salma Jamoussi, Sen Zhang, Filipp Korkmazsky, Joseph di Martino, Joseph Razik, Emmanuel Didiot, Pavel Kral.

Figure 1 shows the different components which are required in the automatic speech recognition process. It also introduces the research topics of automatic speech recognition which we are working on: language modeling, acoustic modeling, robustness and invariance to different speakers or various environments (as noisy or spontaneous speech) with adapting or compensating methods, speech/non speech segmentation. Despite the

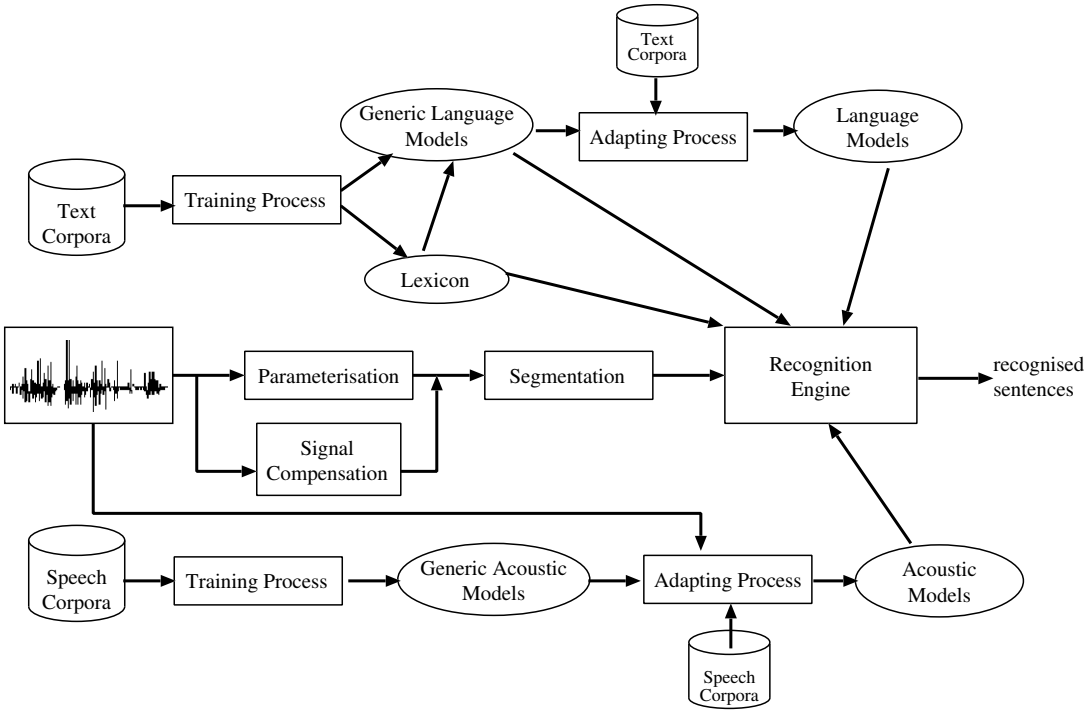

Figure 1. Speech recognition process

fact that all of these components are tightly linked, to be clear we gather our research activities in two sections: “Acoustic features and models” and “Language models”.

3.3.1. Acoustic features and models

3.3.1.1. Acoustic models

Stochastic models are now the most popular approach for automatic speech recognition. We focus our research work on Hidden Markov Models (HMM) and Bayesian Networks (BN). First we have designed a toolkit ESPERE for HMM models (a training tool and a recogniser engine). We have then elaborated several automatic speech recognition and text-to-speech alignment systems with this tool. Now, we use HMM models in order to validate the new algorithms we develop for speaker and noise adaptation, noise robustness and segmentation. We also work on more powerful models: Bayesian Networks. The formalism of Bayesian networks consists in associating a directed acyclic graph and a numerical parameterisation to the joint probability distribution (JPD) of a set of random variables. The nodes of the graph represent the random variables, while the arrows encode the conditional independences which (are supposed to) exist in the JPD. Once the graphical structure is specified, the numerical parameterisation is given by the conditional probabilities of each variable given its parents. HMMs are particular instances of *dynamic* Bayesian networks. Thus, the latter provide a more general theoretical and computational framework to develop and process new models. They are able to represent and handle speech features with higher flexibility than HMMs.

3.3.1.2. Robustness and invariance

Mismatch between the training and testing conditions may result from different sources. But the two most important ones are (i) the background noise and (ii) the speaker variability. Several state of the art methods exist to deal with either one kind of mismatch or both. Amongst those, the following ones serve as basis of our research work:

- 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. At the difference of the two previous methods, it does not require a large amount of adaptation data - about one second 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.

3.3.1.3. Segmentation

Audio indexing and automatic broadcast news transcription need the segmentation of the audio signal. The segmentation task consists in two steps: first homogeneous segments are extracted and classified into speech, noise or music, secondly, speakers turns are detected on 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 centroïd, ...). Except the selection of acoustic features, another point is to find the good classifier. Various classifiers are commonly used: k-Nearest-Neighbours, Hidden Markov Models, Gaussian Mixture Models, Artificial Neural Networks.

As to detect speaker turns, the main approach is to split 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

Acoustically, we can consider that several systems today achieve good results. Nevertheless, some problems due to the complexity of natural language remain without a satisfactory solution. Our group, as others through the world, makes more and more efforts in order to make them more efficient. All the language models we propose are based on information theory and statistics. Some of them use linguistic knowledge to guide the statistical one. The current state of the art in this domain shows that the majority of language models fit for use in speech recognition have a very narrow scope. Some of them use a history which is more or less distant as cache or triggers models. Even if the combination of these models with the baseline ones achieves better results, we have to improve them in order to take into account the wide complexity of natural language. 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.
- Modeling distant relationship. This is necessary because in natural language the relationship between linguistic units is not necessarily contiguous but may be in some cases distant. For instance, a verb could be linked to a subject which occurred n words (with $n > 3$) before. The research activity consists in modelizing these distant relationships and finding the best framework for that.
- Speech understanding process. We believe that speech recognition system will be more efficient when the recognition process will be connected to an understanding one. For that purpose, we work on a new system based on a naïve bayesian classifier which allows for translating a signal to conceptual tags which in turn are translated to SQL requests [47].

Other related activities in language modeling not described here are supported by our team.

4. 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 (e.g., for hearing-impaired persons or for foreign language teaching) as well as for hearing aids. We have developed in the past a set of teaching tools based on the speech analysis and recognition algorithms of the group (cf. the ISAEUS [46] project of the EU that ended in 2000). We are continuing this effort toward 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 tools

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 10 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 filterings) because the spectrogram allows the acoustical consequences of all the modifications to be evaluated. Beside this set of basic functions there are various functionnalities 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 sofware resource for several works in our team (formant tracking, stop identification, perceptive studies,...). Given the interest it represents for speech analysis we distribute it to about fifteen frenchspeaking 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 the reassignment spectrogram calculation proposed by Plante and al. [51] that enables a better localization of the spectrogram energy. The reassignment gives sharper harmonics when it is applied to narrow band spectrograms together with a better time precision for rapid speech events as burst releases. The reassignment also gives a better localization of glottal closure events when it is applied to large band spectrograms.

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 dictionnary (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

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

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 markers which indicate the series of mouth shapes to bedrawn. 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.

5.2. Corpus

The research performed in speech communication needs to record, clean, and label wide text and speech corpora. For example, for an investigation about phonetic cues, it is necessary to record and phonetically label several sentences in order to capture the contextual effect. These sentences have to be pronounced by different speakers to take into account the inter-speaker variabilities.

Several years ago, we developed tools allowing speech corpora to be edited, processed and manually labelled (section 5.1.2).

Another example concerns the constitution and the labelling of speech corpora for automatic speech recognition. These corpora are used to train the acoustic models and to test them. To train the statistical acoustic models, a large number of labelled speech data are necessary. In general, huge corpora are necessary in order we get efficient models. These corpora cannot be annotated manually. Our speech team developed several tools for semi-automatic labelling speech data (5.1.5).

In the same way, training of statistical language models requires huge text corpora. For example, in the scope of the dictation machine (project AUPELF-UREF), bigram and trigram models have been trained using 50 million words corpus extracted from two years issues of “Le Monde”, the French newspaper.

Size of text corpora are constantly increasing. For French, 16 years (300 millions words) extracted from “Le Monde” are now available in our team and used in the new project ESTER in which we are involved (see section 8.2.4).

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*.

6.1.1. Acoustic-to-articulatory inversion

The strength of our inverse method resides at the quasi-uniform acoustic resolution of the articulatory table. This property originates in the construction method that evaluates the linearity of the articulatory-to-acoustic mapping at each step. Articulatory parameters of Maeda's model vary between -3 and 3 σ where σ is the standard deviation. Thus, the codebook inscribes a root hypercube. Sampling the articulatory space amounts to find 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 inverse articulatory parameters generate original formant trajectories with high precision and a realistic sequence of the vocal tract shapes.

This year we continued the evaluation of this inversion method [39] and we mainly focused on the phonetic relevancy of inverse solutions. The main strength of the hypercube approach is to give all the inverse solutions corresponding to an acoustic entry, i.e. a 3-tuple of formants. However, some of these inverse solutions are not realistic from a phonetic point of view because they do not satisfy expected phonetic cues, for instance no protrusion for vowels close to /y/. We thus expressed standard phonetic knowledge in terms of articulatory parameters. This plays the role of phonetic constraints that enable the ranking of inverse solutions. These constraints are about one or several articulatory parameters when there can be some compensatory articulatory effect between several parameters as jaw and tongue positions. We tested these constraints on the recovery of place of articulation for the french vowels. It turns out that this subtantially improves the quality of inverse solutions by removing incorrect places of articulation for vowels. Furthermore, this gives a smooth ranking from very likely to very unlikely when considering the results in term of places of articulation and not only articulatory parameters.

6.1.2. 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 a postdoctoral stage at the Multitel research center (in Belgium) by Vincent Colotte. 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 new 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 to read aloud. Second, linguistic features used for selecting speech units are automatically weighted thanks to an original, entropy-related method [25].

This year, we made an evaluation of the system with the collaboration of MULTITEL center. 50 subjects listened to 25 French sentences. Among them, 20 sentences were synthesized by LiONS, and 5 directly came from the database. Aims of the evaluation were to evalute the *intelligibility*, the *naturalness of the prosody*, the *quality of the concatenation* and the *listening comfort*. A secondary goal was to evaluate the distance between *synthetic* and *original voices*. Explanation given to the subjets and results of the evaluation can be found on the web page http://www.multitel.be/TTS/LiONS/evaluation.html.

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.

6.1.3. Automatic formant tracking

In [49] we showed how active curves could be used to track formants. The underlying idea is to deform initial rough estimates of formants under the influence of the spectrogram to get regular tracks close to lines of spectral maxima which are potential formants. A formant track is thus represented by a curve $t : [t_i, t_f] \rightarrow I\!R^2$, $t \rightarrow (t, F(t))$ in the time frequency domain, t_i and t_f are times of the beginning and end of the formant track,

and $F(t)$ is the frequency of the formant at time t. The compromise between proximity to spectral peaks and regularity is given by the following functional E which has to be minimized

$$\begin{array}{rl} E(F) = & -\int_{t_i}^{t_f} E_{Spectro}(t, F(t))dt \\ & +\lambda \int_{t_i}^{t_f} \alpha|F'(t)|^2 + \beta|F''(t)|^2 dt \end{array} \quad (1)$$

where the overall energy $E(F)$ has to be minimized. The first term represents the spectrogram energy $E_{Spectro}$ along the formant track. It is thus all the bigger since the curve is close to a line of spectral peaks. The second term represents the length and the curvature and is thus all the smaller since the curve is regular. α influences the curve length, β its curvature and λ the compromise between the spectrogram energy explained by formant tracks and the smoothness of the curve.

Each formant curve becomes deformed under the influence of the spectrogram independently of the other formant curves, what requires a complex control strategy [49] to manage interactions between formants. The main difficulty is when two formants are competing with each other to catch the energy of one spectral peak. This problem occurs when one spectral peak is too weak compared to the other and leads the two formants tracks to get closer to the prominent spectral peak. Another difficulty is the initialization of the tracks that requires the construction and the labelling of elementary tracks, i.e. small pieces of formant tracks obtained by applying a simple continuity constraint, in terms of formants.

We thus designed a new strategy [37] that incorporates repulsion forces to keep formant tracks away form each other. This is achieved by incorporating a repulsive term in Eq. 1 which becomes

$$\begin{array}{rl} E(F) = & -\int_{t_i}^{t_f} E_{Spectro}(t, F(t)) \\ & +\mu \sum_n E_{Spectro}(t, F_n(t)) \times exp^{-(\frac{F_n(t)-F(t)}{s_n})^2} dt \\ & +\lambda \int_{t_i}^{t_f} \alpha|F'(t)|^2 + \beta|F''(t)|^2 dt \end{array} \quad (2)$$

where μ is the weight of the repulsive force and n gives the neighboring tracks. For the second formant F_2 n corresponds to F_1 and F_2 and for F_3 n corresponds to F_2. For the first formant F_1 n corresponds to F_2 and an artificial fixed formant (at 0 Hz) to prevent F_1 from becoming negative. This way, formants are deformed by taking into account the deformations of their neighboring formants with two advantages: a better coverage of the spectrogram energy, a simpler and more robust control strategy. Moreover, the initialization stage can be substantially simplified because the interdependency of formant tracks enables an more dynamic exploration of solutions than that possible with the labelling of elementary tracks based on a static strategy. This new strategy turns out to be more efficient than that reported in previous papers.

6.2. Automatic Speech Recognition

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

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

Certainly the most important limiting factors of nowadays speech recognizers are background noise and speaker variability. For example, one privileged application area of automatic speech recognition is cars, in which more and more high-tech devices are embedded, such as navigation systems or hand-free phones. All these technologies rely on speech commands recognition and on its robustness to background noise. One of our objectives is to improve the robustness of the acoustic models (HMM) to noise and to speaker variability. We address these issues through the following.

6.2.1.1. Frequency localized robust feature extraction

State of the art speech feature extraction schemes (Mel frequecy cepstral coefficients (MFCC) and perceptual linear prediction (PLP)) are based on auditory processing on the spectrum of speech signal and cepstral

representation of the resulting features. These algorithms are generally developed considering the properties of HMM formalism. Therefore there is a need for new directions in this field to be able to extract informative features that can later be processed within more generic stochastic modeling paradigms (i.e. dynamic Bayesian networks, c.f. § 6.2.3). With this motivation in mind we seek for potential novel approaches. Precisely, we review wavelet analysis based feature extraction schemes and propose new techniques for frequency localized robust feature extraction using wavelets. The proposed parameterization schemes yield similar or better performances than MFCC in real noise conditions. The results and details of these techniques are summarized in the thesis of Murat Deviren [13]. Our goal now is to use such features to model speech frequency dynamics using dynamic Bayesian networks.

6.2.1.2. Speaker adaptation

Reducing acoustic mismatches due to speaker variability between the training conditions and the testing conditions is a major problem in automatic speech recognition. This problem is particularly difficult for rapid adaptation, when the available amount of adaptation data is small. We have investigated different methods for rapid speaker adaptation. These methods integrate the concepts of both Structural Maximum Likelihood Linear Regression and Eigen-Voices-based technique to adapt the Gaussian means of the speaker independant models for a new speaker.

Two new approaches to rapid speaker adaptation of acoustic models by using genetic algorithms have been proposed in. The first approach consists in using a genetic algorithm to adapt the set of Gaussian means to a new speaker. The second approach uses the genetic algorithm to enrich the set of speaker-dependent systems employed by the EigenVoices. These two approaches have been presented in the PhD thesis of Fabrice Lauri [15].

6.2.1.3. Noise compensation

Two classes of methods exist to deal with noise robustness. The first one, which is called here noise compensation, consists to pre-process the acoustic signal prior to recognition, while the second, model adaptation, rather modifies the acoustic models. We first summarize our work on noise compensation. We developed a frame-synchronous noise compensation algorithm designed to cope with time-varying unknown noise. This method estimates simple mapping function in parallel with Viterbi alignment. We proposed a version of this algorithm that takes into account the abrupt changes in the acoustical environment [18][11]: the environment change is detected using the Shewart Control Charts detection algorithm that searches for the changes in the means of Gaussian sequence. For each noisy environment a specific mismatch function is estimated. A simple bias is used as mismatch function. For various tasks, proposed methods significantly outperform classical compensation/adaptation methods. The results are summerized in the thesis of Vincent Barreaud [11].

A novel approach to speech data normalization by introducing interpolation for histogram equalization is proposed in [34]. Different ways of histogram interpolation that inhence this normalization technique were studied. We found that using a special weighting factor to combine current and past test sentence statistics improved speech recognition performance.

6.2.1.4. Model adaptation

We have proposed several major improvements of the recently appeared Jacobian adaptation method. The most important advantages of Jacobian Adaptation are its very low requirements, both in terms of CPU processing and adaptation data. This is particularly important to be able to quickly adapt the models to a sudden environment change, for example when the car speeds up or down. We have proposed a method to dynamically estimate some parameters of the algorithm. The goal is twofold: to suppress the need of a development corpus, and thus to improve the stability of adaptation. Our work on that topic is summarized in [17].

6.2.1.5. Supervised-predictive noise compensation

We developed a new noise compensation scheme, that we called *supervised-predictive* compensation, which is different in its concept from all known compensation schemes [26]. This scheme can be applied in scenarios where training speech has been recorded in different noise conditions. The principle then is to use a supervised

learning procedure to estimate the parameters of an hypothesized (parametric) model that attempts to describe how matched models vary w.r.t. noise models. This new scheme has many advantages w.r.t. classical adaptive and predictive compensation techniques. Moreover, we showed that it performs significantly better than multi-conditions training, which is the most widely used technique in these kind of scenarios. The results and details of this new approach is summarized in the thesis of Murat Deviren [13].

6.2.1.6. Sentence modality recognition

A PhD student has begun his research work on sentence modality recognition since October 2003. The objective is to recognize automatically whether the input is a question, a declarative or an exclamatory sentence. This information will then be used for example to enrich the talking head when it reproduces/translates someone's talking. This work is realized in collaboration with West Bohemian Czech University. An initial system based on prosodic features such as fundamental frequency and energy has been implemented and tested both in Czech and French. Three different classifiers have been compared: one based on Gaussian Mixture models, another one based on Multi-Layer Perceptron, and the third one combining both previous classifiers. The latter gives the best recognition accuracy, as reported in [35]. The next step consists in including other kinds of information than prosody, such as syntax and semantic to recognise all the questions that have the same prosodic features than declarative sentence.

6.2.1.7. 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 first issue is certainly the most difficult one. After a few preliminary experiments, we have concluded that no single method can solve this problem alone. Our privileged approach is then to combine as many sources of information as possible so as to estimate a probability for every spectro-temporal coefficient to be masked by noise. A PhD student has begun his work on that topic since November 2004. Meanwhile, we have investigated a novel approach to infer the location of the noise in the spectrum, that is based on the information embedded in the clean speech models. Noise models and signal processing techniques have been used so far in the litterature to estimate the spectral bands corrupted by noise. On the other hand, speech energy tends to concentrate into narrow spectral bands, such as formants, which are then caracterized by high local Signal-to-Noise Ratio (SNR) even when the global SNR of the sentence is low. We have shown in [24] that this information can bring effective clues to estimate the optimal masks for MDR.

6.2.2. Core recognition platform

6.2.2.1. Automatic News Transcription System (ANTS)

In the framework of the technolangue project ESTER, we have developped a complete system for French broadcast news transcription. The system is composed of four stages : the broad-band / narrow-band speech segmentation, a speech/music classification, the detection of silences and breathing segments and a large vocabulary speech recognition engine. The aim of the three first stages is to split the audio stream into homogeneous segments with a manageable size and to allow the use of specific algorithms or models according to the nature of the segment [22].

We have carried out several experiments on this system to take into account the specificities of the French language: how accurate should the phones models be and how to deal with the problem of the liaisons between words. As for liaisons, we evaluated different manners to take them into account. Implementing the liaison with a skip transition is the best solution for accuracy as well as for computation time [29].

6.2.2.2. Keyword detection in Broadcast program

In the framework of the CIFRE PhD of Emanuel Didiot with the Presse+ compagny, we begin to implement an automatic system for keywords detection in broadcast news. We chose to use a large vocabulary approach. Sereval problems occur like music in background, advertising, spontaneous speech ...During the first year, a prototype has been realised.

6.2.2.3. Automatic speaker clustering

Automatic speaker clustering is needed for adapting the acoustic models to a speaker in order to improve the recognition accuracy, and for assigning the recognized sentence to the speaker who uttered it in a meeting recording task. So, we have started working on this topic. Our automatic speaker clustering task can be split in two steps:

- an audio stream segmenter which segments the speech signal every time a speaker change occurs. The segmenter uses a distance based on the generalized likelihood ratio ;
- a hierarchical clustering process of these segments based on the Bayesian Information Criterion (BIC).

Ghazi Bousselmi finished his DEA on this subject and now begins a PhD.

6.2.2.4. Keywords detection

Keyword detection allows the detection, in a pronounced sentence, of the keywords characterizing an application and the reject of out-of-vocabulary words as well as hesitations, false starts etc ...Keyword detection is also required for audio indexing. We carried out several studies in this topic.

We propose to use confidence measures in order to make the decision of rejection or acceptance of a given keyword. The confidence measures used are based on the probability of the local acoustic observation. We use these probabilities to calculate the arithmetic, geometric and harmonic means as confidence measures for each keyword.

We present the problem of detection as a classification problem where each keyword can belong to two different classes, namely “correct” and “incorrect”. This classification is carried out by using Support Vector Machines (SVM) which constitute a new technique of statistical training. Each recognized keyword is represented by a characteristic vector which constitutes the entry of the SVM classifier. To determine this vector, we use the probability of the local acoustic observation and then we introduce the duration of each state [20][19][12].

6.2.2.5. 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 criteria like a confidence measure can help the system to determine whether a word sould be kept or not. In the past year, I firstly studied both the confidence measure and the recognition area. And then, we contribute to propose a new confidence measure based on the exploration trellis used by a recognition engine. Experiments have been performed, and preliminary results, compared to reference results using an a posteriori probability, shows that we are on the right track.

6.2.2.6. Speech/Music segmentation

This year, an article was presented at the JEP2004 International conference, dealing with speech/music segmentation in sound documents. It discussed the use of some different parametrizations, based on the MFCC that are often use employed in speech recognition. The problem of robustness was studied and also, a comparison was made between the results of our parametrizations and those of Scheirer and Slaney (on the same corpus they had used) [40].

6.2.3. 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.

6.2.3.1. 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 [13]. Recently, we proposed a new noise robust modeling technique in this framework that takes into account the variation of the acoustic environment [33].

6.3. Language Models

Language modeling is one of the important activities of our team. In spite of all the improvements obtained by the international communauty in this area, the results are not entirely staisfactory. 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.

After showing that using specific vocabularies leads to a large improvement of performance when several topic identification methods are combined [44], we tried to find the limit of the proposed methods. In fact, the combined methods reaches 93.1% of recall. This rate makes us throw the reference labeling back into question. We have then studied the reliability of the topic labeling of our corpora, by using the Kappa statistics and Bayes error. Experiments showed that human labelors tagged differently some paragraphs. This leads to the conclusion that an automatic method can not do better than a human being [23].

The second research domain in language modeling for speech recognition consists in studying relationships between syntagms and components in a text. These relationships can be syntactic or semantic and most of them concern non-contiguous components [48]. For instance, these kind of models are usefull to introduce the gender and number agreements.

We propose a new model (Statistical Feature Langue 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 [42]. The integration of this new language model in a speech recognition system is under work.

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 is defined by a set of words sharing the same semantic properties. Last year, we proposed to use a naïve 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 [32]. 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 [31][14].

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 [48].

The last research area is concerned with the developement of a new framework for combining language models. This framework is based on a dynamic bayesian network (c.f. § 6.2.3). 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 [28][27].

7. Contracts and Grants with Industry

7.1. National Contracts

7.1.1. PRESSE+ project

Presse+ is the French company which monitor all types of media: written press, radio, television, news agencies, Internet. It collects selects, analyses, organises and transmits information. It wants to automatically detect key information useful for its customers. In this framework, we have started the SIFRE thesis of Emmanuel Didiot. During the first year, a prototype has been realised.

7.1.2. NEOLOGOS project

The NEOLOGOS project results from a collaboration in the speech recognition field between French laboratories (IRISA, ENSSAT, LORIA) and industrial companies (TELISMA, DIALOCA, 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 PAIDIALOGOS. For that database, one thousand of different children will be recorded, using both the GSM and PSTN telephone networks in the following proportions: 65% over PSTN and 35% over GSM. The second is an extensive telephone database of grown up voices, called IDIOLOGOS. We have developped a new method to perform the clustering of the 1000 recorded speakers [38].

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 framwork of the 6th PCRD. The HIWIRE project aims to make 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 provoke the necessary 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).

7.2.2. OZONE

OZONE is an IST project funded by the European Commission, whose main topic is "New technologies and services for emerging nomadic societies". Its reference number is IST-2000-30026, and it is leaded by Philips Research Eindhoven. The other partners involved are: INRIA, Interuniversitaires Micro-Electronica Centrum, Laboratoires d'Electronique Philips, EPICTOID, Eindhoven University of Technology, THOMSON multimedia R&D France.

With several other INRIA teams (including Langue & Dialogue and MAIA in Nancy), we are involved in this project, and our role is to develop a generic multimodal user interface designed for nomadic services. The overall objective of OZONE is to develop a framework for ambiant intelligence that can easily support and adapt to different kinds of devices and situations related to nomadic and pervasive computing. The multimodal user interface will use both speech and gesture inputs and shall be able to model the context information about the user and the nomadic services he can interact with. The INRIA teams are involved in the development of a demonstrator embedded in a cybercar that is based at Rocquencourt.

The work realized this year mainly consists of implementation efforts to integrate the different modules together, including speech recognition, dialogue management, user profile and the WSAMI-based Web Service software environment developed at INRIA Rocquencourt, and to finalise the demonstration platform. We have also been involved in the making-up of a DVD to promote this demonstrator, and in the final review of the project that has been successfully completed in October 2004.

7.2.3. 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).

The work realized this year mainly consists of negociations with the other partners to define and plan our contributions within the consortium.

7.2.4. 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.

7.2.5. MIAMM

The IST Project MIAMM (Multidimensional Information Access using Multiple Modalities, http://www.loria.fr/projets/MIAMM/) n° IST-2000-29487, is developed with CANON (Canon Research Centre Europe Limited, UK), DFKI (Germany), SONY Europe, TNO Human Factors (Netherlands), LED Team (Language and Dialogue, France, LORIA).

The objective of the MIAMM project is to provide an integrated and comprehensive framework for the design of modular multidimensional/multimodal dialogue systems. This dialogue system is a musical database query system. The architecture of the whole prototype is modular: one module for each task and/or partner (French/English/German speech recognition, French/German parsing, dialogue manager...). The MPEG7 standard format, based on XML, is used for communication between modules. This project is now finished. Last year, we conducted a demonstration of the system for the final review. This demonstration included all modalities: haptics, graphics and speech. The French speech modality, which was under our responsability, has been successfully tested in terms of usability (Human Factor experiments conducted by the human factor partner) and performance. For this demonstration, we worked in collaboration with the "Langue et Dialogue" team for developing a TAG grammar (used for generating a parse aimed at building a semantic representation of the sentence) and a statistical language model used for speech recognition. This work is described in [41].

7.2.6. The KDD Cup 2003: an International Challenge on Data Mining

The KDD Cup 2003 is a knowledge discovery and data mining competition held in conjunction with the Ninth Annual ACM SIGKDD Conference http://www.acm.org/sigkdd/kdd2003/. This year competition focused on problems motivated by network mining and the analysis of usage logs. This KDD Cup is based on a very large archive of research papers. It provides a framework for testing general network and usage mining techniques, which will be explored via four varied and interesting tasks. Each task is a separate competition with its own specific goals. Martine Cadot (Orpailleur Team) and Joseph di Martino (Parole Team) participated to the second task which consisted in re-creating the citation graph of about 35000 papers with 1.8 gigs of data: it was required for each paper P in the collection, a list of other papers P_1, ..., P_k in the collection such that P cites P_1, ..., P_k. Note that P might cite papers that are not in the collection. All the papers were Latex articles. For doing this task, a very arduous data cleaning process was implemented using Perl scripts. This work is described in [16].

Martine Cadot and Joseph di Martino worked for this challenge during three months. They took up the third place in this international competition: see http://www.cs.cornell.edu/projects/kddcup/results.html for the official results.

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

This project involves the accurate generation of relevant lip deformations and jaw movements of talking heads because artificial talking heads perform significantly more poorly than true speakers. This issue is particularly crucial to enable lip reading by deaf people and to learn the articulation of phonemes that do not exist in the mother tongue in the case of language learning. The expected contributions of this project are to improve Baldi's (talking head developed at PSL) modelling of labial coarticulation in English and French and to evaluate the benefit of using the talking head designed by Dominic Massaro and Michael Cohen 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 ISA project. Within the context of language learning the work will consist in 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 Etat Région" project

The aim of the project is to design a computer-assisted learning system of English prosody for French students [21]. 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. 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 exercices have been designed by teachers of English as a foreign language. These exercices exploit our speech tools (modifications of prosodic cues, filtering of speech signals) and the facilities of SnorriActive X (see below). 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 native speakers of English will serve as a support for the lessons and references for the correction of student's productions.

Software must be simple enough so that teachers and students can adapt themselves to it easily. For that purpose we ported main editing and signal transformation facilities of our speech analysis software WinSnoori in the form of ActiveX controls that can be easily used from any MS Word, PowerPoint document (as well as online web pages). So users can either record or open a signal, display spectrograms, F0 contours, intensity, phonetic or orthographic annotations (whenever they exist) in any PowerPoint slide. They can also modify prosodic cues such as duration, fundamental frequency (independently or not from intensity) at each instant of the signal.

In order to develop a method for the automatic alignement 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.

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 onto the face of one subject. This system has been developed by the ISA project 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 the piloting of 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 Levy (ISA project).

8.2. National Actions

8.2.1. Feedart INRIA cooperative research action (TSI-ENST - ISA and Parole teams

The long term ambition of the cooperative research action Feedart is to offer articulatory feedback to deaf people acquiring language or people learning a foreign language. This project necessitates, on the one hand, recovering articulatory parameters from the speech signal that can be, or not, supplemented by images of speaker's face, and, on the other hand, generating a talking face that produces vocal tract and face deformations consistent with those that could produce a true speaker. The first aspect corresponds to the acoustic-to-articulatory inversion, the second to the synthesis of a talking head.

Modeling coarticulation phenomena, i.e. the way consecutive phonemes influence with each other from an articulatory point of view, is crucial because it affects the audio-visual integration and perception of the "speech plus face" delivered by the talking head.

It turns out that present modeling of labial coarticulation is not sufficient and does not enable correct lip reading by deaf people. We are therefore working with a view of designing better labial coarticulation methods. This works requires recovering of the 3D geometry of speaker's face and especially of lips that should be known with a good precision and the tracking of 3D markers put onto the face. These data are then exploited to derive and evaluate labial coarticulation models.

This year, we focused on the recovering of the 3D geometry of speakers and lips, and the importation of these data in 3D modeling software as Poser, for instance. Furthermore, Shinji Maeda and Jacques Feldmar worked on the adaptation of face deformation modes obtained for one speaker to any arbitrary head given by its 3D mesh.

Tracking 3D markers on the face or even some characteristic points (giving lip aperture, lip protrusion and jaw position) without requiring any marking can be used to reduce the under-determination of the acoustic-to-articulatory inversion. Indeed, the description of the vocal tract shape by articulatory parameters, those of Maeda for instance, requires more parameters - seven - than the number of acoustic parameters that can be recovered from the speech signal - usually the first three formants. Therefore, it can be useful to supplement acoustic parameters with visible parameters.

8.2.2. MathSTIC Project

Probabilistic graphical models for automatic speech recognition

Partners: ENST Paris, ENS-Cachan, Institut Elie-Cartan and LORIA.

This project brings together mathematicians and speech/signal processing specialists to deeply investigate the formalism of dynamic Bayesian networks and robustness problems in speech recognition. The goal is to develop new speech probabilistic models which can lead to robust speech recognition systems.

8.2.3. RAIVES-STIC-SHS Project

The "Invisible Web" is composed of documents which can not currently be accessed by Web search engines. This is due to several reasons, among them we can point out: dynamic URL and no textual format as video and audio documents.

For audio documents, one solution to access the documents is automatic indexing. It consists in finding good descriptors in audio documents which can be used as indexes for archiving and search. Last year we started a French research project on audio indexing, RAIVES (Recherche Automatique d'Informations Verbales Et Sonores). Audio indexing systems can be based on a complete transcription but it is not the only meaning-full information which can be extracted from an audio document. Non-verbal information (as music, jingles or speakers) is also informative for an audio document, and can lead to the extraction of pertinent descriptors. We focus on this kind of information extraction. Therefore, the aim of this project is to automatically separate speech segments from music segments, detect key sounds (like jingles), identify the language of a segment, split the audio signal according to the speakers, detect some keywords and extract the main topics. For the two first years of this project, we participated to the specification of the corpus : 180 hours of the French public radio station RFI (Radio France International). Programs are broadcast news as well as interviews and musical programs. We have developed software to analyse recognition results.

8.2.4. ESTER Project

As, in USA, NIST organises 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 (Evaluation 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 developped 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 [29], [22]. A real-time version of this system was presented in [30].

9. Dissemination

9.1. Animation of the scientific community

The members of Parole are involved in several commitee 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 Etat Region” and member of Eurospeech scientist commitee.
- J.P. Haton is a member of CSL and ICSLP programm commitee, chairman of French Science and Technology Association
- Y. Laprie is a member of (LREC, JEP) scientific commitee. He is an elected member of G.F.C.P, "groupe francophone de la communication" and head of the “Assistant intelligent” project of the PRST “Intelligence Logicielle”.
- O. Mella, D. Fohr, I. Illina, C Cerisara, D. Langlois are involved in several european and national projects.
- K. Smaïli is a member of (Eurospeech, JEP) scientific commitee.

9.2. Distinctions

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

9.3. Invited lectures

- Sylvain Meignier, LIMSI, Paris,
- Julien Pinquier, IRIT, Toulouse,
- Dominic Massaro, Santa Cruz,
- Jean-Baptiste Maj, Leuwen.

9.4. Invited Professors

- Dwayne Paschall, University of Texas,
- Jean Shoentgen, Bruxelles.

9.5. Higher education

- A strong involvement of the team members in education and administration (UHP, univesité Nancy 2, INPL): Master, computer science DEA, IUT, MIAGE, DESS;
- Head of computer science departement STMIA (M. C. Haton),
- Head of MIAGe departement (K. Smaïli),
- Head of Network Speciality of UHP Computer Science DESS (O. Mella).

9.6. Participation to workshops and PhD thesis commitees:

- Members of Phd thesis commitee 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 (see next section).

10. Bibliography

Major publications by the team in recent years

[1] F. BIMBOT, M. EL-BÈZE, S. IGOUNET, M. JARDINO, K. SMAILI, 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° 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° 2, April 2001, p. 151-174.

[4] 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.

[5] 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.

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

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

[8] 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.

[9] Y. LAPRIE, M.-O. BERGER. *Cooperation of Regularization and Speech Heuristics to Control Automatic Formant Tracking*, in "Speech Communication", vol. 19, n° 4, October 1996, 23.

[10] I. ZITOUNI, K. SMAILI, 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.

Doctoral dissertations and Habilitation theses

[11] V. BARREAUD. *Reconnaissance automatique de la parole continue : compensation des bruits par transformation de la parole*, Thèse d'université, Henry Poincaré - Nancy 1, November 2004.

[12] Y. BENAYED. *Détection de mots clés dans un flux de parole*, Ecole, Ecole Nationale Supérieure des Télecommunications (Paris), December 2003.

[13] 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, October 2004.

[14] S. JAMOUSSI. *Méthodes statistiques pour la compréhension automatique de la parole*, Thèse d'université, Henri Poincaré, December 2004.

[15] F. LAURI. *Adaptation au locuteur de modèles acoustiques markoviens pour la reconnaissance automatique de la parole*, Thèse d'université, Nancy 2, October 2004.

Articles in referred journals and book chapters

[16] M. CADOT, J. DI MARTINO. *A Data Cleaning Solution by Perl Scripts for the KDD Cup 2003 Task 2*, in "SIGKDD Explorations", vol. 5, n° 2, January 2004, p. 154–155.

[17] 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° 1, January 2004, p. 25–41.

Publications in Conferences and Workshops

[18] V. BARREAUD, I. ILLINA, D. FOHR, V. COLOTTE. *Compensation en milieu variant abruptement*, in "Journées d'Etudes sur la Parole - JEP'04, Fès, Maroc", April 2004.

[19] Y. BENAYED, D. FOHR, J.-P. HATON, G. CHOLLET. *Comparaison de différentes méthodes de classification pour la détection de mots clés en parole continue*, in "7ème Colloque Africain sur la Recherche en Informatique - CARI'04, Hammamet, Tunisie", November 2004.

[20] Y. BENAYED, D. FOHR, J.-P. HATON, G. CHOLLET. *Using confidence measure for keyword detection in continuous speech recognition*, in "Conférence Internationale sur l'accès Intelligent aux Documents Multimédia sur l'Internet - Medinet'04", November 2004.

[21] A. BONNEAU, M. CAMUS, Y. LAPRIE, V. COLOTTE. *A computer-assisted learning of English prosody for French students*, in "Integrating Speech in Learning (InSTIL 2004), Venise, Italie", June 2004.

[22] A. BRUN, C. CERISARA, D. FOHR, I. ILLINA, D. LANGLOIS, O. MELLA, K. SMAÏLI. *ANTS: le système de transcription automatique du Loria*, in "Journées d'Etude sur la Parole - JEP'04, Fès, Maroc", April 2004, http://www.loria.fr/publications/2004/A04-R-037/A04-R-037.ps.

[23] A. BRUN, K. SMAÏLI. *Fiabilité de la référence humaine dans la détection de thème*, in "Traitement Automatique des Langues Naturelles - TALN'2004, Fès, Maroc", April 2004.

[24] C. CERISARA, D. FOHR, O. MELLA, I. ILLINA. *Exploiting models intrinsic robustness for noisy speech recognition*, in "8th International Conference on Spoken Language Processing - ICSLP'2004, Jeju, South Corea", October 2004, http://www.loria.fr/publications/2004/A04-R-270/A04-R-270.ps.

[25] V. COLOTTE, R. BEAUFORT. *Synthèse vocale par sélection linguistiquement orientée d'unités non-uniformes : LiONS*, in "Journée d'Etudes de la Parole - JEP'04, Fès, Maroc", April 2004.

[26] K. DAOUDI, M. DEVIREN. *Une nouvelle architecture de compensation du bruit pour la reconnaissance robuste de la parole*, in "XXVes Journées d'Etudes sur la Parole - JEP-TALN-RECITAL 2004, Fès, Maroc", April 2004, http://www.loria.fr/publications/2004/A04-R-100/A04-R-100.ps.

[27] M. DEVIREN, K. DAOUDI, K. SMAILI. *Language modeling using dynamic Bayesian networks*, in "4th International Conference on Language Resources and Evaluation - LREC 2004, Lisbonne, Portugal", May 2004, http://www.loria.fr/publications/2004/A04-R-099/A04-R-099.ps.

[28] M. DEVIREN, K. DAOUDI, K. SMAILI. *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", April 2004, http://www.loria.fr/publications/2004/A04-R-098/A04-R-098.ps.

[29] D. FOHR, O. MELLA, I. ILLINA, C. CERISARA. *Experiments on the accuracy of phone models and liaison processing in a French broadcast news transcription system*, in "8th International Conference on Spoken Language Processing - ICSLP'2004, Jeju, South Corea", October 2004.

[30] 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.

[31] 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", May 2004.

[32] S. JAMOUSSI, K. SMAÏLI, D. FOHR, J.-P. HATON. *Un système de compréhension automatique de la parole pour l'interrogation orale d'une base de données de bourse*, in "Journées d'Etudes sur la Parole - JEP'04, Fès, Maroc", April 2004.

[33] F. KORKMAZSKY, M. DEVIREN, D. FOHR, I. ILLINA. *Hidden Factor Dynamic Bayesian Networks for Speech Recognition*, in "8th International Conference on Spoken Language Processing - ICSLP'2004, Jeju, South Corea", October 2004.

[34] F. KORKMAZSKY, D. FOHR, I. ILLINA. *Using Linear Interpolation to Improve Histogram Equalization for Speech Recognition*, in "8th International Conference on Spoken Language Processing - ICSLP'2004, Jeju, South Corea", October 2004.

[35] P. KRÁL, J. KLEČKOVÁ, C. CERISARA. *Analysis of Importance of the prosodic Features for Automatic*

Sentence Modality Recognition in French in real Conditions, in "WSEAS ICECS, Crete, Greece", vol. 3, n° 9, November 2004, p. 1820–1824.

[36] Y. LAPRIE, S. JARIFI, A. BONNEAU, D. FOHR. *Détection automatique de sons bien réalisés*, in "Actes des XXVes Journées d'Étude sur la Parole - JEP'2004, Fès, Maroc", April 2004, http://www.loria.fr/publications/2004/A04-R-284/A04-R-284.ps.

[37] Y. LAPRIE. *A concurrent curve strategy for formant tracking*, in "Interspeech 2004 - International Conference on Spoken Language Processing, Jeju, Corée du sud", October 2004, http://www.loria.fr/publications/2004/A04-R-337/A04-R-337.ps.

[38] E. PINTO, D. CHARLET, H. FRANÇOIS, D. MOSTEFA, O. BOFFARD, D. FOHR, O. MELLA, F. BIMBOT, K. CHOUKRI, Y. PHILIP, F. CHARPENTIER. *Development of new telephone speech databases for French : the NEOLOGOS Project*, in "International Conference on Language Resources and Evaluation - LREC'04, Lisbonne, Portugal", May 2004.

[39] B. POTARD, Y. LAPRIE, S. OUNI. *Expériences d'inversion basées sur un modèle articulatoire*, in "Journées d'Etudes sur la Parole - JEP'04, Fès, Maroc", April 2004, http://www.loria.fr/publications/2004/A04-R-335/A04-R-335.ps.

[40] J. RAZIK, D. FOHR, O. MELLA, N. PARLANGEAU-VALLÈS. *Segmentation Parole/Musique pour la transcription automatique*, in "Journées d'Etude sur la Parole - JEP 2004, F?s, Maroc", April 2004, http://www.loria.fr/publications/2004/A04-R-036/A04-R-036.ps.

[41] 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", May 2004.

[42] 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.

Bibliography in notes

[43] 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.

[44] A. BRUN, K. SMAILI, J.-P. HATON. *Nouvelle approche de la sélection de vocabulaire pour la détection de thème*, in "Traitement Automatique du Langage Naturel - TALN'2003, Batz-sur-Mer, France", Jun 2003, http://www.loria.fr/publications/2003/A03-R-481/A03-R-481.ps.

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

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

[47] S. JAMOUSSI, K. SMAILI, J.-P. HATON. *Understanding process for speech recognition*, in "Eighth European Conference on Speech Communication and Technology - EuroSpeech'03, Gene've, Suisse", Sep 2003.

[48] 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.

[49] Y. LAPRIE, M.-O. BERGER. *Cooperation of Regularization and Speech Heuristics to Control Automatic Formant Tracking*, in "Speech Communication", vol. 19, n° 4, October 1996, p. 255–270.

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

[51] F. PLANTE, G. MEYER, W. AINSWORTH. *Improvement of speech spectrogram accuracy by the method of reassignment*, in "IEEE Transactions on Speech and Audio Processing", vol. 6, n° 3, 1998, p. 282-287.