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 joint project to INRIA, CNRS, Henri Poincaré University and Nancy 2 University through LORIA laboratory (UMR 7503). For more details, we invite the reader to consult the team web site at http://parole.loria.fr/.

Head of project-team

Yves Laprie [Research scientist, CNRS, HdR]

Administrative Assistant

Martine Kuhlmann [Assistant, CNRS]

CNRS Research scientist

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

Faculty member

Armelle Brun [Assistant Professor, Nancy 2 University]
Martine Cadot [PRAG, Henri Poincaré University since 2005, July 7th]
Vincent Colotte [Assistant Professor, Henri Poincaré University]
Joseph di Martino [Assistant Professor, Henri Poincaré University]
Jean-Paul Haton [Professor, Henri Poincaré University, Institut Universitaire de France, HdR]
Marie-Christine Haton [Professor, Henri Poincaré University, HdR]
Irina Illina [Assistant Professor, I.U.T Charlemagne, Nancy 2 University working for INRIA until 2006, September, HdR]
David Langlois [Assistant Professor, IUFM (University Institute for Teacher Training)]
Odile Mella [Assistant Professor, Henri Poincaré University]
Slim Ouni [Assistant Professor, I.U.T Charlemagne, Nancy 2 University]
Kamel Smaïli [Professor, Nancy 2 University, HdR]

Phd Students

Ghazi Bouselmi [TA, thesis to be defended in 2007]
Matthieu Camus [INRIA Grant, thesis to be defended in 2008]
Sébastien Demange [TA, thesis to be defended in 2007]
Emmanuel Didiot [CIFRE grant, ATER since november 2006, thesis to be defended in 2007]
Pavel Král [Czech coPhD, thesis to be defended in 2007]
Blaise Potard [MENRT grant, thesis to be defended in 2007]
Joseph Razik [INRIA grant, ATER since october 2006, thesis to be defended in 2007]
Vincent Robert [High school teacher, thesis to be defended in 2007]
Caroline Lavecchia [EADS foundation grant, thesis to be defended in 2008]
Farid Feïz [CNRS grant (ASPI contract), thesis to be defended in 2008]

Project technical staff

Christophe Antoine [INRIA, until 2006, May 14th]
Alexandre Lafosse [CNRS until 2006, August 31st, INRIA (associate engineer) since 2006, September 1st]
Julien Maire [CNRS]
Viet-Bac Le [CNRS, since 2006, July 19th]

Post Doctoral fellow

Vincent Barreaud [ATER, ESIAL, until 2006, September]

Specialist engineer

Jacques Feldmar [Specialist engineer]

2. Overall Objectives

2.1. Overall Objectives

PAROLE is a common project to INRIA, CNRS, Henri Poincaré University and Nancy 2 University through the LORIA laboratory (UMR 7503). The purpose of our project is to automatically process speech signals to understand their meaning, and to analyze and enhance their acoustic structure. It inscribes within the view of offering efficient vocal interfaces and necessitates works in analysis, perception and automatic speech recognition (ASR).

Our activities are structured in two topics:

- **Speech analysis.** Our works are concerned with automatic extraction and perception of acoustic cues, acoustic-to-articulatory inversion and speech analysis. These themes give rise to a number of ongoing and future applications: vocal rehabilitation, improvement of hearing aids and language learning.
- **Modeling speech for automatic recognition.** Our works are concerned with stochastic models (HMM[1], bayesian networks and missing data models), multiband approach, adaptation of a recognition system to a new speaker or to the communication channel, and with language models. These topics give also rise to a number of ongoing and future applications: automatic speech recognition, automatic speech to speech translation, text-to-speech alignment and audio indexing.

Our pluridisciplinary scientific culture combines works in phonetics, pattern recognition and artificial intelligence. This pluridisciplinarity turns out to be a decisive asset to address new research topics, particularly language learning and multiband approaches that simultaneously require competences in automatic speech recognition and phonetics.

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

3. Scientific Foundations

3.1. Introduction

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

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

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

[1] Hidden Markov Models

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

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

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

3.2. Speech Analysis

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

3.2.1. Acoustic cues

3.2.1.1. Automatic detection of strong cues

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

3.2.1.2. 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 confidence are probably well realized and clearly pronounced. We made the hypothesis that the enhancement of well realized sounds in a sentence gives listeners some islands of confidence during the acoustic decoding stage and improves speech intelligibility.

Previous studies have shown that such an enhancement as well as the slowing down of some classes of sounds (fricatives and stops, in particular) improve the perception of a second language as well as that of the first language for hearing impaired people.

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

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

3.2.2. Oral comprehension

We developed tools to improve speech perception and production, and made perceptual experiments to prove their efficiency in language learning. These tools are also of interest for hearing impaired people, as well as for normally hearing people in noisy environments.

3.2.2.1. Speech signal transformation

In order to improve oral comprehension, we use PSOLA (Pitch Synchronous Overlap and Add), a speech signal transformation method. PSOLA is based on the decomposition of the speech signal into overlapping pitch synchronous frames. Signal modifications consist in manipulating analysis marks to generate new synthesis marks. PSOLA is well known for its easy implementation and the quality of the slowed down signals. However, temporal discrepancies may appear in the region of the synthesis marks and PSOLA may generate noise between harmonics. In order to reduce the loss of quality, the method was improved in the two following ways. First, we have proposed a pruning algorithm to seek analysis marks (for pitch synchronization). It increased the robustness of pitch marking for speech segments with strong formant variation. Second, we improved the localization of analysis and synthesis marks. During the analysis stage, we can either oversample the signal or use F0 detection algorithm which gives an accuracy better than one sample. During the synthesis stage, the improvement is based on a dynamical re-sampling of the speech signal so as to accurately replace the frame on synthesis marks. Both improvements strongly reduced the level of noise between harmonics and we obtained a speech signal of high quality [45].

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

Within the framework of a project concerning language learning, more precisely the acquisition of the prosody of a second language, we are starting a study on the automatic detection and correction of prosodic deviations. This work implies three different tasks: (a) the detection of the prosodic entities of the learner’s realization (lexical accent, intonative patterns), (b) the evaluation of the deviations, by comparison with a model, and (c) their corrections, both verbal and acoustic. This last kind of feedback is directly done on the learner’s realization: the deviant prosodic cues are replaced by the prosodic cues of the model. The identification and correction tasks use speech analysis and modification tools developed in our team. We started our project with the automatic detection of the lexical accent of "transparent" words. For more complex identification tasks, we plan to implement a prosodic model.

3.2.3. Acoustic-to-articulatory inversion

Acoustic-to-articulatory inversion aims at recovering the articulatory dynamics from speech signal that may be supplemented by images of the speaker face. Potential applications concern low bit rate speech coding, automatic speech recognition, speech production disorders assessment, articulatory investigations of phonetics, talking heads and articulatory feedback for language acquisition or learning.

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

Solving acoustic equations. In order to solve the acoustic equations adapted to the vocal tract, one assumes that the sound wave is a plane wave in the vocal tract and that it can be unbend. There are two families of solving methods: **(i)** frequency methods through the acoustical-electrical analogy, **(ii)** spatio-temporal methods, through the direct solving of finite difference equations derived from Webster equations.

Measuring the vocal tract. This represents an important obstacle because there does not exist any reliable method enabling a precise measurement in time and dimension. MRI (Magnetic Resonance Imaging) enables 3D measurements but is not sufficiently fast and X-rays only allows a sagittal slice of the vocal tract to be captured while involving not acceptable health hazards.

Articulatory modeling. Articulatory models aim at describing all the possible vocal tract shapes with a small number of parameters, while preserving deformations observed on a real vocal tract. Present articulatory models often derive from data analysis of cineradiography moving pictures. One of the most widely used is the one built by Maeda [56].

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

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

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

3.2.4. Strategies of labial coarticulation

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

In the long term, our goal is to determine a method of prediction of labial coarticulation adaptable to a virtual speaker. Predicting labial coarticulation is a difficult problem that gave rise to many studies and models. To predict the anticipatory coarticulation gestures (see [48] for an overall presentation of labial coarticulation), three main models have been proposed: the look-ahead model, the time-locked model and the hybrid model.

These models were often compared on their performance in the case of the prediction of anticipation protrusion in VCV or VCCV sequences where the first vowel is unrounded, the consonant(s) is neutral with respect to labial articulation and the last vowel is rounded. There is no general agreement about the efficiency of these models. More recent models have been developed. The one of Abry and Lallouache [38] advocates for the theory of expansion movements: the movement tends to be anticipated when no phonological constraint is imposed on labiality. Cohen and Massaro [43] proposed dominance functions that require a substantial numerical training.

Most of these models derive from the observations of a limited number of speakers. We are developing a more explicative model, i.e., essentially a phonetically based approach that tries to understand how speakers manage to control labial parameters from the sequence of phonemes to be articulated.

3.3. Automatic speech recognition

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

3.3.1. Acoustic features and models

3.3.1.1. Acoustic features

The raw acoustic signal needs to be parameterized to extract the speech information it contains and to reduce its dimensionality. Most of our research and recognition technologies make use of the classical Mel Feature Cepstral Coefficients, which have proven since many years to be amongst the most efficient front-end for speech recognition. However, we have also recently used and explored alternative parameterizations to support some of our recent research progresses. For example, one requirement of missing data recognition is to build masks in a frequency-like domain. Furthermore, depending on the marginalization technique, different properties of the time-frequency feature domain are required. Hence, we have developed two additional feature domains: the first one is the simple Mel-scale filterbank energies, and the second one, called "Frequency filtered coefficients", decorrelates the frequency coefficients to justify the use of diagonal covariance marginalization approaches. Both these feature domains are exploited in the context of missing data recognition. We have further developed a new robust front-end, which is based on wavelet-decomposition of the speech signal. This front-end generalizes the Frequency filtered coefficients. Finally, we also largely exploit the standard ETSI advanced front-end [47], which is famous for its robustness to noise.

3.3.1.2. Acoustic models

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

3.3.1.3. Robustness and invariance

The core of our research activities about ASR aims at improving the robustness of recognizers to the different kinds of variabilities that affect the speech signal and damage the recognition. Indeed, the issue of the lack of robustness of state-of-the-art ASR systems is certainly the most problematic one that still prevents the wide deployment of speech recognizers nowadays. In the past, we have developed a large range of techniques to address this difficult topic, including robust acoustic models (such as stochastic trajectory and multi-band models) and model adaptation techniques (improvements of Parallel Model Combination, such as Jacobian adaptation). These technologies constitute the foundations of our recent developments in this area, such as non-native speaker adaptation, missing data recognition and denoising. The following state-of-the-art approaches thus form our baseline set of technologies:

- MLLR (Maximum Likelihood Linear Regression) Maximum Likelihood Linear Regression adapts the acoustic models to noisy conditions or to a new speaker in the cepstral domain. The method estimates the linear regression parameters associated with Gaussian distributions of the models. The Maximum Likelihood criterion is used for the estimation of the regression parameters.
- MAP and MAPLR (Maximum A Posteriori - Linear Regression) This adaptation is based on Maximum A Posteriori training of HMM parameters, which uses some data from the target condition. This approach uses both the adaptation data and the prior information. The flexibility in incorporating the prior information makes MAP efficient for handling the sparse training data problem.
- PMC (Parallel Model Combination) is an algorithm to adapt the clean speech models to a noisy environment. It basically converts the models back to the power-spectral domain where speech and noise are assumed to be additive. Unlike the two previous methods, it does not require a large amount of adaptation data - about one second speech signal is enough to estimate the noise model.
- CMN (Cepstral Mean Normalization) is an algorithm to compensate for channel mismatch (differences in microphones for example). It is quite effective and very simple to implement, which explains why it is now used in nearly every recognition system.
- Spectral Subtraction subtracts a noise estimated from the incoming signal in the power spectral domain. This "denoising" algorithm is not extremely efficient when used as a pre-processor to a recognition engine.
- Jacobian Adaptation is a linear version of PMC that operates only in the features domain. It is one of the fastest model adaptation algorithms. The original models do not need to be trained in a clean environment. The method works actually better when the models are already slightly noisy.

3.3.1.4. Segmentation

Audio indexing and automatic broadcast news transcription need the segmentation of the audio signal. The segmentation task consists in two steps: firstly, homogeneous segments are extracted and classified into speech, noise or music, secondly, speakers turns are detected in the extracted speech segments.

Speech/music segmentation is often based on the acoustic differences between both kinds of sounds. So discriminative acoustic cues are investigated (FFT, zero crossing rate, spectral centroid, wavelets ...). Except the selection of acoustic features, another point is to find the best classifier. Various classifiers are commonly used: k-Nearest-Neighbors, Hidden Markov Models, Gaussian Mixture Models, Artificial Neural Networks.

As to detect speaker turns, the main approach consists of splitting the audio signal into smaller segments that are assumed to contain only one speaker and then a hierarchical clustering scheme is performed for merging segments belonging to the same speaker.

3.3.2. Language modeling

Apart from the challenges related to acoustic modeling that have just been discussed, some problems due to the complexity of natural language remain without any satisfactory solution. State-of-the-art language models, as n-gram models, are very simple and reach relatively high performance in ASR. Such characteristics explain their predominance in recognition systems. n-gram models assign a probability to the current word, using the only $n-1$ preceding words. For example, given the beginning of sentence "*les pommes que j'ai*", a 4-gram model considers only the 3 previous words "*que j'ai*" to compute the probability of the current word. Due to long distance dependencies, such a model cannot be efficient. However, it is impossible to systematically increase the value of n due to computational constraints and probability reliance.

Our group, as other groups through the world, make an increased effort in order to design more efficient language models. Most of language models we propose are based on information theory and statistics, some of them include linguistic knowledge to improve statistical knowledge. Combining several language models increases performance but unfortunately this is not sufficient and that is why we are still developing new methods to deal with the complexity of language modeling. One way to reach this objective is to deal with a larger history and to consider words as complex unit. Henceforth, a word is redefined as a compound entity represented by a list of features including the orthographic form of the word. To do so, we work in several directions:

- **Feature vector models**. A word is considered not only as an orthographic form but as a complex unit which contains a class tag, a gender and number features, a semantic tag... This makes a statistical language model more realistic and in harmony with linguistic theory. Some tracks were explored and confirmed the feasibility of this approach [61], [46].
- **Phrase-based models** Another way to deal with complex linguistic units consists in using phrase-based models [53], [62]. The idea is to include phrases in the dictionary. These phrases bring up more information for the decoding process. The phrases are retrieved automatically from large corpora
- **Language model adaptation using topic identification**. The objective is first to find out the topic of the uttered sentences, and second, to adapt the baseline language model using the one that corresponds to the retrieved topic. Research concerns both identification and adaptation [42], [37].
- **Selecting the best language model in accordance to the history**. Combining language models is not sufficient to deal with the complexity of natural language, the best way to improve the performance of a speech recognition or translation system is to select dynamically the best language model depending on a history as in the SHP principle [54]. We are pursuing in this research direction in order to obtain an efficient selection of language models.
- **Bayesian Networks**. Bayesian network is a powerful formalism which modelizes the relationship between several events. We exploit this concept in order to construct a more consistent combination of language knowledge. We have developed an unifying approach that processes each knowledge in a unique model and constructs new data-driven language models with improved performances. The principle of this approach is to construct Dynamic Bayesian Networks (DBNs) in which a variable (word, class or any other linguistic unit) may depend on a set of context variables. The details and evaluation of this approach using several datasets are reported in [46].

Our research has been applied to large vocabulary dictation machine, news transcription [50], automatic categorization of mails, dialog systems [60], vocal services [51]...Recently, we initiated another promising research direction for the next ten years: speech-to-speech translation. In fact, this activity concerns not only speech recognition problems, but also machine translation, language model adaptation, speech understanding and decoding problems.

4. Application Domains

4.1. Application Domains

Our research is applied in a variety of fields from ASR to paramedical domains. Speech analysis methods will contribute to the development of new technologies for language learning (for hearing-impaired persons and for the teaching of foreign languages) as well as for hearing aids. In the past, we developed a set of teaching tools based on speech analysis and recognition algorithms of the group (cf. the ISAEUS [49] project of the EU that ended in 2000). We are continuing this effort towards the diffusion of a course on Internet.

Speech is likely to play an increasing role in man-machine communication. Actually, speech is a natural mean of communication, particularly for non-specialist persons. In a multimodal environment, the association of speech and designation gestures on touch screens can, for instance, simplify the interpretation of spatial reference expressions. Besides, the use of speech is mandatory in many situations where a keyboard is not available: mobile and on-board applications (cf. the HIWIRE European project for the use of speech recognition in a cockpit plane), interactive vocal servers, telephone and domestic applications, etc. Most of these applications will necessitate to integrate the type of speech understanding process that our group is presently studying. Furthremore, speech to speech translation concerns all multilingual applications (vocal services, audio indexing of international documents). The automatic indexing of audio and video documents is a very active field that will have an increasing importance in our group in the forthcoming years, with applications such as economic intelligence, automatic transcription, and keyword spotting.

5. Software

5.1. Software

5.1.1. Snorri and WinSnoori

Snorri is a speech analysis software that we have been developing for 15 years. It is intended to facilitate the work of the scientist in automatic speech recognition, phonetics or speech signal processing. Basic functions of Snorri enable several types of spectrograms to be calculated and the fine edition of speech signals (cut, paste, and a number of filters) as the spectrogram allows the acoustical consequences of all the modifications to be evaluated. Beside this set of basic functions, there are various functionalities to annotate phonetically or orthographically speech files, to extract fundamental frequency, to pilot the Klatt synthesizer and to utilize PSOLA resynthesis.

This year we supplemented tools to display spectral analyses used in automatic speech recognition. We developed a graphical interface which enables the impact of PLP (Perceptually Linear Prediction) parameters to be investigated. This interface is incorporated with other tools dedicated to the computation and the display of Mel cepstrally smoothed spectra.

The main improvement concerns automatic formant tracking which is now available with other tools for copy synthesis. It is now possible to determine parameters for the formant synthesizer of Klatt quite automatically. The first step is formant tracking, then the determination of F0 parameters and finally the adjustment of formant amplitudes for the parallel branch of the Klatt synthesizer. The automatic formant tracking that has been implemented is an improved version of the concurrent curve formant tracking [55]. One key point of this tracking algorithm is the construction of initial rough estimates of formant trajectories. The previous algorithm used a mobile average applied onto LPC roots. The window is sufficiently large (200 ms) to remove fast varying variations due to the detection of spurious roots. The counterpart of this long duration is that the mobile average prevents formants fairly far from the mobile average to be kept. This is particularly sensitive in the case of F2 which presents low frequency values for back vowels. A simple algorithm to detect back vowels from the overall spectral shape and particularly energy levels has been added in order to keep extreme values of F2 which are relevant.

Together with other improvements reported during the last four years, this constitutes the new version of WinSnoori (1.34) which has been released on the WinSnoori website.

5.1.2. 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.

5.1.3. Labelling corpora

We developed a labelling tool which allows syntactic ambiguities to be solved. The syntactic class of each word is assigned depending on its effective context. This tool is based on a large dictionary (230000 lemmas) extracted from BDLEX and a set of 230 classes determined by hand. This tool has a labelling error of about 1 %.

5.1.4. Automatic lexical clustering

In order to adapt language models in ASR applications, we have developed a new toolkit to automatically create word classes. This toolkit exploits the simulated annealing algorithm. Creating these classes requires a vocabulary (set of words) and a training corpus. The resulting set of classes minimizes the perplexity of the corresponding language model. Several options are available: the user can fix the resulting number of classes, the initial classification, the value of the final perplexity, etc.

5.1.5. Unsupervised Tagging Tool

A tagger is currently under development. Such a tool is dedicated to tag a text (typically with Parts of Speech). A tagger needs a time-consuming manual pre-tagging to bootstrap the training parameters. It is then difficult to test numerous tag sets as needed for our research activities. However, this stage could be skipped [52].

5.1.6. SALT (Semi-Automatic Labelling Tool)

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

5.1.7. LIPS (Logiciel Interactif de Post-Synchronisation)

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

5.1.8. ESPERE

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

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

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

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

5.1.9. ANTS

The aim of the Automatic News Transcription System (ANTS) is to transcribe radio broadcast news. ANTS is composed of four stages: broad-band/narrow-band speech segmentation, speech/music classification, detection of silences and breathing segments and large vocabulary speech recognition. The three first stages split the audio stream into homogeneous segments with a manageable size and allow the use of specific algorithms or models according to the nature of the segment.

Speech recognition is based on the Julius engine and operates in two passes: in the first pass, a frame-synchronous beam search algorithm is applied on a tree-structured lexicon assigned with bigram language model probabilities. The output of this pass is a word-lattice. In the second pass, a stack decoding algorithm using a trigram language model gives the N-best recognition sentences.

5.1.10. BNTK

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

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

5.1.11. HTK-compliant recognition tools

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

- The HMM parallel training tool distributes the training process over the 25 computers of the PAROLE PC cluster.
- The HMMModelConv toolkit is an extensible software that converts acoustic models between different formats: HTK, ESPERE and Sphinx3.
- The stochastic speech library (old GMMlib) is a JAVA library that integrates most of our recent work on stochastic speech processing. It is highly modular, and supports JUnit testing for most of its functionalities, as well as non-regression tests for speech recognition on standard databases (Aurora4 for now). Its most visible functionalities include:
  - Support load and save formats for HTK models, HTK parameter files and HTK label files.
  - Support visualization and tagging of speech spectrograms, parameter files, missing data masks and label files.
  - It is based on the stochastic processing "pull" paradigm, which allows to easily plug-in and chain several processing modules, while ensuring stream synchronization in the case of lattice-like modules chains.
  - Support advanced Gaussian Mixture Models training and editing (LBG, cross-correlation, marginalization, etc.).
  - Include beta modules for Bayesian denoising of speech signals.
  - Include beta modules for accuracy-based missing data recognition masks inference and training (see section 6.2.1).

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

5.1.12. STARAP

STARAP (Sous-Titrage Aidé par la Reconnaissance Automatique de la Parole) is a toolkit to help the making of sub-titles for TV shows. This toolkit performs:

- Parameterization of speech data;
- Clustering of parameterized data;
- Gaussian Mixture Models (GMM) training;
- Viterbi recognition.

The formats of the input and output files are compatible with HTK toolkit. This toolkit was realised in the framework of the STORECO contract (see section 7.1.2).

6. New Results

6.1. Speech Analysis

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

Participants: Anne Bonneau, Vincent Colotte, Dominique Fohr, Jean-Paul Haton, Yves Laprie, Joseph di Martino, Slim Ouni, Blaise Potard, Matthieu Camus.

6.1.1. Acoustic-to-articulatory inversion

The strength of our inversion method lies on the quasi-uniform acoustic resolution of the articulatory table. The originality is based on the generation method that evaluates the linearity of the articulatory-to-acoustic mapping at each step. Articulatory parameters of Maeda's model vary between -3 σ and 3 σ where σ is the standard deviation. Thus, the codebook inscribes a root hypercube. Sampling the articulatory space amounts to finding reference points that limit linear regions. The inversion procedure then retrieves articulatory vectors corresponding to acoustic entries from the hypercube codebook. A non-linear smoothing algorithm together with a regularization technique is then used to recover the best articulatory trajectory. The inversion ensures that retrieved articulatory parameters produce original formant trajectories accurately and a realistic sequence of the vocal tract shapes [57].

This year, we investigated the impact of phonetic constraints proposed in a previous work [59] on inversion. Since there are few articulatory data available with the corresponding acoustical signal we used X-ray data recorded by IPS (Institut de Phonétique de Strasbourg). The quality of the acoustical signal is not very good but sufficient to determine formant frequencies. The evaluation was carried out on five vowels of the speaker PB whose X-ray contours can be found in [40]. We have analyzed the vocal tract shapes recovered for five vowels (/a,i,y,u,e/) uttered by a female speaker. It turns out that the phonetic constraints derived from standard phonetic knowledge are quite efficient to keep relevant vocal tract shapes and do not penalize realistic vocal tract shapes. The key point is that standard phonetic knowledge enables interdependencies between articulators to be captured efficiently. In addition, this work[32] enables the evaluation of the articulatory model itself. Indeed, the number of vocal tract shapes recovered strongly depends on the vowel. The inversion procedure, and especially the exploration of the null space of the articulatory to acoustic mapping, roughly samples the articulatory space in a uniform fashion. This means that the number of solutions is tightly connected to the extent of the articulatory region corresponding to vowels. Our results show that /u/ requires a more precise articulation than /a/. Other vowels /e,i,y/ are between these two extremes.

We also continued the work on the use of constraints provided by the view of visible articulators (lips and jaw). It is not possible to directly use 3D data into the inversion process. Indeed, the inversion process relies on the Maeda's articulatory model which has been derived from X-ray images of a female speaker. Three of these parameters (jaw opening, lip opening and protrusion) correspond to visible parameters. The exploitation of the parameters derived from speaker face images requires that face and vocal tract articulatory parameters are consistent together. The adaptation consists in expressing the visual parameters in the coordinate system of the vocal tract articulatory model. We tested two adaptation strategies [31], [32]. The first consists in applying exactly the same factor analysis to visual data as that applied by Maeda to X-ray data. Then, articulatory parameters derived from the speaker's face are used in the same manner as the other articulatory parameters of Maeda's model. The underlying hypothesis is that both speakers, i.e. the one used to build the vocal tract articulatory model and the one whose face images are used during inversion, share common articulatory behaviour to prevent mismatches between the two models. This hypothesis is actually very strong. The second solution consists in a superficial adaptation of the vocal tract articulatory model by mapping the average values of the visual measures onto those of the X-ray data. The expected advantage is to keep the internal consistency of the vocal tract articulatory model since there is no model for the face data. It turns out that the second strategy gives more consistent results even if the inversion results are fairly good with the two methods.

6.1.2. Talking Head

We are working with a view of elaborating a prediction algorithm for labial coarticulation. Last year, we investigated speaker variability in labial coarticulation. This allowed us to find out speaker independent strategies and the requirement for a labial coarticulation algorithm. The prediction algorithm we are now developing focuses only on critical parameters for labial coarticulation. This means that other parameters, which do not require a specific evolution, are not considered at a first time. The overall algorithm exploits a concatenation strategy. Transitions between phonemes are represented in the form of sigmoids that can be easily concatenated. The training corpus was recorded last year. It consists of 3D data acquired for 190 markers painted onto a female speaker's face for CVC and 100 sentences. Since a concatenation strategy would require a huge amount of 3D data, covering all transitions, a completion algorithm has been designed in order to derive other transitions from the VCV and VCCV present in the training corpus. For the same reason, this algorithm derives all the VCV from /y,u,a,i/C/y,u,a,i/ recorded. This algorithm will be compared with that proposed by Cohen and Massaro.

6.1.3. Text-to-Speech synthesis

This year, we have started the development of a word syllabification tool based on three phonetic considerations: the sonority principle (all sounds can be sorted by their phonetic sonority and the sequence of sounds in a syllable have to correspond to the increase of the sonority), the maximal onset theory (it is better to separate two consonants in the way that there are more consonants at the beginning of the following syllable than at the end of the current syllable), and the phonotactic existence in the language (in French /tl/ doesn't occur at the beginning of a word, thus, into a word, the sound /t/ and /l/ cannot belong to the same syllable). For synthesis and recognition, we need to syllabify sentences; thus, we have to add some rules to take into account the link between words (especially when a word finishes with a consonant and the following one starts with a vowel). This preliminary work has been used to train the syllable models of our recognition system (also a preliminary work).

In September, an INRIA associate engineer, Alexandre Lafosse, has been hired. He will develop a software platform in the context of Text-To-Speech (TTS) synthesis. The goal is first to obtain a Natural Language Processing system, which is the first step of a TTS system. The second part of TTS systems - the Non Uniform Unit (NUU) selection system - will follow. The Natural Language Processing (NLP) platform will be built from tools that already exist in the team (syntactic tagger, phonetizer, syllabification). A second objective is to obtain a unified set of tools in the team for NLP. For instance, this toolkit shall be re-used to train different recognition modules. In the next phase, this platform will allow us to continue researches about corpus building for synthesis systems based on NUU selection. Meanwhile, research on feature weighting and selection of non uniform units has been temporarily slowed down [44].

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

The automatic evaluation and correction of learners' most important deviations rely upon a comparison between a model (for the moment, a native speaker's production) and the learner's production. This implies the preliminary segmentation and labelling of both productions. To automatically achieve these tasks, we have implemented an algorithm for automatic alignment, developed in our team. The principle is the following: the text, given by the user, is phonetized and the algorithm, based upon HMM models, aligns the phonetic transcription with the acoustic signal. For non-native speakers, the phonetization must be adapted, and we are presently working on this task. We have also implemented an algorithm that realizes an auditory correction of the learner's production. This algorithm modifies the prosodic cues of a learner's deviant realization while keeping all the other characteristics of the realization unchanged. This modification, made on the learners voice, makes them more sensitive to what they have to produce [26], [25].

6.2. Automatic Speech Recognition

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

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

During the reporting period, the team members have published an extensive review about the current state of the art on automatic speech recognition[11]. The major contributions of the team to this state of the art are summarized in the next section.

6.2.1. Robustness of speech recognition

Robustness of speech recognition to noise and to speaker variability is one of the most difficult challenge that limits the development of speech recognition technologies. We are actively contributing to this area via the development of the following advanced approaches:

6.2.1.1. Bayesian denoising

We collaborate with K. Daoudi from IRIT-CNRS to develop a new denoising approach based on statistical models of clean and noisy speech trained on multi-condition databases. The proposed method is similar in essence to the SPLICE algorithm proposed by Microsoft's researchers, but it further supports adaptation to new noisy environments that do not belong to the training databases. We have investigated two denoising methods that support adaptation in this sense: "Gaussian mapping", which assumes a linear relationship between clean and noisy observations within two corresponding Gaussian clusters, and "cross-correlation", which computes the expected value of the clean speech given the noisy speech and a joint model of clean and noisy speech. We have also investigated different techniques to train the Gaussian mixture models jointly or independently. During testing, the denoiser is first adapted to the unknown environment using a simple adaptation bias that guarantees to preserve the correspondence between the clean and noisy Gaussians [16], [27]. More recently, we have applied this technique to the robust ETSI standard Advanced DSR Front-End [47] on the continuous speech database Aurora4. We are also currently investigating more elaborated adaptation procedures that still support this correspondance.

6.2.1.2. Missing data recognition

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

We recently published an in-depth review about the existing approaches that can be used to compute missing data masks [12]. We further proposed an original solution that is based on statistical models of masks. These models are trained on oracle masks that optimize the signal-to-noise ratio criterion. Our mask models exploit contextual dependencies, both along the frequency axis and in the time domain. This is achieved in the time domain by training state transitions of an ergodic HMM, and in the frequency domain by modeling full-band masks. These full-band masks are also clustered into a smaller number of mask representatives, in order to tackle the combinatory explosion that results from the high dimensionality of the feature vectors. This clustering led to an interesting observation, i.e., only a very small proportion of the possible masks are actually used by the oracle masks estimator. Thus, these mask clusters can easily be trained into stochastic mask models. The use of both dependencies also reduces the number of isolated erroneous mask spots. These approaches have been validated on the Aurora2 and Aurora4 databases [18], [17]. These research technologies have also been transfered into the European HIWIRE project.

6.2.1.3. Non-native speakers

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

Our main motivation is to develop a new approach for non-native speech recognition that can automatically handle non-native pronunciation variants without a significant loss in recognition time performance. As non-native speakers tend to realize phones of the spoken language as they would do with similar phones from their native language, we claim that taking into account the acoustic models of the native language in the modified ASR system may enhance performance. We automatically extracted association rules between non-native and native phones models from an audio corpus recorded by non native speakers. Then, new acoustic models were built according to these rules. This year, we developed a new method, based on the idea that the phone realisation produced by non-native speakers may depend on the grapheme of the uttered words. This method has been evaluated on a non-native english speech database uttered by French, Italian, Spanish and Greek speakers [24], [22], [23], in the context of the European HIWIRE project.

6.2.2. Core recognition platform

6.2.2.1. Broadcast News Transcription

In the framework of the Technolangue project ESTER, we have developed a complete system, named ANTS, for French broadcast news transcription (see section 5.1.9).

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

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

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

6.2.2.2. Speech/music/advertisement segmentation

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

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

For the speech/music segmentation, we focused on a new parameterization based on wavelets. We studied different decompositions of the audio signal based on wavelets (Daubechie, Coiflets, symlets) which allow a better analysis of non stationary signals like speech or music. We computed different energy types in each frequency band. Our first results on an audio broadcast corpus gave significant improvement compared to classical MFCC features [19], [21], [20]. During broadcast programs, a lot of advertisement spots are emitted and could yield false keywords detection. In order to remove such advertisements from the audio stream, we developed an approach based on fingerprint.

The objective of audio fingerprinting is to build an efficient mechanism to establish the perceptual equality of two audio files: not by comparing the (typically large) files themselves, but by comparing the associated fingerprints (small by design). The advantage of using fingerprints are: reduced memory/storage requirements as fingerprints are relatively small; efficient comparison as perceptual irrelevancies have already been removed from fingerprints; efficient searching as the dataset to be searched is smaller. The most important perceptual audio features can be observed in the frequency domain. We study the influence of different frequency ranges and two fingerprint methods.
In this framework, we try to automatically detect speech segments that have same fingerprint: they correspond to either advertisement or jingle.

6.2.2.3. Confidence measure

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

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

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

These new measures were evaluated on a 1-hour broadcast news corpus [35].We also designed several measures that can be directly computed within the recognition process, called frame-synchronous confidence measures. Through these measures, we intend to modify the likelihood score computation of the recognition process engine. These measures are divided into two kinds: strictly frame-synchronous measures and measures with a slight delay [36].

We assessed the measures defined above in an on-the-fly keyword spotting application in order to reduce the false-acceptation rate. Indeed, in this kind of application, we do not have access to the whole sentence (no end) and we cannot keep all the information from the beginning of the sentence.

6.2.3. Ubiquitous speech recognition

We have recently initiated a new research area related to Ambient Intelligence: ubiquitous speech processing. In Ambient Intelligence, the main innovation concerning speech interactions is the concept of implicit speech interactions: traditional Human-Computer dialogs assume that the user is directly interacting with the system. Such speech interactions are explicit. On the contrary, when the user intention is not to communicate with the system, every sentence he/she says can be used as an implicit speech interaction by the system. This can happen for example when the user is talking to someone, in a meeting, in a classroom, or simply when he is listening to the radio. In our team, we are already working on some of these research areas, for instance in the ESTER and LABIAO projects, and ubiquitous speech recognition may thus be considered as a new application domain rather than a new fundamental research area.

Concretely, we are addressing this challenging topic in the context of three activities: The first one concerns an extensive study of the state-of-the-art about user interactions and Ambient Intelligence in the national OFTA group. The second one relates to our contribution in the European project Amigo, described in section 7.2.2. The third one deals with our research work on automatic recognition of dialog acts. Indeed, one of the main feature of implicit speech interaction concerns the computation of non-lexical information from the speech stream, such as emotions or dialog acts.

Dialog acts represent the role of successive sentences, or sequences of words, in the course of a dialog, such as statements or questions. The objective of this work is to automatically identify dialog acts from the user's speech signal. Since the end of 2005, we have extended our work to a larger set of dialog acts, and validated our original approach that exploits prosody, lexical and words position informations on two languages, Czech and French, and on two different tasks: reservation application and broadcast news transcription. We have further proposed and developed a semi-automatic training algorithm for dialog acts models that exploits confidence measures [30], [29], [28].

6.3. Language Models

- **Machine Translation based on phrases.** Our purpose is to build a complete speech-to-speech translation system based on phrases. We implemented both models 1 and 2 proposed by Brown in [41] to train the translation component. We also implemented an original method to learn the associated language model. To find the best target sentence given the source sentence, we developed a decoder based on Viterbi algorithm.
  To evaluate the machine translation quality, we chose to adopt the BLEU method proposed by Papineni [58].
  Currently, we are developping a method to automatically acquire bilingual corpora from movies subtitles. The alignement between subtitles is handled by Viterbi algorithm.
- **Dealing with agreement problem in speech recognition**[33]. We introduced an original model called Features-Cache (FC) to estimate the gender and the number of the word to predict. It is a dynamic variable-length Features-Cache for which the size is determined in accordance to syntagm delimitors. This model does not need any syntactic parsing, it is used as any other statistical language model. Several models have been carried out. The best one achieves an improvement of more than 8 points in term of perplexity and slightly improves the word error.
- **Using statistical language models in recommendation systems** We studied recently the integration of statistical language modeling in usage based recommendation systems. A comparative study has been carried out, showing that recommendation systems and language modeling are similar, and leading to the conclusion that SLM can be integrated in usage based recommender systems. This work has been investigated during a research master training period. The use of n-gram models and trigger models has been studied. The Statistical Grammar of Usage takes into account the sequence of consultations in order to propose resources to the active user[15].
- **Crossing context n-grams.** Distant language models are studied, in the scope of using them in an intelligent prediction framework (non distant prediction). Classical models are, in some cases, unable to predict the correct word. Bigram models are also used to predict distant words. Using these models, we show an improvement of 12.9% of the perplexity. We also study contexts for which the models proposed are more performant than classical models.
- **Unsupervised tagger.** An efficient tagger is necessary in our research activity because classifications and labelled corpora are important information exploited by our models (feature models [61], crossing context n-grams, ...).

  A tagger uses two modules: (a) a training parameters tool and (b) a tagging tool. The training parameters tool uses an iterative process to estimate the parameters of the tagging tool. Taggers have two drawbacks: first, the iterative process needs a manually tagged corpus for bootstrap, and second, the out-of-vocabulary rate increases ambiguity.

  We are developping in collaboration with the Langue et Dialogue team and ATILF laboratory an unsupervised tagger [52] which integrates a guesser. A guesser analyses out-of-vocabulary words morphology and proposes a restricted set of Parts of Speech for each of them.

7. Contracts and Grants with Industry

7.1. National Contracts

7.1.1. TNS project

TNS is a French company which monitors all types of media: written press, radio, television, news agencies, Internet. It collects, selects, analyses, organizes and transmits information. It aims to automatically detect key information useful for its customers. The CIFRE thesis of Emmanuel Didiot takes place in this framework. Last year, we improved segmentation modules. This year, we develop tools to mine and to extract information from text and speech data (see section 6.2.2.2).

7.1.2. STORECO project

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

We are involved in three main tasks:

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

Last year, we have implemented software to perform clustering and training of acoustic models for segmentation. This year, the parameterisation and segmentation algorithms have been developed. Moreover, 17 hours of TV programm have been manually labelled in different categories : speech, speech+music, instrumental music, song, lough, applause, silence, noise.

7.1.3. LABIAO project

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

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

Second, a realistic model of a talking head has been produced in collaboration with the Magrit Project. From a stereo corpus of a female speaker, a dense and realistic 3D mesh has been fitted using linear morphing and radial basis function deformations and a learning-based coarticulation algorithm for realistic motion has been implemented.

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

7.1.4. ST&TAP project

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

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

We have developed an algorithm based on DTW (Dynamic Time Warping) to perform alignment between the recognized words and the teleprompter text.

7.1.5. NEOLOGOS project

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

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

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

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

In this framework, the reference speakers should come out of a selection process which guarantees that their recorded voices are non-redundant but keep a balanced coverage of the voice space. Thus, the collection of the Neologos corpus is a three stage process:

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

The extraction of the reference speakers has been interpreted as a *clustering task*, which consists in partitioning the voice space in homogeneous subspaces that can be abstracted by a single reference speaker. First, the academic partners and FRANCE TELECOM formulated this problem in a general framework which remains compatible with a variety of speech/speaker modeling methods. Then, IRISA, FRANCE TELECOM and LORIA designed each a specific inter-speaker dissimilarity measure. The obtained lists of reference speakers were compared and jointly optimized [13].

7.2. International Contracts

7.2.1. HIWIRE

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

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

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

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

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

Two kinds of activities are planed: long-term research and research for the fixed and mobile plateforms [34].

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

During this year we worked on the following subjects:

- Missing data: we propose original approaches to deal with non-stationary noise (see section 6.2.1.2).
- Non-native speech recognition: we modify lexicon to take into account pronounciation variation due to non-native speakers (see section 6.2.1.3).

5 international publications have been accepted during year 2006.

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

During the annual review (sept 2006), we have performed a live demonstration of non-native speech recognition.

7.2.2. Amigo

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

In this project, we are collaborating with the Langue & Dialogue team in Nancy to continue the efforts we have begun in OZONE, with a focus on multimodality (speech, 2D and 3D gestures with VTT), and on adapting our speech technologies to handle implicit user interactions.

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

7.2.3. Muscle

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

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

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

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

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

This project involves the accurate generation of relevant lip deformations and jaw movements of artificial talking heads as the latter perform significantly poorly than real speakers. This issue is particularly crucial to improve lip reading by deaf people and to learn the articulation of phonemes that do not exist in a native language in the case of language learning. The expected contributions of this project are to improve the modeling of labial coarticulation of Baldi (a talking head developed by Dominic Massaro and Michael Cohen at the Perceptual Sciences Laboratory PSL, University of California at Santa Cruz) in English and French, and to evaluate the benefit of using this talking head for 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 acquired by using the tracking system designed by the Magrit team. Within the context of language learning the work will consist of investigating how Baldi can be used to make the learner more sensitive to acoustical and articulatory features of both French and English sounds. This work will exploit standard phonetic knowledge of French and English pronunciation together with the available articulatory data.

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

7.2.5. ASPI–IST FET STREP

The ASPI (Audiovisual to Articulatory Speech Inversion) project is funded by the European Commission in the framework of the 6th PCRD. The HIWIRE project, started on November 2005, aims at recovering the vocal-tract shape (from vocal folds to lips) dynamics from the acoustical speech signal, supplemented by image analysis of the speaker's face. Being able to recover this information automatically would be a major break-through in speech research and technology, as a vocal-tract representation of a speech signal would be both beneficial from a theoretical point of view and practically useful in many speech processing applications (language learning, automatic speech processing, speech coding, speech therapy, film industry...). The design of audiovisual-to-articulatory inversion involves two kinds of interdependent tasks. The first is the development of inversion methods that successfully answer the main acknowledged difficulties (non-unicity of inverse solution, lack of phonetic relevancy of inverse solutions, impossibility of using standard spectral data), and the second is the construction of an articulatory database that comprises dynamic images of the vocal tract together with the speech signal uttered, and that for several male and female speakers. The partners of this project are KTH (Stockholm), ULB (Brussels), ENST LTCI (Paris), and NTUA-ICCS (Athens). Together with INRIA project Magrit we are involved in this project.

This year, the main achievements concern the fusion of ultrasound data and electromagnetic sensors[14], the investigation of phonetic constraints for inversion and the design of fusion algorithm intended to use 3D data of the speaker's face as constraints on visible articulators, i.e. lips and lower jaw. In addition, we started the acquisition of MRI data in order to elaborate a 3D model of the vocal tract. The main contributions of Parole will be about inversion algorithm, especially inversion from standard spectral data (MFCC for instance) and the incorporation of constraints.

8. Dissemination

8.1. Animation of the scientific community

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

- The members of the team frequently review articles and papers for Journal of Phonetics, JASA, Acta Acoustica, Eurospeech, CSL, Speech communication, TAL, IEEE Transaction of Information Theory, Signal Processing.
- A. Bonneau is an elected member of the Instil Board (Integration of speech technology in learning). She is in charge of the project "assistance to language learning" of the "Plan État Région" and member of Eurospeech scientist committee.
- J.P. Haton is a member of Speech Communication and Computer Speech and Language editorial boards, and of ICSLP program commitee, chairman of French Science and Technology Association.
- Y. Laprie is a member of (LREC, JEP) scientific committee. He is in charge of the "Assistant intelligent" project of the PRST "Intelligence Logicielle" and associate editor of the EURASIP Journal on Audio, Speech, and Music Processing.
- O. Mella, D. Fohr, I. Illina and C. Cerisara are involved in several European and national projects.
- K. Smaïli has reviewed several papers for different journals and conferences.
- K. Smaïli has been invited as lecturer at the Text, Image and Speech Recognition (TISR) Conference.
- I. Illina is a member of the evaluation commission of l'INRIA.
- I. Illina is a member of AFCP[2] board.
- S. Ouni is an area chair at the International Conference on Multimodal Interfaces (ICMI'06).
- C. Cerisara is a member of the program committee of ICMI'06.
- The Parole team has participated to the event "journées de la science" and presented the ANTS system (see section 5.1.9).
- Two members of Parole have been reviewers for the AFCP PhD Thesis award.

8.2. Distinctions

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

8.3. Invited lectures

- Virginie Govaere, INRS,
- Thierry Aubin, 'Acoustic Communication' Group, Université Paris-Sud,
- Didier Desor, Neurosciences Comportementales, UHP/INPL/INRA,
- Ouriel Grynszpan, LIMSI-CNRS / THIM, Université Paris 8,
- Jean Lieber, Orpailleur Group, LORIA,
- Matthieu Chabanas, ICP Grenoble,
- Olov Engwall, KTH,
- Bernd Kröger, UK-Aachen, Germany.

8.4. Higher education

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

[2] Association Française pour la Communication Parlée (French Association for Oral Communication)

8.5. Participation to workshops and PhD thesis committees:

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

9. Bibliography

Major publications by the team in recent years

[1] F. BIMBOT, M. EL-BÈZE, S. IGOUNET, M. JARDINO, K. SMAÏLI, I. ZITOUNI. *An alternative scheme for perplexity estimation and its assessment for the evaluation of language models*, in "Computer Speech and Language", vol. 15, n^{o} 1, Jan 2001, p. 1-13.

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

[3] C. CERISARA, S. DEMANGE, J.-P. HATON. *On noise masking for automatic missing data speech recognition: a survey and discussion*, in "Computer Speech and Language", in press, 2006.

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

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

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

[7] J.-P. HATON, C. CERISARA, D. FOHR, Y. LAPRIE, K. SMAÏLI. *Reconnaissance Automatique de la Parole. Du signal à son interprétation*, Dunod, 2006, http://hal.inria.fr/inria-00105908/en/.

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

[9] S. OUNI, Y. LAPRIE. *Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion*, in "Journal of the Acoustical Society of America (JASA)", PACS numbers: 43.70.h, 43.70.Bk, 43.70.Aj [DOS], vol. 118 (1), 2005, p. 444–460, http://hal.archives-ouvertes.fr/hal-00008682/en/.

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

Year Publications

Books and Monographs

[11] J.-P. HATON, C. CERISARA, D. FOHR, Y. LAPRIE, K. SMAÏLI. *Reconnaissance Automatique de la Parole Du signal à son interprétation*, Dunod, 2006, http://hal.inria.fr/inria-00105908/en/.

Articles in refereed journals and book chapters

[12] C. CERISARA, S. DEMANGE, J.-P. HATON. *On noise masking for automatic missing data speech recognition: a survey and discussion*, in "Computer Speech and Language", in press, 2006.

[13] S. KRSTULOVIC, F. BIMBOT, O. BOËFFARD, D. CHARLET, D. FOHR, O. MELLA. *Optimizing the coverage of a speech database through a selection of representative speaker recordings*, in "Speech Communication", vol. 48, 10 2006, p. 1319-1348, http://hal.archives-ouvertes.fr/hal-00110509/en/.

Publications in Conferences and Workshops

[14] M. ARON, M.-O. BERGER, E. KERRIEN, Y. LAPRIE. *Coupling electromagnetic sensors and ultrasound images for tongue tracking: acquisition setup and preliminary results*, in "International Seminar on Speech Production", 2006, http://hal.inria.fr/inria-00110634/en/.

[15] A. BRUN, D. LANGLOIS, K. SMAÏLI. *Exploration et utilisation d'informations distantes dans les modèles de langage statistiques*, in "13ème Conférence sur le Traitement Automatique des Langues Naturelles - TALN'2006, 10/04/2006, Leuven/Belgique", 2006, p. 425-434, http://hal.inria.fr/inria-00103459/en/.

[16] C. CERISARA, K. DAOUDI. *Evaluation of the SPACE denoising Algorithm on Aurora2*, in "IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2006, France Toulouse", 2006, http://hal.archives-ouvertes.fr/hal-00020669/en/.

[17] S. DEMANGE, C. CERISARA, J.-P. HATON. *Mask Estimation For Missing Data Recognition Using Background Noise Sniffing*, in "IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2006, 19/05/2006, Toulouse/France", 2006, http://hal.inria.fr/inria-00080562/en/.

[18] S. DEMANGE, C. CERISARA, J.-P. HATON. *Missing data mask models with global frequency and temporal constraints*, in "Ninth International Conference on Spoken Language Processing - Interspeech 2006 - ICSLP, 17/09/2006, Pittsburgh, Pennsylvania/USA", 2006, http://hal.inria.fr/inria-00103574/en/.

[19] E. DIDIOT, D. FOHR, J.-P. HATON, I. ILLINA, O. MELLA. *A Wavelet-Based Parameterization for Speech/Music Segmentation*, in "Ninth International Conference on Spoken Language Processing - INTERSPEECH 2006", ISCA, 2006, 653, http://hal.archives-ouvertes.fr/hal-00103569/en/.

[20] E. DIDIOT, I. ILLINA, O. MELLA, D. FOHR, J.-P. HATON. *Speech/music discrimination based on wavelets for broadcast programs*, in "International Conference on Signal Processing and Multimedia Applications - SIGMAP 2006", INSTICC PRESS, 2006, 151, http://hal.archives-ouvertes.fr/hal-00103554/en/.

[21] E. DIDIOT, I. ILLINA, O. MELLA, D. FOHR, J.-P. HATON. *Une nouvelle approche fondée sur les ondelettes pour la discrimination parole/musique*, in "XXVIes Journées d'Etude sur la Parole - JEP 2006", AFCP, 2006, 209, http://hal.archives-ouvertes.fr/hal-00103565/en/.

[22] B. GHAZI, D. FOHR, J.-P. HATON, I. ILLINA. *Reconnaissance de parole non native fondée sur l'utilisation de confusion phonétique et de contraintes graphèmiques*, in "XXVIes Journées d'Etude sur la Parole - JEP'06, 12/06/2006, Saint-Malo, France", 2006, http://hal.inria.fr/inria-00110495/en/.

[23] B. GHAZI, D. FOHR, I. ILLINA, J.-P. HATON. *Fully Automated Non-Native Speech Recognition Using Confusion-Based Acoustic Model Integration And Graphemic Constraints*, in "IEEE International Confer-

ence on Acoustics, Speech, and Signal Processing - ICASSP 2006, 15/05/2006, Toulouse/France", 2006, http://hal.inria.fr/inria-00110492/en/.

[24] B. GHAZI, D. FOHR, I. ILLINA, J.-P. HATON. *Multilingual Non-Native Speech Recognition using Phonetic Confusion-Based Acoustic Model Modification and Graphemic Constraints*, in "The Ninth International Conference on Spoken Language Processing - ICSLP 2006, 17/09/2006, Pittsburgh, PA/USA", 2006, http://hal.inria.fr/inria-00110496/en/.

[25] G. HENRY, A. BONNEAU, V. COLOTTE. *Détection et correction automatique des déviations dans la réalisation de l'accent lexical anglais par des apprenants français*, in "XXVIes Journées d'Etude sur la Parole - JEP 2006", AFCP, 2006, p. 41–44, http://hal.archives-ouvertes.fr/hal-00103643/en/.

[26] G. HENRY, A. BONNEAU, V. COLOTTE. *Making learners aware of the prosody of a foreign language*, in "IV International Conference on Multimedia and Information and Communication Technologies in Education - m-ICTE 2006", FORMATEX, 11 2006, http://hal.archives-ouvertes.fr/hal-00112055/en/.

[27] D. KHALID, C. CERISARA. *An improved version of the SPACE algorithm for noise robust speech recognition*, in "IEEE-EURASIP ISCCSP, 13/03/2006, Marrakech, Morroco", 2006, http://hal.inria.fr/inria-00111910/en/.

[28] P. KRAL, C. CERISARA, J. KLECKOVA. *Automatic dialog acts recognition based on sentence structure*, in "IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP/2006", IEEE (editor). , IEEE, 2006, p. 61-64, http://hal.archives-ouvertes.fr/hal-00078245/en/.

[29] P. KRAL, C. CERISARA, J. KLECKOVA, T. PAVELKA. *Sentence structure for dialog act recognition in Czech*, in "2nd IEEE International Conference on Information et Communication Technologies: from Theory to Applications - ICTTA´06", Syrian Computer Society, SCS, 2006, ., http://hal.archives-ouvertes.fr/hal-00078247/en/.

[30] P. KRAL, J. KLECKOVA, C. CERISARA. *Automatic Dialog Acts Recognition based on Words Clusters*, in "9th Western Pacific Acoustics Conference - WESPAC IX 2006", The Acoustical Society of Korea, 2006, http://hal.archives-ouvertes.fr/hal-00086310/en/.

[31] Y. LAPRIE, B. POTARD. *Adapting visual data to a linear articulatory model*, in "7th International Seminar on Speech Production - ISSP 2006, Sao Paulo/Brazil", 2006, http://hal.inria.fr/inria-00112223/en/.

[32] Y. LAPRIE, B. POTARD. *Adjonction de contraintes visuelles pour l'inversion acoustique-articulatoire*, in "Journées d'Études sur la Parole - JEP 2006, Dinard/France", 2006, http://hal.inria.fr/inria-00112219/en/.

[33] C. LAVECCHIA, K. SMAÏLI, J.-P. HATON. *How to handle gender and number agreement in statistical language models?*, in "Ninth International Conference on Spoken Language Processing - INTERSPEECH 2006, 17/09/2006, Pittsburgh, Pennsylvania/USA", 2006, http://hal.inria.fr/inria-00103497/en/.

[34] A. POTAMIANOS, G. BOUSELMI, D. DIMITRIADIS, D. FOHR, R. GEMELLO, I. ILLINA, F. MANA, P. MARAGOS, M. MATASSONI, V. PITSIKALIS, J. RAMIREZ, E. SANCHEZ-SOTO, J. SEGURA, P. SVAIZER. *Towards Speaker and Environmental Robustness in ASR: The HIWIRE Project*, in "SRIV'06 ITRW on Speech Recognition and Intrinsic Variation, France Toulouse", HIWIRE, 05 2006, http://hal.archives-ouvertes.fr/hal-00110502/en/.

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

[36] J. RAZIK, O. MELLA, D. FOHR, J.-P. HATON. *Mesures de confiance trame-synchrone*, in "XXVIes Journées d'Etude sur la Parole - JEP 2006", AFCP, 2006, p. 135–138, http://hal.archives-ouvertes.fr/hal-00080848/en/.

References in notes

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

[38] C. ABRY, T. LALLOUACHE. *Le MEM: un modèle d'anticipation paramétrable par locuteur: Données sur l'arrondissement en français*, in "Bulletin de la communication parlée", vol. 3, n^{o} 4, 1995, p. 85–89.

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

[40] A. BOTHOREL, P. SIMON, F. WIOLAND, J.-P. ZERLING. *Cinéradiographies des voyelles et consonnes du Français*, Travaux de l'institut de Phonétique de Strasbourg, 1986.

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

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

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

[44] V. COLOTTE, R. BEAUFORT. *Linguistic features weighting for a Text-To-Speech system without prosody model*, in "proceedings of EUROSPEECH/INTERSPEECH 2005", 2005, p. 2549-2552, http://hal.ccsd.cnrs.fr/ccsd-00012561/en/.

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

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

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

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

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

[50] 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.

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

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

[53] D. LANGLOIS, K. SMAÏLI, J.-P. HATON. *Retrieving phrases by selecting the history : application to Automatic Speech Recognition*, in "7th International Conference on Spoken Language Processing - ICSLP'2002, Denver, USA", vol. 1, September 2002, 721.

[54] 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.

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

[56] 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.

[57] S. OUNI, Y. LAPRIE. *Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion*, in "Journal of the Acoustical Society of America (JASA)", PACS numbers: 43.70.h, 43.70.Bk, 43.70.Aj [DOS], vol. 118 (1), 2005, p. 444–460, http://hal.archives-ouvertes.fr/hal-00008682/en/.

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

[59] B. POTARD, Y. LAPRIE. *Using phonetic constraints in acoustic-to-articulatory inversion*, in "Proceedings of the 9th European Conference on Speech Communication and Technology - Interspeech - Eurospeech 2005", 2005, p. 3217-3220, http://hal.archives-ouvertes.fr/hal-00014057/en/.

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

[61] 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.

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