INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team parole

Analysis, Perception and speech recognition

Lorraine

THEME 3A

Activity Report

2003

Table of contents

1. Team

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

Head of project-team
Yves Laprie [Research scientist HDR, CNRS]

Administrative Assistant
Martine Kuhlmann [CNRS]

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

INRIA Research scientist
Khalid Daoudi [Research scientist]

Faculty member
Armelle Brun [Assistant Professor, U. Nancy 2]
Vincent Colotte [Assistant Professor, U. H. Poincaré]
Joseph di Martino [Assistant Professor, U. H. Poincaré]
Jean-Paul Haton [Professor, U. H. Poincaré, Institut Universitaire de France]
Marie-Christine Haton [Professor, U. H. Poincaré]
Irina Illina [Assistant Professor, I.U.T Charlemagne, U. Nancy 2]
David Langlois [Assistant Professor, IUFM]
Odile Mella [Assistant Professor, U. H. Poincaré,working for CNRS since 1 septembre 2002]
Nathalie Parlangeau-Vallès [Maître de conférences, I.U.T Charlemagne, U. Nancy 2]
Kamel Smaïli [Professor, U. Nancy 2]

Phd Students
Vincent Barreaud [TA]
Yassine Benayed [TA]
Murat Deviren [TA]
Salma Jamoussi [TA]
Fabrice Lauri [CIFRE grant]
Vincent Robert [High school teacher]

Project technical staff
Christophe Antoine [Research scientist partner, DIALOCA]
Sen Zhang [project technical staff INRIA]

Research scientist partner INRIA
Filipp Korkmazsky

2. Overall Objectives

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

Our activities are structured in two topics:

- **Speech analysis** Our works are about automatic extraction and perception of acoustic cues, acoustic-to-articulatory inversion and speech analysis. These themas give rise to a number of ongoing or future applications: vocal rehabilitation, improvement of hearing aids, language learning.

- **Modeling speech for automatic recognition** Our works are about stochastic models (HMM[1], bayesian networks and missing data models), multiband approach, adaptation of a recognition system to a new speaker or to the communication channel and on language models. These topics give also rise to a number of ongoing or future applications: automatic speech recognition, automatic translation, text-to-speech alignment, audio indexing.

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

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

3. Scientific Foundations

3.1. Introduction

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

Taken as a whole research in speech gave rise to two kinds of approach:

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

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

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

These two approaches are complementary; the second borrows theoretical results on speech from the first, which, in its turn borrows some numerical methods. Spectral analysis methods are undoubtedly the

[1] Hidden Markov Models

domain where exchanges are most marked. The simultaneous existence of these two approaches is one of the particularities of speech research conducted in Nancy and we intend to enhance exchanges between them. These exchanges will probably grow in number since automatic speech recognition systems (especially those for automatic dictation) are now available for every consumer: their acoustical robustness (against noise and speaker variation) and their linguistic reliability must be increased.

Our activities are structured according to these two approaches:

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

3.2. Speech Analysis

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

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

3.2.1. Acoustic cues

We have introduced the notion of strong and weak cues to palliate a weakness of ASR systems: the lack of certitude. Indeed, due to the lack of variability of speech signals, acoustical regions representing different sounds overlap one with another. Nevertheless, we know, from previous perceptual experiments [38][37], that some realizations of a given sound can nevertheless be discriminated with a high level of confidence. That is why we have developed a system for the automatic detection of strong cues, devoted to the reliable recognition of stop place of articulation. Strong cues, as we call them, identify or eliminate a phonetic feature with certainty (no error is allowed). Such a decision is possible in few cases, when an acoustic cue has a high power of discrimination and is well marked. During strong cue detection, we must fulfil two requirements: to make no error on the one hand, and to obtain a relatively high firing rate, on the other hand. The notion of strong cue must not be merged into that of "robust" cue or landmark which are systematically fired and can make some errors. On a corpus, made up of approximately 2000 stops, we obtained a firing rate for stop bursts and transitions in one case out of four [21].

Strong cues can be exploited either in language learning in order to enhance the most reliable cues or in ASR to provide "confidence islands" so as to reduce the search space during the lexical access.

3.2.1.1. Automatic detection of "well realized" sounds

The detection of strong cues confirms that a same sound, depending on its realization, can be identified with a very different level of confidence. Sounds that are identified with certitude are probably well realized and well pronounced sounds. We made the hypothesis that the enhancement of well realized sounds in a sentence gives listeners some islands of confidence during the acoustic decoding stage and improves speech intelligibility. That is why we decided to detect these sounds in an entirely automatic way, with HMM models. This detection and the results obtained for stops are detailed in the new results.

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

[2] Hidden Markov Models

for normally hearing people in noisy environments. Our project concerning the design and development of computer-assisted learning of prosody is presented in section 7 (national projects).

3.2.2.1. Speech signal transformation

In order to improve oral comprehension, we use a speech signal transformation method called PSOLA (Pitch Synchronous Overlap and Add). This method is based on the decomposition of the speech signal into overlapping pitch synchronous frames. Signal modifications consist in manipulating analysis marks to generate new synthesis marks. This method is well known for its easy implementation and the quality of the slowed down signals. However, temporal discrepancies can appear in the region of the synthesis marks and noise can be generated between harmonics. In order to reduce the loss of quality, we improved the method in the two following ways. Firstly, we introduced a pruning algorithm to seek analysis marks (for pitch synchronization). It increases the robustness of pitch marking for speech segments with strong formant variation. Secondly, we improved the localization of analysis and synthesis marks. During the analysis stage, we can either oversample the signal or use F0 detection algorithm which gives an accuracy better than one sample. During the synthesis stage, the improvement is based on a dynamical re-sampling of the speech signal so as to accurately replace the frame on synthesis marks. Both improvements strongly reduced the level of noise between harmonics and we obtained a high quality speech signal [41].

3.2.2.2. Perceptual experiments

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

3.2.3. Acoustic-to-articualtory inversion

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

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

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

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

Articulatory modeling Articulatory models aim at describing all the possible vocal tract shapes with a small number of parameters, while preserving deformations observed on a real vocal tract. Present articulatory models often derive from data analysis of cineradiography moving pictures. One of the most widely used is that built by Maeda [45].

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

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

As our contribution only concerns inversion, we accepted widely used articulatory synthesis methods. We therefore chose Maeda's articulatory model, the acoustical-electrical analogy to compute the speech spectrum and the spatio-temporal method proposed by Maeda to compute the speech signal. As regards inversion, we chose Maeda's model to constrain vocal tract shapes because this model guarantees that synergy and compensation articulatory phenomena are still possible, and consequently, that articulatory deformations close to those of a human speaker may be recovered.

3.3. Automatic speech recognition

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

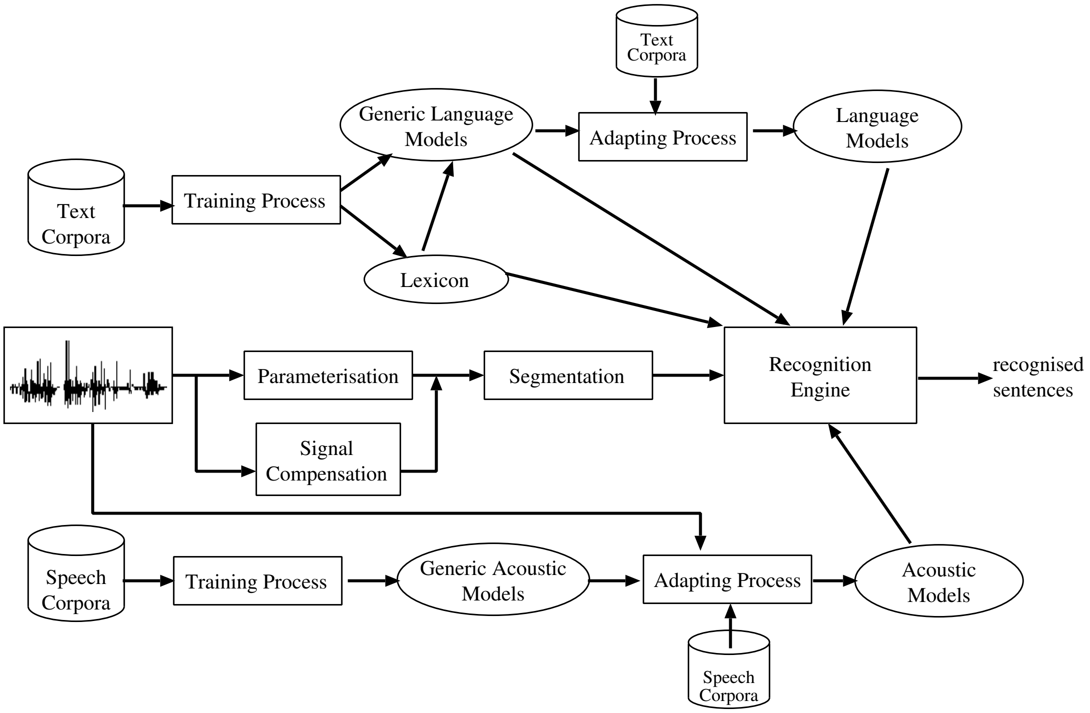

Figure 1. Speech recognition process

Figure 1 shows the different components which are required in the automatic speech recognition process. It also introduces the research topics of automatic speech recognition which we are working on: language modeling, acoustic modeling, robustness and invariance to different speakers or various environments (as noisy or spontaneous speech) with adapting or compensating methods, speech/non speech segmentation. Despite the fact that all of these components are tightly linked, to be clear we gather our research activities in two sections: "Acoustic features and models" and "Language models".

3.3.1. Acoustic features and models

3.3.1.1. Acoustic models

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

3.3.1.2. Robustness and invariance

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

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

3.3.1.3. Segmentation

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

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

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

3.3.2. Language modeling

Acoustically, we can consider that several systems today achieve good results. Nevertheless, some problems due to the complexity of natural language remain without a satisfactory solution. Our group, as others through the world, makes more and more efforts in order to make them more efficient. All the language models we propose are based on information theory and statistics. Some of them use linguistic knowledge to guide the statistical one. The current state of the art in this domain shows that the majority of language models fit for use in speech recognition have a very narrow scope. Some of them use a history which is more or less distant as cache or triggers models. Even if the combination of these models with the baseline one achieves better results, we have to improve them in order to take into account the wide complexity of natural language. To do so, we work through several directions:

- Language model adaptation using topic identification. The objective of this research area is first to find out the topic of the uttered sentences, second, to adapt the baseline language model using the one which corresponds to the retrieved topic. Research is in both identification and adaptation.
- Modeling distant relationship. This is necessary because in natural language the relationship between linguistic units is not necessarily contiguous but may be in some cases distant. For instance, a verb could be linked to a subject which occurred n words (with $n > 3$) before. The research activity consists in modelizing these distant relationships and finding the best framework for that.
- Speech understanding process. We believe that speech recognition system will be more efficient when the recognition process will be connected to an understanding one. For that purpose, we work on a new system based on a naïve bayesian classifier which allows for translating a signal to conceptual tags which in turn are translated to SQL requests [29].

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

4. Application Domains

Our research works are applied in a variety of fields from automatic speech recognition to paramedical domains. Speech analysis methods will contribute to the development of new technologies for language learning (e.g., for hearing-impaired persons or for foreign language teaching) as well as for hearing aids. We have developed in the past a set of teaching tools based on the speech analysis and recognition algorithms of the group (cf. the ISAEUS [28] project of the EU that ended in 2000). We are continuing this effort toward the diffusion of a course on Internet. Speech is likely to play an increasing role in man-machine communication. Actually, speech is a natural mean of communication, particularly for non-specialist persons. In a multimodal environment, the association of speech and designation gestures on touch screens can for instance simplify the interpretation of spatial reference expressions. Besides, the use of speech is mandatory in many situations where a keyboard is not available: mobile and on-board applications (cf. the IVOMOB project of RNRT for the use of speech recognition in a car), interactive vocal servers, telephone and domestic applications, etc.

Most of these applications will necessitate to integrate the type of speech understanding process our group is presently studying. The automatic indexing of audio and video documents is a very active field that will have an increasing importance in our group in the forthcoming years, with applications such as economic intelligence, automatic transcription, or key word spotting.

5. Software

5.1. Software tools

5.1.1. PhonoLor

PhonoLor is a phonetizer enabling the word transcription or a sentence to be translated into a sequence of phonemes. This software exploits phonetization rules learnt from a corpus of examples.

5.1.2. Snorri and WinSnoori

Snorri is a speech analysis software we have been developing since 10 years. It is intended to facilitate the work of the scientist in automatic speech recognition, phonetics or speech signal processing. Basic functions of Snorri enable several types of spectrograms to be calculated and the fine edition of speech signals (cut, paste, and a number of filterings) because the spectrogram allows the acoustical consequences of all the modifications to be evaluated. Beside this set of basic functions there are various functionnalities to annotate speech files phonetically or orthographically, to extract fundamental frequency, to pilot the Klatt synthesizer and to utilize PSOLA resynthesis.

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

5.1.3. Labelling corpora

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

5.1.4. Automatic lexical clustering

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

5.1.5. SALT

SALT (Semi-Automatic Labelling Tool).

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

5.1.6. LIPS

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

5.1.7. ESPERE

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

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

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

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

5.2. Corpus

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

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

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

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

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

6. New Results

6.1. Speech Analysis

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

6.1.1. Acoustic cues

We have developped a method to detect automatically “well realized sounds”. The aim of this work is to improve language learning through speech signal modifications. Previous studies have shown that such an enhancement and/or slowing down of some sounds improve the perception of a second language as well as that of the first language for hearing impaired people. Our studies have confirmed these results for language learning. Indeed, we have enhanced the spectral characteristics of non-voiced stops and fricatives consonants, and slowed down transitions of French sentences; the results showed that these modifications improved the perception of learners of French as a foreign language. The specificity of our approach relies in its entirely automatic process.

Another approach consists in enhancing only the sounds which are well realized, so as to provide listeners some islands of confidence during the acoustic decoding stage. But the detection of these well realized sounds in an automatic manner is not obvious. On one hand, it is possible to find well realized features with a speech recognition system based upon phonetic knowledge (acoustic cues). We have shown the existence of strong cues, which allow features to be identified with certainty for stops. But this method cannot be entirely automatic, especially because of segmentation problems. Stochastic methods, such as the Hidden Markov Models(HMM), can recognize sentences in en entirely automatic way. But, if these systems obtained very high overall recognition scores, they do not give any indication to the way one sound in particular has been realized.

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

6.1.2. Acoustic-to-articulatory inversion

The strength of our inverse method resides at the quasi-uniform acoustic resolution of the articulatory table. This property originates in the construction method that evaluates the linearity of the articulatory-to-acoustic mapping at each step. Articulatory parameters of Maeda’s model vary between -3 and 3 σ where σ is the standard deviation. Thus, the codebook inscribes a root hypercube. Sampling the articulatory space amounts to find reference points that limit linear regions. Therefore we evaluate linearity on segments between any two vertexes in the hypercube considered: acoustic values are linearly interpolated at the middle point between two vertexes from the acoustic values calculated at these vertexes and the result is compared against that directly given by the articulatory synthesizer. If the difference between the acoustic values synthesized and those interpolated is less than a predefined threshold, the hypercube is considered to be linear concerning the articulatory-to-acoustic mapping. Otherwise this hypercube is decomposed into sub-hypercubes and the linearity test is repeated for every new hypercube. Actually, the cube decomposition is also applied if one or several vertexes of the cube lead to a non valid vowel area function, i.e. an area function with a complete constriction. Such cubes delimitate the articulatory space. As for linearity, the decomposition is stopped if the edge becomes smaller than a predefined value and invalid cubes at the boundary are discarded. A small edge value was thus used to guarantee that the boundary of the articulatory space is known with a sufficient precision. Unfortunately, it turned out that the boundary decomposition leaded to a vast number of small cubes since each decomposition gives 128 (2^7) new cubes even if only a very small number of vertexes among them are invalid. Conversely, accepting a coarse edge to discard cubes with some invalid vertexes would lead to the elimination of articulatory regions that are important to produce cardinal vowels and particularly /u/. Therefore we decided to decompose cubes with invalid vertexes only if the ratio of invalid vertexes with respect to the total number of vertexes, i.e. 128, is greater than a predefined threshold, or if the jacobian at the hypercube centre cannot be calculated.

6.1.3. Sinusoidal modeling

Sinusoidal modeling has been successfully applied to a wide range of audio signal processing problems, such as coding or time and frequency stretching. While many methods have been proposed for the analysis part of the process, it seems that there is some general agreement concerning the synthesis part in the non-overlapping case: it is very often achieved using the well-known McAulay-Quatieri method, which consists of an 3 order polynomial reconstruction of the phases of the sinusoidal model partials. In the paper we proposed [27], we compare this “classical” approach with both a simpler (order 1, that is linear interpolation) and a more complex (order 5) polynomial model for phase interpolation of quasi-harmonic signals. A gain has been measured in the signal-to-noise ratio at the synthesis stage, although the performance is limited by the amplitude model and by the imprecision in the analysis stage.

6.2. Automatic Speech Recognition

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

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

6.2.1. Robustness of speech recognition

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

6.2.1.1. Speaker adaptation

Reducing acoustic mismatches due to speaker variability between the training conditions and the testing conditions is a major problem in automatic speech recognition. This problem is particularly difficult for rapid adaptation, when the available amount of adaptation data is small. We have investigated different methods for rapid speaker adaptation. These methods integrate the concepts of both Structural Maximum Likelihood Linear Regression and Eigen-Voices-based technique to adapt the Gaussian means of the speaker independant models for a new speaker [33].

Two new approaches to rapid speaker adaptation of acoustic models by using genetic algorithms have been proposed in [32]. The first approach consists in using a genetic algorithm to adapt the set of Gaussian means to a new speaker. The second approach uses the genetic algorithm to enrich the set of speaker-dependent systems employed by the EigenVoices.

6.2.1.2. Noise compensation

Two classes of methods exist to deal with noise robustness. The first one, that is called here noise compensation, consists to pre-process the acoustic signal prior to recognition, while the second, model adaptation, rather modifies the acoustic models. We first summarize our work on noise compensation. We developed a frame-synchronous noise compensation algorithm designed to cope with time-varying unknown noise [14][13]. This method estimates simple mapping function in parallel with Viterbi alignment. The hierarchical mapping function for this methodis proposed in [16][15]: the transformation tree is built from the states of acoustic models. The objective of this hierarchical transformation is to better compensate for non-linear distortions of the feature space. We also proposed a version of this algorithm that takes into account the abrupt changes in the acoustical environment [12] : the environment change is detected using the Shewart Control Charts detection algorithm that searches for the changes in the means of Gaussian sequence. For each noisy environment a specific mismatch function is estimated. A simple bias is used as mismatch function. For various tasks, proposed methods significantly outperform classical compensation/adaptation methods.

6.2.1.3. Model adaptation

We have proposed several major improvements of the recently appeared Jacobian adaptation method. The most important advantages of Jacobian Adaptation are its very low requirements, both in terms of CPU processing and adaptation data. This is particularly important to be able to quickly adapt the models to a sudden environment change, for example when the car speeds up or down. We have proposed a method to dynamically estimate some parameters of the algorithm [40]. The goal is twofold: to suppress the need of a development corpus, and thus to improve the stability of adaptation. Another interesting property of Jacobian adaptation is that it is achieved only in the features domain and could in theory work with any kind of parameterization. This is however not completely true, because the Jacobian matrices are derived for now only for the cepstral coefficients. We are currently working on a solution addressing this issue and making Jacobian adaptation totally independent of the acoustic front-end. At our knowledge, no other standard adaptation method offers this possibility, although it might however become very important in the near future, when model adaptation techniques will be combined with robust parameterizations.

Another work in the field of model adaptation concerns the comparison of several methods dedicated to fast adaptation to convolutional or additive noise. The following methods have been tested on the VODIS database (speech corpus recorded in a car): PMC (Parallel Model Combination), CMS (Cepstral Mean Subtraction) and a new algorithm that combines both PMC and CMS in the spectral domain [39].

We further collaborated with Granada University on a comparative study of noise compensation and adaptation approaches dedicated to improve robustness to background noise in a car [43].

6.2.1.4. Supervised-predictive noise compensation

We developed a new noise compensation scheme, that we called *supervised-predictive* compensation, which is different in its concept from all known compensation schemes [25]. This scheme can be applied in scenarios where training speech has been recorded in different noise conditions. The principle then is to use a supervised learning procedure to estimate the parameters of an hypothesized (parametric) model that attempts to describe how matched models vary w.r.t. noise models. This new scheme has many advantages w.r.t. classical adaptive and predictive compensation techniques. Moreover, we showed that it works perform significantly better than multi-conditions training, which is the most widely used technique in these kind of scenarios.

6.2.1.5. Missing data recognition

Finally, we very recently began some research works with the objective of adapting acoustic models to “highly” non-stationary noises, like musical noise, which often occur simultaneously with speech in radio and television talks. The basic principle consists in decomposing the recorded signal into different streams, for example one for music and one for speech. We studied two different approaches to deal with musical noise: (i) missing data recognition, which masks the spectro-temporal coefficients that do not come from the main speaker [23], and (ii) a modification to the PMC algorithm that takes into account the non-stationary properties of music [24]. Much work has still to be done in this area, for example to develop new algorithms that dynamically infer which parts of the spectrum are corrupted by noise.

6.2.2. Core recognition platform

6.2.2.1. Discriminative training

We have started research work on discriminative training. Tests were conducted on speech/nonspeech classification for broadcast news. Six models were trained : pure speech, speech with music in the background, speech with song in the background, music, song and noise. In the conducted experiments, discriminatively trained models achieve higher accuracy than the maximum likelihood trained models for a speech/music classification task. We also have developed an approach that significantly reduced needed time for discriminative training by using a subset of training data and an optimal step size estimated for this subset. For the task of speech/music classification, a substantial reduction of the time needed was achieved for discriminative training. Using discriminatively trained phoneme models, a higher phoneme string recognition accuracy was obtained in comparison to the corresponding maximum likelihood trained phoneme models.

6.2.2.2. Automatic speaker clustering

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

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

6.2.2.3. Keywords detection

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

First, we investigated the detection approach based on keyword and filler models [17]. We proposed the use of confidence measures in order to make the decision of rejection or acceptance of a given keyword. The various confidence measures used are based on the probability of the local acoustic observation or on the loop of phonemes recognition method. We also considered the problem of detection as a classification problem where each keyword can belong to two different classes, namely “correct” and “incorrect”. This classification is carried out using Support Vector Machines (SVM) which constitutes a new technique of statistical training. Each recognized keyword is represented by a characteristic vector which constitutes the entry of the SVM classifier. In order to improve the performances, we proposed hybrid approaches combining the garbage models with the confidence measures and the confidence measures with the SVM [18],[19].

We have also carried on with our study on keyword detection in radio programs as broadcast news, interviews, in the framework of the project Raives (cf. section 8.2.3. Last year we implemented a keyword detection method based on keyword and filler models [46]. This year we supplemented this approach with some tests comparing two filler models and various weighting factors of the keyword models [34]. Above all, we elaborated, evaluated and compared on a read speech corpus, two 1000-keyword detection systems based on two different approaches: the previous one (keyword and filler models) and a second approach based on a large vocabulary continuous speech recognition system (LVCSR) [36].

6.2.2.4. Speech/Music segmentation

Audio indexing and automatic broadcast news transcription need the segmentation of the audio signal into speech and music portions. We have started a study about this topic. For that purpose, we have chosen a competing modeling approach based on 4 Gaussian Mixture Models (GMM) which models speech, instrumental music, songs and speech with music in background. Different kinds of parameterisation of the signal have been evaluated on real data (broadcast news, interviews and musical programs from a French radio). This corpus is considered as quite difficult because :

- there are a lot of superimposed segments (speech with music in background),
- broad-band speech segments often alternate with telephone speech,
- some interviews are very noisy, etc.

Good results have been obtained with the variances of the 12 first static MFCC or with the 12 first derivative MFCC. We have also compared our approach to a class/non-class modeling approach in collaboration with the IRIT laboratory [35].

6.2.3. Dynamic Bayesian networks (DBNs)

One of the main weaknesses of HMM-based ASR systems is the fact that, while speech temporal dynamics are well captured by HMMs, the frequency dynamics (which are phonetically very informative) are weakly modeled. One of our main objectives has been then to develop DBN models capable of capturing *both* temporal and frequency dynamics of speech. The DBN-based multiband [42] approach has been an attempt in this direction. The promising results we obtained motivated us to take a deeper look at the speech parameterization itself. Indeed, in order to address the problem of modeling frequency dynamics, one needs to work with speech features that are localized in the frequency domain and, preferably, that have a meaningful physical interpretation. The Mel-frequency cepstral coefficients (MFCC), which are the most widely used features in current speech recognition systems, do not have such properties. Surprisingly, MFCC-based parameterizations lead to the best recognition performances and it is difficult to compete with them. Recently however, a new and promising approach to extract frequency-localized features, called *Frequency Filtering* (FF), has been proposed. It has been reported that FF features lead to similar or better performances when compared to MFCC in clean and artificial noise conditions. But when we tested FF in real noise conditions (using the Aurora3 database), the performances of MFCC were significantly higher in general. By looking at FF from a wavelet analysis perspective, we provided an explanation to this behavior and a unifying framework for frequency filters design. More importantly, this lead us to design new wavelet-based speech features which are localized in the frequency domain and yield similar or better performances than MFCC in *all* conditions [26]. Our goal now is to use such features to model speech frequency dynamics using DBNs.

Another aspect of our work concern fast inference of dynamic Bayesian networks. When operating a structure learning procedure to model the acoustic units, the resulting DBN structures may be different for each unit. Thus, in continuous speech recognition, the problem is to find the most probable sequence of DBNs given the observation. We developed an efficient algorithm to solve this problem based on a parallel dynamical programing technique [11].

We are also interested in the field of discriminative learning of graphical models. The motivation behind this interest is our desire to develop DBN speech models that are discriminatingly trained instead of the usual generatively trained models. Actually, as compared to generative learning, there has not been so much work done in discriminative learning of graphical models. A first version of a survey on this subject can be found in http://www.glue.umd.edu/~acardena/Research.html. This work has been realized during the summer internship of Alvaro Cardenas, a PhD student in Electrical and Computer Engineering at the university of Maryland, College Park.

Currently, a research theme that is retaining our attention is language modeling. Indeed, all actual statistical language models (n-gram and class-based models in particular) can be unified within the framework of (discrete) dynamic Bayesian networks. We thus developed new DBN-based language models which rely essentially on the statistical data and which combine word and class information in a more natural way than classical techniques. The preliminary experimental results we obtained are very promising and our investigation in this field is continuing.

6.3. Language Models

Language modeling is one of the important activities of our team. In spite of all the improvements obtained by the international communauty in this area, the results are not entirely staisfactory. This is due to the high complexity of natural language. To cope with these limits, our group proposes several different and complementary solutions.

We are highly interested in language model adaptation to improve speech recognition quality. In our case, language models are adapted to the topic of the utterance. Topic identification consists in assigning a label to an utterance, among a set of predefined topics. During speech recognition, given the set of words recognized, the topic is identified and the corresponding language model is used for the next words to be recognized.

During this year, our work has put forward two main parameters in topic detection: the topic detection method and the vocabulary. The vocabulary is made up of the most significant words for topic detection. The topic detection method defines the way the vocabulary words are treated. The use of the appropriate couple method/vocabulary has lead to an improvement of topic detection performance. Moreover, we have showed that the use of specific vocabularies leads to a large improvement of performance when several topic identification methods are combined [22].

The second research domain in language modeling consists in studying relationships between syntagms and components in a text. These relationships can be syntactic or semantic and most of them concern non-contiguous components. As precised above, such relationships are distant. To study this linguistic feature, we adapted classical n-gram models to use distant n-gram models. We used a linear combination between several distant models. First results are encouraging. To improve these results, we adapted the linear combination following two ways: first, we adapted the linear combination to the history using classifications of histories [31] (linear dependent combination). Second, we replaced the linear combination paradigm with a selection paradigm: the different models are not statistically combined in a whole model, but one model is dynamically selected among others according to a measure of its prediction capacity. This selection method is as efficient as a linear dependent combination while requiring less parameters. Another important consequence of this work is a precise study which allows to find out a strong relationship between distant language models and automatic retrieving of phrases. These units have been integrated with profit into a baseline language model [44].

Speech understanding is another research activity in which we are involved. The automatic speech understanding problem could be considered as an association problem between two different languages. The request expressed in natural language is transformed in terms of concepts. A concept represents a given meaning, it is defined by a set of words sharing the same semantic properties. Last year, we propose to use a naïve bayesian classifier to automatically extract the underlined concepts. We also propose a new approach for the vector representation of words. This step allows to validate our speech understanding approach. In fact, on a test corpus automaticaly rewritten in terms of concepts has been transformed on SQL requests and achieved a result of $92,5\%$ of well formed SQL requests [30]. The understanding process has been integrated in our speech recognition system ESPERE. The first results are very encouraging [29].

Another point is concerned with the integration of impossible events in statistical language models. This new and original concept tries to cut off all the impossible linguistic events from the classical language models. The challenge consists in finding out automatically these events [31].

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

7. Contracts and Grants with Industry

7.1. National Contracts

7.1.1. RNRT IVOMOB Project

The main objective of this project is to build up a speech recognition engine which could be used in a car. Three compagnies DIALOCA, Mémodata and Technium are involved.

Currently, we are mainly working on speaker adaptation (6.2.1.1) and keyword detection (6.2.2.3). Another axis of research in this project consists in introducing understanding. This module aims at understanding and answering to a request formulated by a driver. A request could be for example: *Would you please show me the next gaz station ?* [29]

7.1.2. NEOLOGOS project

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

The aim of NEOLOGOS is to create new kinds of speech databases. The first one is an extensive telephone database of children's voices, called PAIDIALOGOS. For that database, one thousand of different children will be recorded, using both the GSM and PSTN telephone networks in the following proportions: 65% over PSTN and 35% over GSM. The second is an extensive telephone database of grown up voices, called IDIOLOGOS. For the first year of this project, we participated to the specification of both databases (speakers, sentences, calls) and to the preliminary study of the appropriate methods which allow the selection of the 200 relevant speakers among the thousand ones. These reference speakers, named “eigenspeakers”, have to maximise the coverage of the whole speaker space.

7.2. International Contracts

7.2.1. *OZONE*

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

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

The work realized this year essentially concerned with the design of the Service Enabling Layer, and more specifically the speech recognition component and its interface with the linguistic parser. These modules are now of being implemented and tested for the demonstration.

7.2.2. *MIAMM*

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

The objective of the MIAMM project is to provide an integrated and comprehensive framework for the design of modular multidimensional/multimodal dialogue systems. This dialogue system is a musical database query system. The architecture of the whole prototype is modular: one module for each task and/or partner (French/English/German speech recognition, French/German parsing, dialogue manager...). The MPEG7 standard format, based on XML, is used for communication between modules.

In this project, this year, we provided the speech modality using the ESPERE system, a speech recognition system developed in our team and dedicated to small and medium vocabularies (several hundred words). The system is based on classical Hidden Markov Models technology. We integrated ESPERE into the whole system using the common data exchange format MPEG7. We also adapted/created a language model dedicated to such innovative application, which is a classical bigram. But, we trained this language model using a training corpus generated with a context-free grammar. This non-classical way to build a model comes from the fact that no real-life training corpus is available for multidimensional/multimodal applications.

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

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

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

8. Other Grants and Activities

8.1. Regional Actions

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

The aim of the project is to design a computer-assisted learning system of English prosody for French students [20]. The development of this system has been achieved in the framework of a project supported by our region and gathering scientists from different domains (phonetics, automatic speech processing, ergonomy and language learning).

The system exploits signal visualization and transformation techniques that are intended to be used by teachers of foreign languages in their courses. Besides signal processing and automatic speech recognition tools, our system includes a course on prosody designed for teachers and will contain a database of characteristic sentences. Our objective is twofold: first to train teachers, then, through the experience of teachers, to make students aware of foreign language prosody by listening, visualisating and exaggerating errors and targets to be reached from their own productions.

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

We are currently recording sentences extracted from the Timit corpus and uttered by young French speakers of a high college and two universities of Nancy. About 1000 sentences already have been collected.

8.2. National Actions

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

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

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

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

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

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

8.2.2. MathSTIC Project

Probabilistic graphical models for automatic speech recognition

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

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

8.2.3. RAIVES-STIC-SHS Project

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

For audio documents, one solution is automatic indexing. It consists in finding good descriptors in audio documents which can be used as indexes for archiving and search. Last year we started a French research project on audio indexing, RAIVES (Recherche Automatique d'Informations Verbales Et Sonores). Audio indexing systems can be based on a complete transcription but it is not the only meaning-full information which can be extracted from an audio document. Non-verbal information (as music, jingles or speakers) is also informative for an audio document, and can lead to the extraction of pertinent descriptors. We focus on this kind of information extraction. Therefore, the aim of this project is to automatically separate speech segments from music segments, detect key sounds (like jingles), identify the language of a segment, split the audio signal according to the speakers, detect some keywords and extract the main topics. For the two first years of this project, we participated to the specification of the corpus : 180 hours of the French public radio station RFI (Radio France International). Programs are broadcast news as well as interviews and musical programs. We also hand-labelled 2 hours of audio signal. We elaborated, evaluated and compared two 1000-keyword detection systems based on two different approaches. The first one uses keyword and filler models [46], [34]. The second approach is based on a large vocabulary continuous speech recognition system (LVCSR) to produce a word string. Then, search algorithms are applied for keyword detection in that string [36] (cf. section 6.2).

8.2.4. ESTER Project

As, in USA, NIST organises every year an annual evaluation of the systems performing an automatic transcription of radio and television broadcast news, the French association AFCP (Association Francophone de la Communication Parlée) has initiated such an evaluation for the French language, in collaboration with ELRA (European Language Resources Association and DGA (Délégation Générale pour l'Armement). The ESTER (Evaluation des Systèmes de Transcriptions Enrichies des émissions Radiophoniques) project is supported by the French research ministry (CNRS-Technolangue-EVALDA) for two years. ESTER is composed of two evaluation phases. Phase one evaluates the segmentation systems, as speech/music segmentation, the speaker tracking systems and the orthographic transcription systems. We have decided to participate in the evaluation of the orthographic transcription task and of the acoustic segmentation task (as speech/music/noise). We are now working on the "dry run" test. It consists providing the segmentation and the orthographic transcription of about five hours of French radio programs. The available data for this task are 30 hours of transcribing radio broadcast news for training acoustic models, 15 years of text corpora from the newspaper "Le Monde" to train the language models, and 5 hours of transcribing radio data for the development.

9. Dissemination

9.1. animation of the scientific community

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

- The members of the team frequently review articles and papers for Journal of Phonetics, JASA, Acta Acoustica, Eurospeech, CSL, Speech communication, TAL, IEEE Transaction of Information Theory, Signal Processing.
- A. Bonneau is an elected member of the Instil Board (Integration of speech technology in learning). She is in charge of the project "assistance to language learning" of the "Plan Etat Region" and member of Eurospeech scientist commitee
- J.P. Haton is a member of CSL and ICSLP programm commitee. Chairman of French Science and Technology Association
- M.C. Haton chairman of CNRS scientist movie 2003
- Y. Laprie is a member of (LREC, JEP) scientific commitee. He is an elected member of G.F.C.P, groupe francophone de la communication and head of the "Assistant intelligent" project of the PRST "Intelligence Logicielle"
- O. Mella and D. Fohr are involved in several European and National projects.
- K. Smaïli is a member of (Eurospeech, JEP) scientific commitee.

9.2. Distinctions

- Jean-Paul Haton is Professor at IUF (Institut Universitaire de France)
- Khalid Daoudi received the 2003 LORIA medal

9.3. The team members invited to give lectures

- K. Smaïli has been invited to give a talk at Information Processing Laboratory, Tokyo Japan
- J.P. Haton has been invited to give a talk at IBM Yorktown, USA

9.4. Invited lectures

- Sylvain Marchand, LaBRI, Université Bordeaux 1
- Dietmar Janetzko, Dept. of Cognitive Science University of Freiburg
- François Régnier, doctor and economist
- François Pellegrino, Laboratoire Dynamique Du Langage Université Lumière Lyon 2
- Jérôme Lang, IRIT Toulouse
- Stéphane Gérard, Directeur des systèmes d'information chez Presse+
- Yannick Esteve, LIA
- Jean ERCEAU, ONERA

9.5. Higher education

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

9.6. Participation to workshops and PhD thesis commitees

- Members of Phd thesis commitee D. Fohr, J.-P. Haton, M.-C. Haton, Y. Laprie, K. Smaïli;
- All the members of the team have been on workshops and gave talks (see next section).

10. Bibliography

Major publications by the team in recent years

[1] F. BIMBOT, M. EL-BÈZE, S. IGOUNET, M. JARDINO, K. SMAILI, I. ZITOUNI. *An alternative scheme for perplexity estimation and its assessment for the evaluation of language models.* in « Computer Speech and Language », number 1, volume 15, Jan, 2001, pages 1-13.

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

[3] C. CERISARA, D. FOHR. *Multi-band automatic speech recognition.* in « Computer Speech and Language », number 2, volume 15, April, 2001, pages 151-174.

[4] K. DAOUDI, D. FOHR, C. ANTOINE. *Dynamic Bayesian Networks for Multi-Band Automatic Speech Recognition.* in « Computer Speech and Language », volume 17, 2003, pages 263-285.

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

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

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

[8] D. LANGLOIS, A. BRUN, K. SMAÏLI, J.-P. HATON. *Événements impossibles en modélisation stochastique du langage.* in « Traitement Automatique des Langues », number 1, volume 44, Jul, 2003, pages 33-61.

[9] Y. LAPRIE, M.-O. BERGER. *Cooperation of Regularization and Speech Heuristics to Control Automatic Formant Tracking.* in « Speech Communication », number 4, volume 19, October, 1996, pages 23.

[10] I. ZITOUNI, K. SMAILI, J.-P. HATON. *Statistical Language Modeling Based on Variable-Length Sequences.* in « Computer Speech and Language », number 1, volume 17, Jan, 2003, pages 27-41.

Articles in referred journals and book chapters

[11] M. DEVIREN, K. DAOUDI. *Advances in Bayesian Networks.* Physica Verlag (Springer), 2003, chapter Continuous Speech Recognition Using Dynamic Bayesian Networks : A Fast Decoding Algorithm.

Publications in Conferences and Workshops

[12] V. BARREAUD, I. ILLINA, D. FOHR. *On-line compensation for non-stationary noise.* in « Automatic Speech Recognition and Understanding Workshop - ASRU'2003, St Thomas, US Virgin Islands », Nov, 2003.

[13] V. BARREAUD, I. ILLINA, D. FOHR. *Un algorithme de compensation de bruit en ligne synchrone a‘ la trame.* in « Journe'es de Jeunes Chercheurs en Parole », Sept, 2003.

[14] V. BARREAUD, I. ILLINA, D. FOHR. *On-Line Frame-Synchronous Compensation of Non-Stationary noise.* in « The 2003 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP 2003, Hong Kong, Chine », Apr, 2003.

[15] V. BARREAUD, I. ILLINA, D. FOHR. *On-Line Frame-Synchronous Noise Compensation.* in « The 15th International Congress of Phonetic Sciences - ICPhS 2003, Barcelone, Espagne », Aug, 2003.

[16] V. BARREAUD, I. ILLINA, D. FOHR, F. KORKMAZSKI. *Structural State-Based Frame Synchronous Compensation.* in « 8th European Conference on Speech Communication and Technology - Eurospeech'03, Gene‘ve, Suisse », Sept, 2003.

[17] Y. BENAYED, D. FOHR, J.-P. HATON, G. CHOLLET. *A New Keyword Spotting Approach Based on Reward Function.* in « Eventh International Symposium on Signal Processing and Its Applications - ISSPA'2003, Paris, France », Jul, 2003.

[18] Y. BENAYED, D. FOHR, J.-P. HATON, G. CHOLLET. *Confidence Measures for Keyword Spotting using Suport Vector Machines.* in « IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP'2003, Hong Kong, Chine », Apr, 2003.

[19] Y. BENAYED, D. FOHR, J.-P. HATON, G. CHOLLET. *Improving the Performance of a Keyword Spotting System by Using Support Vector Machines.* in « IEEE Automatic Speech Recognition and Understanding Workshop - ASRU'2003, St. Thomas, U.S. Virgin islands », Dec, 2003.

[20] A. BONNEAU, K. BALCI, Y. LAPRIE, V. COLOTTE. *Design and development of computer-assisted learning of prosody,.* in « Proceedings of the International Congress of Phonetic Sciences », 2003.

[21] A. BONNEAU, Y. LAPRIE. *Elitist identification of stops from formant transitions Proceedings of the International Congress of Ph onetic Sciences,.* in « Proceedings of the International Congress of Phonetic Sciences », 2003.

[22] A. BRUN, K. SMAI"LI, J. HATON. *Nouvelle approche de la se'lection de vocabulaire pour la de'tect ion de the‘me.* in « Traitement Automatique des Langues Naturelles (TALN2003) », pages 45-54, Nantes, France, 2003.

[23] C. CERISARA. *Towards Missing Data Recognition with Cepstral Features.* in « 8th European Conference on Speech Communication and Technology - EUROSPEECH'03, Geneva, Switzerland », Sep, 2003.

[24] C. CERISARA, I. ILLINA. *Robust speech recognition to non-stationary and unpredictable noise based*

on model-driven approaches. in « 8th European Conference on Speech Communication and Technology - EUROSPEECH'03, Geneva, Switzerland », Sep, 2003.

[25] K. DAOUDI, M. DEVIREN. *A new supervised-predictive compensation scheme for noisy speech recognition.* in « Proceedings of Eurospeech'03 », Geneva, Switzerland, September, 2003.

[26] M. DEVIREN, K. DAOUDI. *Frequency Filtering or Wavelet Filtering?.* in « Joint 13th Int. Conf. on Artificial Neural Networks and 10th Int. Conf. on Neural Information Processing (ICANN/ICONIP) », Istanbul, Turkey, June, 2003.

[27] L. GIRIN, S. MARCHAND, J. DI MARTINO, A. RO"BEL, G. PEETERS. *Comparing the Order of a Polynomial Phase Model for the Synthesis of Quasi-Harmonic Audio Signals.* in « Workshop on Applications of Signal Processing to Audio and Acoustics - WASPAA'03, Mohonk Mountain House, New Paltz, New York », Oct, 2003.

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

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

[30] S. JAMOUSSI, K. SMAILI, J.-P. HATON. *Vers la compréhension automatique de la parole : extraction des concepts par réseaux bayésiens.* in « Dixième Conférence en Traitement Automatique des Langues Naturelles - TALN'03, Batz-sur-Mer, France », Jun, 2003.

[31] D. LANGLOIS, K. SMAI"LI, J.-P. HATON. *Efficient linear combination for distant n-gram models.* in « 8th European Conference on Speech Communication and Technology - Eurospeech'03, Gene've, Suisse », volume 1, pages 409-412, Sep, 2003.

[32] F. LAURI, I. ILLINA, D. FOHR, F. KORKMAZSKI. *Using Genetic Algorithm for Rapid Speaker Adaptation.* in « 8th European Conference on Speech Communication and Technology - EUROSPEECH'03, Geneva, Switzerland », Sept, 2003.

[33] F. LAURI, I. ILLINA, D. FOHR. *Combining EigenVoices and Structural MLLR for Speaker Adaptation.* in « IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP'03, Hong Kong, China », Apr, 2003.

[34] N. PARLANGEAU-VALLÈS, J. FARINAS, D. FOHR, I. ILLINA, I. MAGRIN-CHAGNOLLEAU, O. MELLA, J. PINQUIER, J.-L. ROUAS, C. SÉNAC. *Audio Indexing on the Web : a Preliminary Study of Some Audio Descriptors.* in « 7th World Multiconference on Systematics, Cybernetics and Informatics - SCI'2003, Orlando, Florida, USA », Jul, 2003.

[35] J. RAZIK, C. SÉNAC, D. FOHR, O. MELLA, N. PARLANGEAU-VALLÈS. *Comparison of Two Speech/Music Segmentation Systems For Audio Indexing on the Web.* in « 7th World Multiconference on Systemics, Cybernetics and Informatics - SCI'2003, Orlando, Florida, USA », Jul, 2003.

Internal Reports

[36] N. Parlangeau-Vallès, I. Magrin-Chagnolleau, D. Fohr, I. Illina, O. Mella, K. Smai"li, C. Sénac, R. André-Obrecht, J. Farinas, J. Pinquier, J.-L. Rouas, F. Pellegrino. *Projet RAIVES (Recherche Automatique d'Informations Verbales Et Sonores) : vers l'extraction et la structuration de données radiophoniques sur Internet.* Rapport Intermédiaire, Nov, 2003.

Bibliography in notes

[37] A. Bonneau. *Identification of vocalic features from French stop bursts.* in « Journal of Phonetics », number 28, Dec, 2000, pages 495-502.

[38] A. Bonneau, L. Djezzar, Y. Laprie. *Perception of the Place of Articulation of French Stop Bursts.* in « Journal of the Acoustical Society of America », number 1, volume 100, Jul, 1996, pages 555-564.

[39] C. Cerisara, D. Fohr. *Fast Channel and Noise Compensation in the Spectral Domain.* in « XI European Signal Processing Conference - EUSIPCO 2002, Toulouse, France », September, 2002, http://www.loria.fr/publications/2002/A02-R-180/A02-R-180.ps.

[40] C. Cerisara, J.-C. Junqua, L. Rigazio. *Dynamic estimation of a noise over estimation factor for Jacobian-based adaptation.* in « IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2002, Orlando, Florida », IEEE, May, 2002, http://www.loria.fr/publications/2002/A02-R-179/A02-R-179.ps.

[41] V. Colotte, Y. Laprie. *Higher precision pitch marking for TD-PSOLA.* in « XI European Signal Processing Conference EUSIPCO, Toulouse, France », September, 2002.

[42] K. Daoudi, D. Fohr, C. Antoine. *Dynamic Bayesian Networks for Multi-Band Automatic Speech Recognition.* in « Computer Speech and Language », 2003, pages 263-285.

[43] A. de la Torre, D. Fohr, J.-P. Haton. *Statistical Adaptation of Acoustic Models to Noise Conditions for Robust Speech Recognition.* in « International Conference on Spoken Language Processing - ICSLP 2002, Denver, USA », pages 1437-1440, September, 2002.

[44] 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 », volume 1, pages 721, September, 2002.

[45] S. Maeda. *Un modèle articulatoire de la langue avec des composantes linéaires.* in « Actes 10èmes Journées d'Etude sur la Parole », pages 152-162, Grenoble, Mai, 1979.

[46] N. Parlangeau-Vallès, I. Magrin-Chagnolleau, D. Fohr, I. Illina, O. Mella, K. Smaïli, C. Sénac, J. Farinas, J. Pinquier, J.-L. Rouas, R. André-Obrecht, F. Pellegrino, D. Janiszek. *Projet RAIVES (Recherche Automatique d'Informations Verbales Et Sonores) vers l'extraction et la structuration de données radiophoniques sur Internet.* Rapport Intermédiaire, Dec, 2002.