IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2011

Project-Team PAROLE

Analysis, perception and recognition of speech

IN COLLABORATION WITH: Laboratoire lorrain de recherche en informatique et ses applications (LORIA)

RESEARCH CENTER
Nancy - Grand Est

THEME
Audio, Speech, and Language Processing

Table of contents

Project-Team PAROLE

Keywords: Natural Language, Speech, Recognition, Statistical Methods, Perception, Signal Processing

1. Members

Research Scientists

Yves Laprie [Team Leader, Senior Researcher, CNRS, HdR]
Anne Bonneau [Senior Researcher, CNRS]
Christophe Cerisara [Junior Researcher, CNRS, HdR]
Dominique Fohr [Senior Researcher, CNRS]
Denis Jouvet [Senior Researcher, INRIA, HdR]

Faculty Members

Vincent Colotte [Associate Professor, Henri Poincaré University]
Joseph di Martino [Associate Professor, Henri Poincaré University]
Jean-Paul Haton [Professor emeritus, Henri Poincaré University, Institut Universitaire de France]
Marie-Christine Haton [Professor emeritus, Henri Poincaré University, HdR]
Irina Illina [Associate Professor, I.U.T. Charlemagne, Nancy 2 University, HdR]
David Langlois [Associate Professor, IUFM, Henri Poincaré University]
Agnès Piquard-Kipffer [Associate Professor, IUFM, Henri Poincaré University]
Odile Mella [Associate Professor, Henri Poincaré University]
Slim Ouni [Associate Professor, I.U.T. Charlemagne, Nancy 2 University]
Kamel Smaïli [Professor, Nancy 2 University, HdR]

Technical Staff

Jean-François Grand [ADT JSnoori]
Larbi Mesbahi [Allegro, until October 2011]
Luiza Orosanu [Allegro, since October 2011]
Caroline Lavecchia [ANR Visac since November 2011]
Sébastien Demange [Emospeech]

PhD Students

Christian Gillot [MENRT grant, thesis to be defended in 2012]
Sylvain Raybaud [MENRT grant, thesis to be defended in 2012]
Othman Lachhab [COADVISE-FP7 program since November 2010]
Fadoua Bahja [COADVISE-FP7 program since May 2009]
Julie Busset [CNRS since 1st September 2009]
Utpala Musti [INRIA Cordi grant since 1st October 2009]
Arseniy Gorin [INRIA Cordi grant since 1st October 2011]

Post-Doctoral Fellows

Ingmar Steiner [since January 2011]
Asterios Toutios [until April 2011]

Administrative Assistant

Hélène Zganic [INRIA]

2. Overall Objectives

2.1. Overall Objectives

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

Our activities are structured in three topics:

- **Speech analysis and synthesis.** Our works are concerned with automatic extraction and perception of acoustic and visual cues, acoustic-to-articulatory inversion and speech synthesis. These themes give rise to a number of ongoing and future applications especially in the domain of foreign language learning.
- **Enriched automatic speech recognition.** Our works are concerned with stochastic models (HMM[1] and Bayesian networks), semi-supervised and smoothed training of these stochastic models, adaptation of a recognition system to important variabilities, and with enriching the output of speech recognition with higher-level information such as syntactic structure and punctuation marks. These topics give also rise to a number of ongoing and future applications: automatic transcription, speech/text alignment, audio indexing, keyword spotting, foreign language learning, dialog systems, vocal services...
- **Speech to Speech Translation and Langage Modeling.** This axis concerns statistical machine translation. The objective is to translate speech from a source language to any target language. The main activity of the group which is in charge of this axis is to propose an alternative method to the classical five IBM's models. This activity should conduct to several applications: e-mail speech to text, translation of movie subtitles.

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 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 an ANR project about audiovisual speech synthesis, another about acoustic-to-articulatory inversion of speech (ARTIS), another about the processing of articulatory data (DOCVACIM) and in a national evaluation campaign of automatic speech recognition systems (ESTER). We also coordinated until January 2009 the 6th PCRD project ASPI about acoustic-to-articulatory inversion of speech, and the Rapsodis ARC until october 2009. Additionally, we are also participating to a number of regional projects.

3. Scientific Foundations

3.1. Introduction

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 because of new applications like: **(i)** computer aided foreign language learning which requires both reliable automatic speech recognition and fine acoustic and articulatory speech analysis, **(ii)** automatic recognition of spontaneous speech which requires robustness against noise and speaker variability.

3.2. Speech Analysis and Synthesis

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

3.2.1. Oral comprehension

We developed tools to improve speech perception and production, and made perceptual experiments to prove their efficiency in language learning. These tools are also of interest for hearing impaired people, as well as for normally hearing people in noisy environments and also for children who learn to read (children who have language disabilities without cognitive deficit or hearing impairment and "normal" children).

3.2.1.1. Computer-assisted learning of prosody

We are studying automatic detection and correction of prosodic deviations made by a learner of a foreign language. 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 reference. The identification and correction tasks use speech analysis and modification tools developed in our team.

Within the framework of a new project (see 7.2.2), we also investigate the impact of a language intonational characteristics on the perception and production of the intonation of a foreign language.

3.2.1.2. Phonemic discrimination in language acquisition and language disabilities

We keep working on a project concerning identification of early predictors of reading, reading acquisition and language difficulties, more precisely in the field of specific developmental disabilities : dyslexia and dysphasia. A fair proportion of those children show a weakness in phonological skills, particularly in phonemic discrimination. However, the precise nature and the origin of the phonological deficits remain unspecified. In the field of dyslexia and normal acquisition of reading, our first goal was to contribute to identify early

indicators of the future reading level of children.We based our work on the longitudinal study - with 85 French children - of [55], [56] which indicates that phonemic discrimination at the beginning of kindergarten is strongly linked to sucess and specific failure in reading acquisition. We study now the link between oral discrimination both with oral comprehension and written compréhension. Our analyses are based on the follow up of a hundred chrildren for 4 years from kindergarten to end of grade 2 (from age 4 to age 8). Publications in progress.

3.2.1.3. Esophageal voices

It is possible for laryngectomees to learn a substitution voice: the esophageal voice. This voice is far from being natural. It is characterized by a weak intensity, a background noise that bothers listening, and a low pitch frequency. A device that would convert an esophageal voice to a natural voice would be very useful for laryngectomees because it would be possible for them to communicate more easily. Such natural voice restitution techniques would ideally be implemented in a portable device.

3.2.2. 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 [63].

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.

3.2.3. 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 [51] 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 [44] advocates for the theory of expansion movements: the movement tends to be anticipated when no phonological constraint is imposed on labiality. Cohen and Massaro [49] proposed dominance functions that require a substantial numerical training.

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

3.2.4. Speech Synthesis

Data-driven speech synthesis is widely adopted to develop Text-to-Speech (TTS) synthesis systems. Basically, it consists of concatenating pieces of signal (units) selected from a pre-recorded sentence corpus. Our ongoing work on acoustic TTS was recently extended to study acoustic-visual speech synthesis (bimodal units).

3.2.4.1. Text-to-speech synthesis

Data-driven text-to-speech synthesis is usually composed of three steps to transform a text in speech signal. The first step is Natural Language Processing (NLP) which tags and analyzes the input text to obtain a set of features (phoneme sequence, word grammar categories, syllables...). It ends with a prosodic model which transforms these features into acoustic or symbolic features (F0, intensity, tones...). The second step uses a Viterbi algorithm to select units from a corpus recorded beforehand, which have the closest features to the prosodic features expected. The last step amounts to concatenate these units.

Such systems usually generate a speech signal with a high intelligibility and a naturalness far better than that achieved by old systems. However, building such a system is not an easy task [48] and the global quality mainly relies on the quality of the corpus and prosodic model. The prosodic model generally provides a good standard prosody, but, the generated speech can suffer from a lack of variability. Especially during the synthesis of extended passages, repetition of similar prosodic patterns can lead to a monotonous effect. Therefore, to avoid this problem due to the projection of linguistic features onto symbolic or acoustic dimensions (during NLP), we [50] proposed to perform the unit selection directly from linguistic features without incorporating any prosodic information. To compensate the lack of prosodic prediction, the selection needs to be performed with numerous linguistic features. The selection is no longer restrained by a prosodic model but only driven by weighted features. The consequence is that the quality of synthesis may drop in crucial instants. Our works deal to overcome this new problem while keeping advantage of the lack of prosodic model.

These works have an impact on the construction of corpus and on the NLP engine which needs to provide as much information as possible to the selection step. For instance, we introduced a chunker (shallow parser) to give us information on a potential rhythmic structure. Moreover, to perform the selection, we developed an algorithm to automatically weight the linguistic features given by the NLP. Our method relies on acoustic clustering and entropy information [50]. The originality of our approach leads us to design a more flexible unit selection step, constrained but not restrained.

3.2.4.2. Acoustic-visual speech synthesis

Audiovisual speech synthesis can be achieved using 3D features of the human face supervised by a model of speech articulation and face animation. Coarticulation is approximated by numerical models that describe the synergy of the different articulators. Acoustic signal is usually synthetic or natural speech synchronized with the animation of the face. Some of the audiovisual speech systems are inspired by recent development in speech synthesis based on samples and concatenative techniques. The main idea is to concatenate segments of recorded speech data to produce new segments. Data can be video or motion capture. The main drawback of these methods is that they focus on one field, either acoustic or visual. But (acoustic) speech is actually generated by moving articulators, which modify the speaker's face. Thus, it is natural to find out that acoustic and face movements are correlated. A key point is therefore to guarantee the internal consistency of the acoustic-visual signal so that the redundancy of these two signals acknowledged as a determining perceptive factor, can really be exploited by listeners. It is thus important to deal with the two signals (acoustic and visual) simultaneously and to keep this link during the whole process. This is why we make the distinction between audiovisual speech synthesis (where acoustic is simply synchronized with animation) and acoustic-visual speech where speech is considered as a bimodal signal (acoustic and visual) as considered in our work. Our long-term goal is to contribute to the fields of acoustic speech synthesis and audiovisual speech synthesis by building a bimodal corpus and developing an acoustic-visual speech synthesis system using bimodal unit concatenation.

3.3. Automatic speech recognition

Automatic speech recognition aims at reproducing the cognitive ability of humans to recognize and understand oral speech. Our team has been working on automatic speech recognition for decades. We began with knowledge-based recognition systems and progressively made our research works evolve towards stochastic approaches, both for acoustic and language models. Regarding acoustic models, we have especially investigated HMM (Hidden Markov Models), STM (Stochastic Trajectory Models), multi-band approach and BN (Bayesian Networks). Regarding language models, our main interest has concerned ngram approaches (word classes, trigger, impossible ngram, etc).

The main challenge of automatic speech recognition is its robustness to multiple sources of speech variability [54]. Among them, we have been focusing on acoustic environment, inter- and intra-speaker variability, different speaking styles (prepared speech, spontaneous, etc.) and non-native pronunciations.

Another specifity of automatic speech recognition is the necessity to combine efficiently all the research works (in acoustic modeling, langage modeling, speaker adaptation, etc.) into a core platform in order to evaluate them, and to go beyond pure textual transcriptions by enriching them with punctuation, syntax, etc., in order to make them exploitable by both humans and machines.

3.3.1. Acoustic features and models

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 explored alternative parameterizations to support some of our recent research progresses. For example, prosodic features such as intonation curves and vocal energy give important cues to recognize dialog acts, and more generally to compute information that relates to supra-phonemic (linguistic, dialog, ...) characteristics of speech. Prosodic features are developed jointly for both the Speech Analysis and Speech Recognition topics. We also developed a new robust front-end, which is based on wavelet-decomposition of the speech signal.

Concerning 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). In fact, HMMs are mainly used to model the acoustic units to be recognized (usually triphones) in all of our recognition engines (ESPERE, ANTS...). Besides,we have investigated Bayesian Networks (BN) to explicitly represent random variables and their independence relationships to improve noise robustness.

3.3.2. Robustness and invariance

Part of our research activities about ASR aims at improving the robustness of recognizers to the different sources of variability 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 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 (such as missing data theory). The following state-of-the-art approaches thus form our baseline set of technologies: MLLR (Maximum Likelihood Linear Regression), MAP (Maximum A Posteriori), PMC (Parallel Model Combination), CMN (Cepstral Mean Normalization), SAT (Speaker Adaptive Training), HLDA (Heteroscedastic Linear Discriminant Analysis), Spectral Subtraction and Jacobian Adaptation.

These technologies constitute the foundations of our recent developments in this area, such as non-native speaker adaptation, out-of-vocabulary words detection and adaptation to pronunciation variations. Handling speech variabilities may also benefit from exploiting additional external or contextual sources of information to more tightly guide the speech decoding process. This is typically the role of the language model, which shall in this context be augmented with higher-level knowledge, such as syntactic or semantic cues. Yet, automatically extracting such advanced features is very challenging, especially on imperfect transcribed speech.

The performance of automatic speech recognition (ASR) systems drastically drops when confronted with non-native speech. If we want to build an ASR system that takes into account non-native speech, we need to modify the system because, usually, ASR systems are trained on standard phone pronunciations and designed to recognize only native speech. In this way, three method categories can be applied: acoustic model transformation, pronunciation modeling and language modeling. Our contribution concerns the first two methods.

3.3.3. 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 requires to extract discriminant acoustic parameters. Our contribution concerns the MFCC and wavelet parameters. Another point is to find a good classifier. Various classifiers are commonly used: k-Nearest-Neighbors, Hidden Markov Models, Gaussian Mixture Models, Artificial Neural Networks.

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

3.3.4. Speech/text alignment

Speech/text alignment consists in finding time boundaries of words or phones in the audio signal knowing the orthographic transcription. The main applications of speech/text alignment are training of acoustic models, segmentation of audio corpus for building units for speech synthesis or segmentation of the sentence uttered by a learner of a foreign language. Moreover, speech/text alignement is a useful tool for linguistic researchers.

Speech/text alignment requires two steps. The first step generates the potential pronunciations of the sentence dealing with multiple pronunciations of proper nouns, liaisons, phone deletions, and assimilations. For that, the phonetizer is based on a phonetic lexicon, and either phonological rules or an automatic classifier as a decision tree. The second step finds the best pronunciation corresponding to the audio signal using acoustic HMM models and an alignment algorithm. The speech team has been working on this domain for a long time.

3.4. Speech to Speech Translation and Langage Modeling

Speech-to-Speech Translation aims at translating a source speech signal into a target speech signal. A sequential way to adress this problem is to first translate a text to another one. And after, we can connect a speech recognition system at the input and a text to speech synthesis system at the output. Several ways to adress this issue exist. The concept used in our group is to let the computer learning from a parallel text all the associations between source and target units. A unit could be a word or a phrase. In the early 1990s [47] proposes five statistical translation models which became inescapable in our community. The basic idea of the model 1 is to consider that any word of the target language could be a potential translation of any source word. The problem is then to estimate the distribution probability of a target word given a source one. The translation problem is similar to the speech recognition one. Indeed, we have to seek the best foreign sentence given a source one. This one is obtained by decoding a lattice translation in which a language and translation models are used. Several issues have to be supported in machine translation as described below.

3.4.1. Word translation

The first translation systems identify one-to-one associations between words of target and source languages. This is still necessary in the present machine translation systems. In our group we develop a new concept to learn the translation table. This approach is based on computing all the inter-lingual triggers inside a parallel corpus. This leads to a pertinent translation table [62]. Obviously, this is not sufficient in order to make a realistic translation because, with this approach, one word is always translated into one word. In fact, it is possible to express the same idea in two languages by using different numbers of words. Thus, a more general one-to-one alignement has to be achieved.

3.4.2. Phrase translation

The human translation is a very complex process which is not only word-based. A number of research groups developed phrase-based systems which are different from the baseline IBM's model in training. These methods, deals with linguistic units which consists in more than one word. The model supporting phrase-based machine translation uses reordering concept and additional feature functions. In order to retrieve phrases, several approaches have been proposed in the litterature. Most of them require word-based alignments. For example, Och and al. [65] collected all phrase pairs that were consistent with the word alignment provided by Brown's models.
We developed a phrase based algorithm which is based on finding first an adequate list of phrases. Then, we find out the best corresponding translations by using our concept of inter-lingual triggers. A list of the best translations of sequences is then selected by using simulated annealing algorithm.

3.4.3. Language model

A language model has an important role in a statistical machine translation. It ensures that the translated words constitute a valid linguistic sentence. Most of the community uses n-grams models, that is what we do also.

3.4.4. Decoding

The translation issue is treated as an optimization problem. Translating a sentence from English into a Foreign language involves finding the best Foreign target sentence f^* which maximizes the probability of f given the English source sentence e. The Bayes rule allows to formulate the probability $P(f|e)$ as follows:

$$f^* = \arg\max_f P(f|e) = \arg\max_f P(e|f)P(f)$$

The international community uses either PHARAOH [58] or MOSES [57] based on a beam search algorithm. In our group we started decoding by PHARAOH but we moved recently to MOSES.

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 [53] 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 (for instance in the framework of 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. Furthermore, 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, keyword spotting and automatic categorization of mails.

5. Software

5.1. WinSnoori

contact : Yves Laprie (Yves.Laprie@loria.fr)

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

The main improvement concerns automatic formant tracking which is now available with other tools for copy synthesis. It is now possible to determine parameters for the formant synthesizer of Klatt quite automatically. The first step is formant tracking, then the determination of F0 parameters and finally the adjustment of formant amplitudes for the parallel branch of the Klatt synthesizer enable a synthetic speech signal to be generated. The automatic formant tracking that has been implemented is an improved version of the concurrent curve formant tracking [60]. 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 years, formant tracking enables copy synthesis. The current version of WinSnoori is available on http://www.winsnoori.fr.

5.2. SUBWEB

contacts : David Langlois (langlois@loria.fr) and Kamel Smaïli (smaili@loria.fr).

We published in 2007 a method which allows to align sub-titles comparable copora [61]. In 2009, we proposed an alignment web tool based on the developed algorithm. It allows to: upload a source and a target files, obtain an alignment at a sub-title level with a verbose option, and and a graphical representation of the course of the algorithm. This work has been supported by CPER/TALC/SUBWEB [2].

5.3. SELORIA

contact : Odile Mella (Odile.Mella@loria.fr).

SELORIA is a toolbox for speaker diarization.

The system contains the following steps:

- Speaker change detection: to find points in the audio stream which are candidates for speaker change points, a distance is computed between two Gaussian modeling data of two adjacent given-length windows. By sliding both windows on the whole audio stream, a distance curve is obtained. A peak in this curve is thus considered as a speaker change point.
- Segment recombination: too many speaker turn points detected during the previous step results in a lot of false alarms. A segment recombination using BIC is needed to recombine adjacent segments uttered by the same speaker.
- Speaker clustering: in this step, speech segments of the same speaker are clustered. Top-down clustering techniques or bottom-up hierarchical clustering techniques using BIC can be used.
- Viterbi re-segmentation: the previous clustering step provides enough data for every speaker to estimate multi-gaussian speaker models. These models are used by a Viterbi algorithm to refine the boundaries between speakers.
- Second speaker clustering step (called cluster recombination): This step uses Universal Background Models (UBM) and the Normalized Cross Likelihood Ratio (NCLR) measure.

This toolbox is derived from mClust designed by LIUM.

5.4. ANTS

contact : Dominique Fohr (fohr@loria.fr).

The aim of the Automatic News Transcription System (ANTS) is to transcribe radio broadcast news. ANTS is composed of five stages: broad-band/narrow-band speech segmentation, speech/music classification, speaker segmentation and clustering, detection of silences/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.

A real time version of ANTS has been developped. The transcription is done in real time on a quad-core PC.

5.5. JSafran

Contact : Christophe Cerisara (Christophe.Cerisara@loria.fr).

[2] http://wikitalc.loria.fr/dokuwiki/doku.php?id=operations:subweb

J-Safran is the "Java Syntaxico-semantic French Analyser". Its development has started in june 2009 from the collaboration between Parole and Talaris in the context of the RAPSODIS project. It is an open-source dependency parsing platform that is dedicated to oral speech. Its main interesting features, as compared to other similar software, are:

- It is designed for both manual and semi-automatic edition of dependency graphs, as well as for fully automatic parsing. To this end, it integrates two of the best state-of-the-art automatic parsers of the litterature, the Malt Parser and the MATE parser, as well as a third experimental Maximum Entropy Markov Model-based parser developed from November 2011 in the team. It further integrates three automatic Part-of-speech taggers: the TreeTagger, the OpenNLP and MATE taggers.
- It is smoothly interfaced with the JTrans platform, thus enabling the user to directly listen to the aligned speech segments when annotating, which is an important added value to help disambiguation. The interface between both software goes well beyond simple method calls, as they both share for instance parts of the tokenization process and access a common immutable text source from the disk or on the Web.
- It supports multi-layer annotations, such as dependency relations, semantic role labeling, named entities and coreference links for instance, as well as inter-layer projection facilites.
- It offers a powerful rule-based search and tree manipulation language to transform for instance the annotation schema of a large corpus with a few commands only.
- As it is written in pure Java, it can run on any modern computer, either as a standalone application or embedded in a web page.

A description of JSafran is published in [16]. JSafran is distributed under the Cecill-C licence, and can be downloaded at http://synalp.loria.fr/?n=Research.Software

5.6. JTrans

Contact : Christophe Cerisara (Christophe.Cerisara@loria.fr).

JTrans is an open-source software for semi-automatic alignement of speech and textual corpus. It is written 100% in JAVA and exploits libraries developed since several years in our team. Two algorithms are available for automatic alignment: a block-viterbi and standard forced-alignement Viterbi. The latter is used when manual anchors are defined, while the former is used for long audio files that do not fit in memory. It is designed to be intuitive and easy to use, with a focus on GUI design. The rationale behind JTrans is to let the user control and check on-the-fly the automatic alignment algorithms. It is bundled for now with a French phonetic lexicon and French models.

Recent improvments include its integration within the JSafran platform and its release as a Java applet that can be demonstrated on web pages. During the last three months, JTrans has been downloaded about 120 times and seven users of JTrans, outside LORIA, have directly contacted the team for requests about JTrans.

JTrans is developed in the context of the CPER MISN TALC project, in collaboration between the Parole and Talaris INRIA teams, and CNRS researchers from the ATILF laboratory. It is distributed under the Cecill-C licence, and can be downloaded at http://synalp.loria.fr/?n=Research.Software

5.7. STARAP

contact : Dominique Fohr (fohr@loria.fr).

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.

This toolkit was realised in the framework of the STORECO contract and the formats of the input and output files are compatible with HTK toolkit.

5.8. TTS SoJA

contact : Vincent Colotte (Vincent.Colotte@loria.fr).

TTS SoJA (Speech synthesis platform in Java) is a software of text-to-speech synthesis system. The aim of this software is to provide a toolkit to test some steps of natural language processing and to provide a whole system of TTS based on non uniform unit selection algorithm. The software performs all steps from text to the speech signal. Moreover, it provides a set of tools to elaborate a corpus for a TTS system (transcription alignment, ...). Currently, the corpus contains 1800 sentences (about 3 hours of speech) recorded by a female speaker.

Most of the modules are developed in Java. Some modules are in C. The platform is designed to make easy the addition of new modules. The software runs under Windows and Linux (tested on Mandriva, Ubuntu). It can be launch with a graphical user interface or directly integrated in a Java code or by following the client-server paradigm.

The software license should easily allow associations of impaired people to use the software. A demo web site has been built: http://soja-tts.loria.fr

5.9. Corpus Recorder

contact : Vincent Colotte (Vincent.Colotte@loria.fr).

Corpus Recorder is a software for the recording of audio corpora. It provides a easy tool to record with a microphone. The gain of the audio input is controlled during the recording. From a list of sentences, the output is a set of wav files automatically renamed with textual information given in input (nationality, speaker language, gender...). An easy syntactic tagging allows to display a textual context of the sentence to pronounce. This software is suitable for recording sentences with information to guide the speaker.

The software is developed in Tcl/Tk (tested under Windows and Linux). It was used for the recording of sentences for the TTS system SOJA and during the Intonale Project (Prosody Modeling).

6. New Results

6.1. Speech Analysis and Synthesis

Participants: Anne Bonneau, Vincent Colotte, Dominique Fohr, Yves Laprie, Joseph di Martino, Slim Ouni, Asterios Toutios, Sébastien Demange, Fadoua Bahja, Agnès Piquard-Kipffer, Utpala Musti.

6.1.1. Acoustic-to-articulatory inversion

6.1.1.1. Building new articulatory models

The possibility of generating the same sounds as those uttered by the speaker (or at least vocal tract transfer functions not too far from those observed) via the articulatory model and the acoustic simulation constitutes the underlying hypothesis of an analysis by synthesis method of acoustic-to-articulatory inversion. The articulatory model, and consequently its construction, thus plays a crucial role in inversion. An geometrical adaptation procedure has been developed in order to account for new speakers [28], [29]. It uses two scaling factors, one for the mouth cavity and the second for the pharyngeal cavity. In addition the model can be rotated globally and a second rotation controls the relative position of the pharynx with respect to the mouth cavity. In order to ensure a smooth transition from the mouth cavity to the pharynx cavity the angle of the rotation is a function of the distance with respect to the mouth axis.

The adaptation and model have been tested by using the X-ray data used by Maeda to construct his model. It should be noted that there are very few X-ray data with articulatory contour information available. These data correspond to a female speaker. The RMS reconstruction error reached by the adapted articulatory model is 0.550 mm what is very good for this particular speaker. Other data will be used in the future to validate the model and the adaptation procedure as soon as the contours will be delineated. An anatomical adpation procedure will also be developed in the future.

6.1.1.2. Determination of the vocal tract centerline

The connection of the articulatory model with the acoustic simulation requires the area function to be decomposed into elementary uniform tubes. The decomposition should respect the plane wave propagation. For that purpose the central line of the vocal tract has to be determined. The quality of the centerline stongly influences the closeness between natural and artificial formant frequencies.

We designed two complementary algorithms. The first exploits a dynamic programming approach to select points on interior and exterior walls of the vocal tract which minimize a global criterion combining the length of the centerline and the angle between the normal to the segments linking the points selected on both walls and the centerline [29]. It turned out that this first algorithm provides an insufficient smoothness of the centerline. A second algorithm has been designed by using an active curve which maximizes the smoothness of the centerline and the distance from any point of the centerline with exterior and interior walls. This second algorithm provides very good results.

6.1.1.3. Adaptation of cepstral coefficients for inversion

The inversion of speech requires spectra of natural speech to be compared with spectra synthesized via the articulatory synthesizer. This comparison cannot be carried out directly because the source is not taken into account in the synthetic spectra. Last year we thus investigated an affine adaptation of all the cepstral coefficients. This adaptation brings the spectral peaks of natural and synthetic spectra closer but at the same time tends to flatten the spectra. Moreover, it also appears that adaptation of only the very first cepstral coefficients (the first two except C_0 which represents energy) were sufficient to capture the spectral tilt. Since it is important to keep clear spectral peaks to explore the articulatory space, we used the bilinear transform in order to bring the two spectra closer [15]. The results are now better and the bilinear transform will be used to recover inverse solutions.

6.1.1.4. Acoustic-to-articulatory inversion using a generative episodic memory

We have developed an episodic based inversion method. Episodic modeling is interesting for two reasons. First, it does not rely on any assumption about the mapping relationship between acoustic and articulatory, but rather it relies on real synchronized acoustic and articulatory data streams. Second, the memory structurally embeds the naturalness of the articulatory dynamics as speech segments (called episodes) instead of single observations as for the codebook based methods. Estimating the unknown articulatory trajectories from a particular acoustic signal, with an episodic memory, consists in finding the sequence of episodes, which acoustically best explains the input acoustic signal. We refer to such a memory as a concatenative memory (C-Mem) as the result is always expressed as a concatenation of episodes. Actually a C-Mem lacks from generalization capabilities as it contains only several examples of a given phoneme and fails to invert an acoustic signal, which is not similar to the ones it contains. However, if we look within each episode we can find local similarities between them. We proposed to take advantage of these local similarities to build a generative episodic memory (G-Mem) by creating inter-episodes transitions. The proposed G-Mem allows switching between episodes during the inversion according to their local similarities. Care is taken when building the G-Mem and specifically when defining the inter-episodes transitions in order to preserve the naturalness of the generated trajectories. Thus, contrary to a C-Mem the G-Mem is able to produce totally unseen trajectories according to the input acoustic signal and thus offers generalization capabilities. The method was implemented and evaluated on the MOCHA corpus, and on a corpus that we recorded using an AG500 articulograph. The results showed the effectiveness of the proposed G-Mem which significantly outperformed standard codebook and C-Mem based approaches. Moreover similar performances to those reported in the literature with recently proposed methods (mainly parametric) were reached. [18]

The paradigm of episodic memories was also used for speech recognition. We do not extend the acoustic feature with any explicit articulatory measurements but instead we used the articulatory-acoustic generative episodic memories (G-mem). The proposed recognizer is made of different memories each specialized for a particular articulator. As all the articulators do not contribute equally to the realization of a particular phoneme, the specialized memories do not perform equally regarding each phoneme. We showed, through phone string recognition experiments that combining the recognition hypotheses resulting from the different articulatory specialized memories leads to significant recognition improvements. [19].

6.1.2. Using Articulography for Speech production

Since we have an articulograph (AG500, Carstens Medizinelektronik) available, we can easily acquire articulatory data required to study speech production. The articulograph is used to record the movement of the tongue (this technique is called electromagnetography - EMA). The AG500 has a very good time resolution (200Hz), which allows capturing all articulatory dynamics. The articulograph was used in a study about inversion (see the previous section) and to investigate pharyngealization.

Pharyngealized phonemes are commonly described as having the same place of articulation (dental) as their non-pharyngealized counterparts, but differ by the presence of a secondary articulation involving mainly the back of the tongue.

To study pharyngealized phonemes in Arabic from an articulatory point of view, our articulograph was used to record the movement of the tongue. Although EMA is not known as an optimal technique to cover the back of the tongue, good placement of the sensors and good interpretation of their positions can help to define pharyngealization relevantly. In fact, it is important to set one sensor as far as possible on the tongue (in our case, at 7cm from the tongue tip).

A corpus of several CVCVCVs was recorded using this articulograph, then phonetically labeled, and analyzed. The main finding of this work is that the coarticulation effect of the pharyngealized phonemes extends the immediate surrounding phonemes to influence the phonemes up to four-phoneme distance from the pharyngealized phoneme. The pharyngealization affects indifferently the previous and the following vowels and consonants.

We also investigated the effect of pharyngealization in Modern Standard Arabic (MSA) and Dialectal Arabic (DA). The acoustic material was more important than EMA. Although, we studied one speaker for EMA, the obtained results are encouraging to record more arabic speakers. [42]

6.1.3. Labial coarticulation

Results show that protrusion is a fragile cue to the rounding feature. Although we observe for each speaker a clear (but not large) separation between vowels /i/ and /y/ produced in isolation, many realizations of /i/ and /y/ come very close together and even overlap in few cases for vowels in contexts. The efficiency of the parameter depends on speakers and contexts. The distance between the corners is probably the most fragile cue to vowel roundedness. Many overlapping areas are observed for vowels in context. This is not good news for speech specialists since this parameter is easy to measure (with cameras and markers painted on the speaker's face) and its evaluation can be fully automatic. Each of the three lip opening parameters constitutes a very efficient cue to the rounding feature. For vertical opening, the opposition between /i/ and /y/ in initial position appears to be endangered in bilabial context, due to the anticipation of lip closing during /i/. Nevertheless, the temporal variations of lip opening during the initial /i/ are very important, and more analyses, taking into account these variations, will be necessary to analyse /i/ vs. /y/ phonetic distinction more thoroughly.

6.1.4. Speech synthesis

Visual data acquisition was performed simultaneously with acoustic data recording, using an improved version of a low-cost 3D facial data acquisition infrastructure. The system uses two fast monochrome cameras, a PC, and painted markers, and provides a sufficiently fast acquisition rate to enable an efficient temporal tracking of 3D points. The recorded corpus consisted of the 3D positions of 252 markers covering the whole face. The lower part of the face was covered by 70% of all the markers (178 markers), where 52 markers were covering

only the lips so as to enable a fine lip modeling. The corpus was made of 319 medium-sized French sentences uttered by a native male speaker and corresponding to about 25 minutes of speech,.

We designed a first version of the text to acoustic-visual speech synthesis based on this corpus. The system uses bimodal diphones (an acoustic component and a visual one) and unit selection techniques (see 3.2.4). We have introduced visual features in the selection step of the TTS process. The result of the selection is the path in the lattice of candidates found in the Viterbi algorithm, which minimizes a weighted linear combination of three costs: the target cost, the acoustic joined cost, and the visual joined cost.

Finding the best set of weights is a difficult problem by itself mainly because of their highly different nature (linguistic, acoustic, and visual considerations). This year, we added the first derivative of the visual trajectories in the visual join cost and we developed a method to determine automatically the weights applied to each cost, using a series of metrics that assess quantitatively the performance of synthesis [37].

This year, more progress have been made regarding the definition of the target cost. Now, The target cost includes both acoustic target cost and visual target cost.

The visual target cost includes visual and articulatory information. We implemented and evaluated two techniques [32]: (1) Phonetic category modification, where the purpose was to change the current characteristics of some phonemes which were based on phonetic knowledge. The changes modified the target and candidate description for the target cost to better take into account their main characteristics as observed in the audio-visual corpus. The expectation was that their synthesized visual speech component would be more similar to the real visual speech after the changes. (2) Continuous visual target cost, where the visual target cost component is now considered as real value, and thus continuous, based on the articulatory feature statistics.

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

6.1.5.1. Phonemic segmentation in reading and reading-related skills acquisition in dyslexic children and adolescents

Our computerized tool EVALEC was published [67] after the study of reading level and reading related skills of 400 hundred children from grade 1 to grade 4 (from age 6 to age 10) [69]. This research was supported by a grant from the French Ministery of Health (Contrat 17-02-001, 2002-2005). This first compurerized battery of tests in french language assessing reading and related skills (phonemic segmentation, phonological short term memory) comparing results both to chronological age controls and reading level age control in order to diagnostic Dyslexia. Both processing speed and accuracy scores are taken into account. This battery of tests is used by speech and langage therapists. We keep on examining the reliability (group study) and the prevalence (multiple case study) of 15 dyslexics' phonological deficits in reading and reading related skills in comparaison with a hundred reading level children [68], and by the mean of longitudinal studies of children from age 5 to age 17 [66]. This year, we started the development of a project which examined multimodal speech both with SLI, dyslexics and control children (30 children). Our goal is to examine visual contribution to speech perception accross differents experiments with a natural face (syllables with several conditions). Our goal is to search what can improve intelligibility in children who have sévère langague acquisition difficulties.

6.1.5.2. Langage acquisition and langage disabilities (deaf chidren, dysphasic children)

Providing help for improving french language acquisition for hard of hearing (HOH) children or for children with language disabilities was one of our goal : ADT (Action of Technological Developpement) Handicom [piquardkipffer:2010:inria-00545856:2]. The originality of this project was to combine psycholinguistical and speech analyses researchs. New ways to learn to speak/read were developed. A collection of three digital books has been written by Agnès Piquard-Kipffer for both 2-6, 5-9, 8-12 year old children (kindergarten, 1-4th grade) to train speaking and reading acquisition regarding their relationship with speech perception and audio-visual speech perception. A web interface has been created (using Symfony and AJAX technologies) in order to create others books for language impaired children. A workflow which transforms a text and an audio source in a video of digital head has been developed. This worklow includes an automatic speech alignment, a phonetic transcription, a speech synthetizer, a French cued speech coding and speaking digital head. A series of studies (simple cases studies, 5 deaf children and 5 SLI children and group studies with 2 kindergarten

classes) were proposed to investigate the linguistical, audio-visual processing.... presumed to contribute to language acquisition in deaf children. Publication are submitted.

6.1.6. Enhancement of esophageal voice

6.1.6.1. Detection of F0 in real-time for audio: application to pathological voices

The work first rested on the CATE algorithm developed by Joseph Di Martino and Yves Laprie, in Nancy, 1999.The CATE (Circular Autocorrelation of the Temporal Excitation) algorithm is based on the computation of the autocorrelation of the temporal excitation signal which is extracted from the speech log-spectrum. We tested the performance of the parameters using the Bagshaw database, which is constituted of fifty sentences, pronounced by a male and a female speaker. The reference signal is recorded simultaneously with a microphone and a laryngograph in an acoustically isolated room. These data are used for the calculation of the contour of the pitch reference. When the new optimal parameters from the CATE algorithm were calculated, we carried out statistical tests with the C functions provided by Paul BAGSHAW. The results obtained were very satisfactory and a first publication relative to this work was accepted and presented at the ISIVC 2010 conference [46]. At the same time, we improved the voiced / unvoiced decision by using a clever majority vote algorithm electing the actual F0 index candidate. A second publication describing this new result was published at the ISCIT 2010 conference [45].

6.1.6.2. Voice conversion techniques applied to pathological voice repair

Voice conversion is a technique that modifies a source speaker's speech to be perceived as if a target speaker had spoken it. One of the most commonly used techniques is the conversion by GMM (Gaussian Mixture Model). This model, proposed by Stylianou, allows for efficient statistical modeling of the acoustic space of a speaker. Let "x" be a sequence of vectors characterizing a spectral sentence pronounced by the source speaker and "y" be a sequence of vectors describing the same sentence pronounced by the target speaker. The goal is to estimate a function F that can transform each source vector as nearest as possible of the corresponding target vector. In the literature, two methods using GMM models have been developed: In the first method (Stylianou), the GMM parameters are determined by minimizing a mean squared distance between the transformed vectors and target vectors. In the second method (Kain), source and target vectors are combined in a single vector "z". Then, the joint distribution parameters of source and target speakers is estimated using the EM optimization technique. Contrary to these two well known techniques, the transform function F, in our laboratory, is statistically computed directly from the data: no needs of EM or LSM techniques are necessary. On the other hand, F is refined by an iterative process. The consequence of this strategy is that the estimation of F is robust and is obtained in a reasonable lapse of time. This interesting result was published and presented at the ISIVC 2010 conference [70].

6.1.7. Perception and production of prosodic contours in L1 and L2

6.1.7.1. Language learning (feedback on prosody)

Feedback on L2 prosody based upon visual displays, speech modifications and automatic diagnosis has been elaborated and a pilot experiment undertaken to test its immediate impact on listeners. Results show that the various kinds of feedback provided by the system enable French learners with a low production level to improve their realisations of English lexical accents more than (simple) auditory feedback. These results shoud be confirmed with a large number of speakers but based upon the important differences between results obtained for speakers in test and control conditions, we are confident in the interest of the system presented here [41]. In particular, the system analyses learners' realisations and provide indications on what they should correct, a guidance which is considered as necessary by specialists in the oral aspects of language learning.

6.1.7.2. Production of prosody contour

We report here relevant observations for the study continuation in French. These observations were obtained in an ongoing project about non-conclusive prosodic patterns in French and English ("Intonale" project 7.2.2). We specifically discuss slope variations, estimated in semitones, concerning two kinds of non-conclusive configurations, which are inside a clause, or at the end of a clause, respectively : (i) the final segment of a subject NP in an assertive sentence, followed or not by another syntagm ended by a continuation contour (ii)

the final segment of a A clause, in a two clause utterance AB, where A and B are assertive clauses connected by an discourse relation, marked or not with a conjunction.

Intonation slopes are computed as regression slopes using F0 values in semitones estimated every 10 ms. Slopes are calculated on the two last syllables of the target segments of every sentence. Results show that slopes for segments which are not at the end of a clause, and segments at the end of a clause followed by a conjunction are typically rising, and not significantly different the ones from the others. On the contrary, slopes for ends of clauses not followed by a conjunction are significantly different from the previous ones. More than 50We are presently studying English sentences, in particular continuation contours, produced by French speakers, in order to determine the impact of their native language (French) on their English pronounciations.

6.1.8. Pitch detection

Over the last two years, we have proposed two new real time pitch detection algorithms (PDAs) based on the circular autocorrelation of the glottal excitation, weighted by temporal functions, derived from the CATE [64] original algorithm (Circular Autocorrelation of the Temporal Excitation), proposed initially by J. Di Martino and Y. Laprie. In fact, this latter algorithm is not constructively real time because it uses a post-processing technique for the Voiced/Unvoiced (V/UV) decision. The first algorithm we developed is the eCATE algorithm (enhanced CATE) that uses a simple V/UV decision less robust than the one proposed later in the eCATE+ algorithm.

We propose a recent modified version called the eCATE++ algorithm which focuses especially on the detection of the F0, the tracking of the pitch and the voicing decision in real time. The objective of the eCATE++ algorithm consists in providing low classification errors in order to obtain a perfect alignment with the pitch contours extracted from the Bagshaw database by using robust voicing decision methods. The main improvement obtained in this study concerns the voicing decision, and we show that we reach good results for the two corpora of the Bagshaw database.

6.2. Automatic Speech Recognition

Participants: Christophe Cerisara, Sébastien Demange, Dominique Fohr, Christian Gillot, Jean-Paul Haton, Irina Illina, Denis Jouvet, Odile Mella, Luiza Orosanu, Othman Lachhab, Larbi Mesbahi.

6.2.1. Core recognition

6.2.1.1. 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 5.4).

Extensions of the ANTS system have been studied, including the possibility to use the sphinx recognizers. Training scripts for building acoustic models for the Sphinx recognizers are now available and take benefit of the computer cluster for a rapid optimization of the model parameters. The Sphinx models are also used for speech/text alignment on both French and English speech data. A new speech decoding program has been developed for efficient decoding on the computer cluster, and easy modification of the decoding steps (speaker segmentation and clustering, data classification, speech decoding in one or several passes, ...). It handles both the Julius and Sphinx (versions 3 and 4) decoders.

This year, we have proposed an approach to grapheme-to-phoneme conversion based on a probabilistic method: Conditional Random Fields (CRF). CRF gives a long term prediction, and assume a relaxed state independence condition. Moreover, we proposed an algorithm to the one-to-one letter to phoneme alignment needed for CRF training. This alignment is based on discrete HMMs. The proposed system was validated on two pronunciation dictionaries. Different set of input features were studied: POS-tag, context size, unigram versus bigram. Our approach compared favorably with the performance of the state-of-the-art Joint-Multigram Models (JMM) for the quality of the pronunciations, but provided better recall and precision measures for multiple pronunciation variants generation [22] [21].

As the pronunciation lexicon is one the key-points of a speech recognition system, we have investigated to which extent wiktionary data can be used to build such a lexicon. Collecting the pronunciations available for many entries of the wiktionary make possible the creation of an initial pronunciation lexicon. Such initial lexicon is then used for training grapheme-to-phoneme conversion systems (either CRF-based of JMM-based), in order to obtain pronunciation variants for words that are not in the initial pronunciation lexicon extracted from the web wiktionary data. Combining the pronunciation variants generated by the 2 grapheme-to-phoneme systems provides the best results. Although the achieved results are not as good as those obtained with a hand-made pronunciation lexicon, this automatic approach makes possible an easy creation of a pronunciation lexicon for a new language [26].

Confidence measures aim at estimating the confidence of a hypothesis result provided by the speech recognition engine. Two word confidence measures were proposed, which can be computed without waiting for the end of the audio stream; one frame-synchronous and one local. Our local measures achieved performance very close to a state-of-the-art measure which requires the recognition of the whole sentence. A preliminary experiment to assess the contribution of our confidence measure in improving the comprehension of automatic transcription results by hearing impaired was also conducted [10].

6.2.1.2. Speech recognition for interaction in virtual worlds

Automatic speech recognition is investigated for vocal interaction in virtual worlds, in the context of serious games in the EMOSPEECH project. This year, a wizard-of-oz experiment was carried out to collect speech data corresponding to the dialogs from 5 players interacting with a serious game. The players were invited to speak freely to any character of the game with whom it is possible to interact, while the wizard of Oz (a game expert localized in the same room) answered them. Hence, the recorded interactions between the player and the characters of the game are natural dialogs. The audio sessions have been manually transcribed. Each session comprises roughly 30 speech turns (one player's sentence plus one wizard's sentence).

For training the language models, the text dialogs recorded by the TALARIS team (Midiki corpus) on the same serious game (but in a text-based interaction), have been used on addition of available broadcast news corpus. For this purpose we have also manually corrected the Midiki sentences, in order to handle the numerous typos and misspellings as well as chat specific "words" such as smileys (“mdr” or “lol”), emphasized punctuations (“!!!!!”) or over-segmentations such as “é-lec-tro-nique”. This normalization step is a strong requirement for speech recognition models. Different language models have then been created using different vocabulary sizes.

The acoustic models are adapted from the radio broadcast news models, using state-of-the-art Maximum A Posteriori adaptation algorithm. This reduces the mismatch in recording conditions between the game devices and the original models trained on radio streams. We are currently investigating solutions to integrate this adaptation within the speech recognition component and perform it online. At runtime, the targeted strategy is to ask the player to utter some few predefined sentences and to use these sentences to adapt the generic acoustic models to the player's voice.

6.2.2. Speech recognition modeling

Robustness of speech recognition to multiple sources of speech 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 modeling approaches.

6.2.2.1. Detection of Out-Of-Vocabulary words

One of the key problems for large vocabulary continuous speech recognition is the occurrence of speech segments that are not modeled by the knowledge sources of the system. An important type of such segments are so-called Out-Of-Vocabulary (OOV) words (words are not included in the lexicon of the recognizer). Mostly OOV words yield more than one error in the transcription result because the error can propagate due to the language model.

We have investigated, with Frederik Stouten (postdoctoral), to what extent OOV words can be detected. For this we used a classifier that makes a decision about each speech frame whether it belongs to an OOV word or not. Acoustic features for this classifier are derived from three recognition systems. On top of the acoustic features we also used four language model features: the ngram probability, the order of the gram that was used to calculate the language model probability, the unigram probability for the current word and a binary indicator that takes the value one if the word is preceded by a first name.

We propose to exploit the fact that 38% of the OOV word observations in the broadcast news data are pronounced more than one time in a time period of less than 1 minute. To improve the detection of repeated OOV words, we design a clustering module working on the detected OOV word segments. This algorithm is based on the estimation of the entropy. The proposed incremental clustering algorithm has been evaluated on the broadcast news corpus ESTER and gave better performance than a classical baseline incremental clustering algorithm based on a distance threshold [36].

6.2.2.2. Detailed modeling exploiting uncertainty

Modeling pronunciation variation is an important topic for automatic speech recognition. It has been widely observed that speech recognition performance degrades notably on spontaneous speech, and more precisely, that the word error rate increases when the degree of spontaneity increases. The rate of speech is also an important variability source which impacts notably on the acoustic realization of the sounds as well as on the pronunciation of the words, and consequently affects recognition performance. Large increases in word error rates are observed when speaking rate increases. And, it should be noted that rate of speech and spontaneous speech are not completely independent as the rate of speech is an important cue for detecting spontaneous speech.

This year, we have investigated further the detailed modeling of the probabilities of pronunciation variants for large vocabulary continuous speech recognition, and evaluated it on broadcast news transcriptions. In particular we have refined the modeling of the probabilities of the pronunciation variants dependent on the speaking rate. This was achieved by taking into account the uncertainty in the estimation of the speaking rate that results from the word and phoneme boundary uncertainty (speech signal - phoneme alignment errors). Such uncertainty was handled both in the training process and in the decoding step, leading to speech recognition performance improvements [25].

Detailed acoustic modeling was also investigated using automatic classification of speaker data. With such an approach it is possible to go beyond the traditional four class models (male vs female, studio quality vs telephone quality). However, as the amount of training data for each class gets smaller when the number of classes increases, this limits the amount of classes that can efficiently be trained. Hence, this year we have investigated introducing a classification marging in the classification process. With such a marging, which handle boundary classification uncertainty, speech data at the class-boundary may belong to several classes. This increases the amount of training data in each class, which makes the class acoustic model parameters more reliable, and finally improved the overall recognition performance.

6.2.2.3. Speech recognition using distant recording

Speech recognition of distant recording of speech commands was investigated. A set of domotic commands were recorded from a few speakers using a far talking microphone. Acoustic models were adapted to this context using some training data played with a loud speaker, and recorded using a distant microphone. Among other results, preliminary experiments showed the benefit of adapting the models, as well as using a noise robust acoustic analysis when dealing with noisy data.

6.2.2.4. Training HMM acouctic models

At the beginning of his second internship at INRIA Nancy research laboratory, Othman Lachhab focused on the finalization of a speech recognition system based on context-independent HMMs models, using bigram probabilities for the phonotactic constraints and a model of duration following a normal distribution $\mathcal{N}(\mu, \sigma^2)$ incorporated directly in the Viterbi search process. Currently, he built a reference system for speaker-independent continuous phone recognition using Context- Independent Continuous Density HMM (CI-CDHMM) modeled by Gaussian Mixture Models (GMMs). In this system he developed his own training

technique, based on a statistical algorithm estimating the classical optimal parameters. This new training process compares favorably with already published HMM technology on the same test corpus (TIMIT).

6.2.3. Speech/text alignment

6.2.3.1. Alignment with native speech

Speech to text alignement is a research objective that is derived from speech recognition. While it seems easier to solve at first sight, expectations are also higher and new problems appear, such as how to handle very large audio documents, or how to handle out-of-vocabulary words. Another important challenge that motivated our work in this area concerns how to improve our results and meet the user expectation by exploiting as much as possible the interactions and feedback loop between the end-user and the system. This year, we kept on improving the open-source JTrans software platform for this task as described in Section (see section 5.6). We further submited an ANR Corpus proposal in collaboration with University Paris 3. We also sent a new version of the software to the "Timecode" company to help them investigating the usefulness of this approach in the application context of foreign film dubbing (see section 7.4.1).

6.2.3.2. Alignment with non-native speech

Non-native speech alignment with text is one critical step in computer assisted foreign language learning. The alignement is necessary to analyze the learner's utterance, in view of providing some prosody feedback (as for example bad duration of some syllables - too short or too long -). However, non-native speech alignement with text is much more complicated than native speech alignment. This is due to the pronunciation deviations observed on non-native speech, as for example the replacement of some target language phonemes by phonemes of the mother tongue, as well as errors in the pronunciations. Moreover, these pronunciation deviations are strongly speaker dependent (i.e. they depend on the mother tongue of the speaker, and on its fluency in the target foreign lanaguage) which makes their prediction difficult.

In this application context, the precision of phoneme boundaries is critical. Hence, speech-text alignment was investigated on non-native speech. A large non-native speech corpus has been manually segmented for building a reference corpus. Then automatic phonetic segmentation (resulting from the speech-text alignment) has been analyzed. The results shows that rather reliable boundaries are obtained for some phonetic classes [31] and that better results are obtained when only frequent pronunciation deviations are kept as variants in the pronunciation lexicon [27]. Further work is on-going to determine automatically a confidence value on the proposed alignments.

6.2.4. Computing and merging linguistic information on speech transcripts

The raw output of speech recognition is difficult to read for humans, and difficult to exploit for further automatic processing. We thus investigated solutions to enrich speech recognition outputs with non-lexical information, such as dialog acts, punctuation marks and syntactic dependencies. Computing such a linguistic information requires a corpus to train stochastic models, and we also worked out new semi-supervised training algorithms for building a French corpus dedicated to syntactic parsing of oral speech. The creation of this corpus is realized in collaboration with the TALARIS team. Finally, we designed a new solution to improve our core language models by integrating into them lexical semantic distances.

An important information for post-processing speech transcripts concerns dialog acts and punctuation marks. We initiated some work in this area several years ago with the Ph.D. thesis of Pavel Kral. Since then, we continued our collaboration in this domain by successively investigating specific challenges, such as finding the most relevant features, models and testing the adaptation of our approaches in two languages, Czech and French [59]. We further proposed this year an approach to improve commas generation with the help of syntactic features [17].

Infering syntactic dependencies is an extremely important step towards structuring the text and an absolute prerequisite for working with relations between words and next interpreting the utterance. Yet, no state-of-the-art solutions designed for parsing written texts can be reliably adapted to parsing speech, and even less transcribed speech. The lack of such methods and resources is especially blatant in French. We started, in collaboration with the TALARIS team, to address this issue by building a new French treebank dedicated to

speech parsing [52], as well as a software platform dedicated to working with this corpus (see section 5.5). We exploited this year this corpus to study specific syntactic structures, such as negations (Master internship in 2011) and left dislocations in French [13].

While a large part of our work is dedicated to enriching the output of our speech recognition system, we also tried integrating within the speech decoding process itself new information coming from the higher-levels. We thus extended the new approach proposed in 2010 about language model smoothing with a new probabilistic smoothing that takes into account much longer words history thanks to a Levenshtein-based clustering of the training sentences [20].

6.3. Speech-to-Speech Translation and Langage Modeling

Participants: Kamel Smaïli, David Langlois, Sylvain Raybaud.

Our work on Confidence Measures is now published in Machine Translation [9]. Now we are working on Speech-to-Text translation. We have proposed a method to segment audio input stream for machine translation. First, audio stream is splitted into overlapping segments useful for speech recognition; then, these segments are transcribed and regrouped; last, the obtained text flow is segmented into machine translation-friendly segments and translated. We incorporated this work into our speech-to-text machine translation system, and we evaluated our system for french-english broadcast news translation [34]. The following step consists in integrating Confidence Measures into the system in order to improve the integration between the both recognition-translation processes.

Moreover, we pursued our collaboration with Chiraz Latiri from the URPAH Team, University of Tunis. Running on our previous works (based on word-based machine translation system) we compared our respective methods in the scope of phrase-based machine translation [30].

7. Contracts and Grants with Industry

7.1. Introduction

Our policy in terms of technological and industrial partnership consists in favoring contracts that quite precisely fit our scientific objectives. We are involved in an ANR project about audiovisual speech synthesis, another about acoustic-to-articulatory inversion of speech (ARTIS), another about the processing of articulatory data (DOCVACIM) and in a national evaluation campaign of automatic speech recognition systems (ETAPE). We also coordinated until January 2009 the 6th PCRD project ASPI about acoustic-to-articulatory inversion of speech, and the Rapsodis ARC until october 2009.

In addition, we are involved in several regional projects.

7.2. Regional Actions

7.2.1. CPER MISN TALC

The team is involved in the management of the Contrat Plan Etat-Région (CPER) contract. In particular, Christophe Cerisara is co-responsible, with Claire Gardent, of the CPER MISN TALC, which objective is to leverage collaborations between regional academic and private partners in the domain of Natural Language Processing and Knowledge engineering. The TALC action involves about 12 research teams and 40 researchers for a budget of about 240,000 euros per year.

In addition to the co-management of this project, our team is also involved in two scientific collaborative operations:

- An operation about text-to-speech alignement, in collaboration with the TALARIS research team and the ATILF laboratory. This operation aims at proposing semi-supervised solutions to facilitate the transcription and processing of large bimodal text and speech corpora. The main outcomes of this operation are (1) the JTrans software described in section 5.6, and a concordancer that was developped in Java by two BSc students in the framework of their final year project.
- An operation about syntactic analysis of speech transcripts, in collaboration with the TALARIS research team and the ATILF laboratory. This operation aims at adapting state-of-the-art stochastic parsers to the specificities of manual and automatic transcriptions of speech, and at building a French treebank of broadcast news speech transcripts. The main outcome of this operation is the J-Safran software, described in section 5.5.

7.2.2. "Intonale": Perception and production of prosodic contours in L1 and L2

This action, launched by the CCOSL, aims at developping collaboration between academic partners from Lorraine laboratories and universities. It has started in september 2009 and should last until the end of 2011. The speech team from LORIA is associated with the laboratory ATILF (Mathilde Dargnat). The project deals with the perception and production of prosodic contours in the first language (L1) and in a second language (L2). We have chosen two radically different languages with respect to prosody : French and English. We have collected a corpus recorded by 34 French speakers and made up of sentences with different modalities: assertions, questions, major and minor continuations. French speakers uttered these sentences both in French (their native language) and in English (the "targeted" non native language). The English part of the corpus is used by the project ALLEGRO, presented hereafter. The French part of the corpus is currently segmented, whilst its English part is segmented under the framework of the INTERREG project ALLEGRO. In order to record corpora in other languages, we improved the Corpus Recorder software (see 5.9). The previous corpus had also been recorded by native english speakers (French and English sentences).

7.3. National Contracts

7.3.1. ADT Handicom

An ADT (Action of Technological Development), was led from 2008 till 2010, managed by Agnès Piquard-Kipffer. The aim of this project is to provide help for improving French Language Acquisition for hard of hearing (HOH) chidren or for chidren with language disabilities.

A collection of three digital books has been written by Agnès Piquard-Kipffer and a web interface has been created in order to create others books for language impaired children.

A workflow which transforms a text and an audio source in a video of digital head has been developed. This workflow includes:

- An automatic speech alignment has been integrated. This process can retrieve from an acoustic signal and a text transcription, the length and the position of each phoneme and of each word. This allows a synchronization of the articulation of the head with acoustic signal and text display. This technology is a recognition engine, result of a previous work called ESPERE from EPI PAROLE.
- A Phonetic transcription designed in the EPI Parole has been integrated and adapted.
- A Speech synthetizer has been integrated. This technology can create an artificial voice from a text. It's a part of tools provided to make a digital book. Several software programs are tested in order to find the best result.
- A French cued speech coding and talking head has been improved in order to generate videos on a server. The animation consists in animating a 3D talking head, in association with a 3D hand which can code cued speech. This technology was created from a previous RIAM project called LABIAO.

A digital book written in FLASH has been developed. It integrates videos of the digital head, which are synchronized with texts displayed for each page. Digital books can be created manually with a text editor (to create XML file) or automatically with software which can be easily used to add all necessary multimedia elements in pages.

Data (audio source and text) are provided from a web interface. This web site allows users to create digital books. Through this interface, the books can be easily modified, shared and read. This website has been developed with Symfony (PHP 5 web framework) and AJAX (Dojo toolkit API) technologies. A linguistical study and a case study analysis of the current version of the talking head and of the digital books were conducted in collaboration (for feasability studies), both with the Speech Therapy School of Nancy (with 8 students : Floriane Jacques, Amélie Dumont, Sophie Bardin, Elodie Racine, Claire Nostrenoff, Anaïs Laurenceau, Hélène Thiollier and Marie Gabet) and with National Education with two schools and specialized teachers (Hélène Adam-Piquard and Sylvie Nussbaum).

7.3.2. ANR DOCVACIM

This contract, coordinated by Prof. Rudolph Sock from the Phonetic Institute of Strasbourg (IPS), addresses the exploitation of X-ray moving pictures recorded in Strasbourg in the eighties. Our contribution is the development of tools to process X-ray images in order to build articulatory model [35]. This year we incorporated tools to withdraw jumps in X-ray films, which are due to the driving of the film during recording. We also developed an analysis procedure to delineate velum contours and to analyze its deformations.

7.3.3. ANR ARTIS

This contract started in January 2009 in collaboration with LTCI (Paris), Gipsa-Lab (Grenoble) and IRIT (Toulouse). Its main purpose is the acoustic-to-articulatory inversion of speech signals. Unlike the European project ASPI the approach followed in our group will focus on the use of standard spectra input data, i.e. cepstral vectors. The objective of the project is to develop a demonstrator enabling inversion of speech signals in the domain of second language learning.

This year the work has focused on the development of the inversion infrastructure using cepstral data as input. We checked that the codebook represents the articulatory to acoustic mapping correctly and we also developed the optimization of the bilinear transform in order to make the comparison of natural and synthetic spectra possible.

7.3.4. ANR ViSAC

This ANR Jeunes Chercheurs started in 2009, in collaboration with Magrit group. The main purpose of ViSAC (Acoustic-Visual Speech Synthesis by Bimodal Unit Concatenation) is to propose a new approach of a text-to-acoustic-visual speech synthesis which is able to animate a 3D talking head and to provide the associated acoustic speech. The major originality of this work is to consider the speech signal as bimodal (composed of two channels acoustic and visual) "viewed" from either facet visual or acoustic. The key advantage is to guarantee that the redundancy of two facets of speech, acknowledged as determining perceptive factor, is preserved.

Currently, we designed a complete system of the text to acoustic-visual speech synthesis based on a relatively small corpus. The system is using bimodal diphones (an acoustic component and a visual one) and it is using unit selection techniques. Although the database for the synthesis is small, however the first results seem to be very promising. The developed system can be used with a larger corpus. We are trying to acquire/analyze an 1-2 hours of audiovisual speech. With a larger corpus, the quality of the synthesis will be obviousely much better.

The next year, we will mainly evaluate the system using both subjective and objective perceptual evaluation.

7.4. Grants with Industry

7.4.1. Timecode

We begin a collaboration with the Timecode company that works in dubbing (recording and replacing voices on a motion picture or television soundtrack). We want to use tools developped in our team to speed up the process of making a rythmo band (or "lip-sync band"). The band is actually a clear 35 mm film leader on which the dialogue is written, along with numerous additional indications for the actor (laughs, cries, length of syllables, mouth sounds, breaths, mouth openings and closings, etc.). The rythmo band is projected in the studio and scrolls in perfect synchronization with the picture. We have designed a tool for automatic alignment of the rythmo band and the audio file.

8. Partnerships and Cooperations

8.1. International Contracts

8.1.1. CMCU - Tunis University

This cooperation involves the LSTS (Laboratoire des systèmes et Traitement du Signal) of Tunis University headed by Prof. Noureddine Ellouze and Kais Ouni. This new project involves the investigation of automatic formant tracking, the modelling of peripheral auditory system and more generally speech analysis and parameterization that could be exploited in automatic speech recognition.

8.1.2. The Oesovox Project 2009-2011: 4 international groups associated...

It is possible for laryngectomees to learn a substitution voice: the esophageal voice. This voice is far from being natural. It is characterized by a weak intensity, a background noise that bothers listening, and a low pitch frequency. A device that would convert an esophageal voice to a natural voice would be very useful for laryngectomees because it would be possible for them to communicate more easily. Such natural voice restitution techniques would ideally be implemented in a portable device. In order to answer the INRIA Euromed 3+3 Mediterranean 2006 call, the INRIA Parole group (Joseph Di Martino, LORIA senior researcher, Laurent Pierron, INRIA engineer and Pierre Tricot, Associated Professor at INPL-ENSEM) associated with the following partners:

- **Spain**: Begoña Garcia Zapirain, Deusto University (Bilbao-Spain), Telecommunication Department, PAS-"ESOIMPROVE" research group.
- **Tunisia**: Sofia Ben Jebara, TECHTRA research group, SUP'COM, Tunis.
- **Morocco**: El Hassane Ibn-Elhaj, SIGNAL research group, INPT, Rabat.

This project named LARYNX has been subsidized by the INRIA Euromed program during the years 2006-2008. Our results have been presented during the INRIA 2008 Euromed colloquium (Sophia Antipolis, 9-10 October 2008). During this international meeting, The French INRIA institute decided to renew our project with the new name "OESOVOX". This new project will be subsidized during the years 2009-2011.

In the framework of the European COADVISE-FP7 program, two PhD students have assigned to the Euromed 3+3 Oesovox project. These students are, Miss Fadoua Bahja from INPT-Rabat (Morocco) whose PhD thesis title is "Detection of F0 in real-time for audio: application to pathological voices" and Mr. Ammar Werghi from SUP'COM-Tunis (Tunisia) whose PhD thesis title is "Voice conversion techniques applied to pathological voice repair". The activity reports of these two students for the year 2009 is described in 6.1.6.

8.2. European Initiatives

8.2.1. Collaborations in European Programs, except FP7

Program: Interreg

Project acronym: Allegro

Project title: Adaptive Language LEarning technology for the Greater Region

Duration: 01/01/2009 to 31/12/2012

Coordinator: Saarland University

Other partners: Supélec Metz and DFK Kaiserslautern

Abstract: Allegro is an Interreg project (in cooperation with the Department of COmputational LInguistics and Phonetics of the Saarland University and Supélec Metz) which started in April 2010. It is intended to develop software for foreign language learning. Our contribution consists of developing tools to help learners to master the prosody of a foreign language, i.e. the prosody of English by french learners, and then prosody of French by german learners. We started by recording (with the project Intonale) and segmentating of a corpus made up of English sentences uttered by French speakers and we analyzed specific problems encountered by French speakers when speaking English.

Program: Eurostar

Project acronym: Emospeech

Project title: Interagir naturellement et émotiennellement avec des environnements virtuels

Duration: 01/06/2009 to 01/06/2012

Coordinator: Artefacto

Other partners: Acapela Speech group

Abstract: The Emospeech project is an Eurostar project started on 1st June 2010 in cooperation with SMEs Artefacto (France) and Acapela (Belgium). This project comes within the scope of serious games and virtual worlds. If existing solutions reach a satisfying level of 3D physical immersion, they do not provide satisfactory natural language interactions. The objective is thus to add spoken interactions via automatic speech recognition and speech synthesis. EPI Parole and Talaris take part in this project and the contribution of Parole will be about the interaction between the virtual world, automatic speech recognition and the dialogue management.

8.3. International Initiatives

8.3.1. Visits of International Scientists

8.3.1.1. Internships

Nicolas VINUESA (from Mar 2011 until Aug 2011)

Subject: Dealing with automatic classification uncertainty in training acoustic models for speech recognition

Institution: Universidad Nacional de Rosario (Argentina)

9. Dissemination

9.1. Animation of the scientific community

- The members of the team frequently review articles and papers for Journal of Phonetics, JASA, Acta Acoustica, CSL, Speech communication, TAL, IEEE Journal of Selected Topics in Signal Processing, IEEE Transaction of Information Theory, Signal Processing, Multimedia Tools, Pattern Recognition Letters, ICASSP, INTERSPEECH, EURASIP, JEP.
- Member of editorial boards :
  - Speech Communication (J.P. Haton, D. Jouvet)
  - Computer Speech and Language (J.P. Haton)

  - EURASIP Journal on audio, Speech, and Music Processing (Y. Laprie)
- Member of scientific commitee of conference :
  - TAIMA, SIIE (K. Smaïli)
  - AVSP2011 (international conference on Auditory-Visual Speech Processing), (S. Ouni)
- Member of organisational conference commitees :
  - Interspeech 2013, (C. Cerisara)
- Chairman of French Science and Technology Association (J.P. Haton)
- Member of "Association Française pour la Communication Parlée" (French Association for Oral Communication) board (I. Illina)
- Member of the lorrain network on specific language and Learning disabilities and in charge of the speech and language therapy expertise in the Meurthe-et-Moselle House of Handicap (MDPH) (A. Kipffer-Piquard)
- The members of the team have been invited as lecturer:
  - Slim Ouni; "Acquisition de donnies articulatoires par un articulographe" at Sorbonne Nouvelle, Journies d'itudes "Typologie des rhotiques", 28-29 juin 2011
  - Agnès Piquard-Kipffer, Finnish Center of Excellence in Learning and Motivation - (Finlande, Jyväskylä).
  - Agnès Piquard-Kipffer, EHESP, Ecole des Hautes Etudes de Santé Publique - Rennes, Sorbonne Paris Cité Université.

9.2. Invited lectures

- Hugo Van Hamme (Katholieke Universiteit Leuven), TALC "Vocabulary acquisition by machines"
- Elisabeth Delais-Roussarie (LLF-Paris 7, Directeur de Recherche au CNRS), ATILF Seminar (Intonale Project).
- Philippe Martin (University Paris-Diderot Paris 7, Professor), ATILF Seminar (Intonale Project).
- Yi Xu (Department of Speech, Hearing and Phonetic Sciences, University College London), TALC seminar, on the topic of Intonale Project (Prosody modeling).
- Peter Birkholz (Clinic for Phoniatrics, Pedaudiology, and Communication Disorders, University Hospital Aachen and RWTH Aachen University), PAROLE Seminar.

9.3. Teaching

cours spécifiques en rapport avec les thèmes de l'équipe : Analyse, Traitement et reconnaissance de la parole, 30HETD, M1, université Henri Poincaré (V. Colotte and D. Langlois), Natural Langue Processing (in English), 24HETD, M2 Erasmus Mundus, université Nancy2 (K. Smaï),
the professors and associate professors of the team are teaching at Henri Poincaré University, at nancy 2 university and INPL.

In addition to courses, we highlight the following activities:

- A strong involvement of the team members in education and administration (University Henri Poincaré, University Nancy 2, INPL): Master of Computer Science, IUT, MIAGE, Speech and Language Therapy School of Nancy;
- Coordinator of C2i (Certificat Informatique et Internet) at Henri Poincaré University (V. Colotte).
- Head of MIAGE Maroc (students of University Nancy 2 but having their courses in Morocco)(K. Smaïli),
- Head of UFR Math-Info at University Nancy2 (K. Smaïli),
- Head of Networking Speciality of University Henri Poincaré Master of Computer Science until 1st September (O. Mella).
- co-Director of DU, « Troubles du Langage et des Apprentissages », Université de Nancy 1, Faculté de Médecine (Agnès Piquard-Kipffer)

PhD & HdR (Les thèses soutenues doivent figurer dans la bibliographie) :

PhD in progress : Christian Gillot, Modèles de langue à base d'exemples pour la reconnaissance de la parole, september 2008, Christophe Cerisara

PhD in progress : Raybaud Sylvain, Traduction automatique par apprentissage discriminant, september 2008, K. Smaïli (supervisor) and D. Langlois (co-supervisor)

PhD in progress : Julie Busset, Acoustic to articulatory inversion from cepstral data., september 2009, Yves Laprie

PhD in progress : Utpala Musti, Bimodal Acoustic-Visual Synthesis, september 2009, S. Ouni and V. Colotte

PhD in progress : Fadoua Bahja, Détection du fondamental de la parole : application au rehaussement de la voix pathologique, Elhassane Ibn Elhaj (supervisor) and Joseph Di Martino (co-supervisor)

PhD in progress : Othman Lachhab, Reconnaissance de la parole continue pour voix naturelle et pathologique, Elhassane Ibn Elhaj (supervisor) and Joseph Di Martino (co-supervisor)

PhD in progress : Arseniy Gorin, Handling trajectories and speaker consistency in automatic speech recognition, october 2011, Denis Jouvet

10. Bibliography

Major publications by the team in recent years

[1] M. ABBAS, K. SMAÏLI, D. BERKANI. *Multi-category support vector machines for identifying Arabic topics*, in "Journal of Research in Computing Science", 2009, vol. 41.

[2] A. BONNEAU, Y. LAPRIE. *Selective acoustic cues for French voiceless stop consonants*, in "The Journal of the Acoustical Society of America", 2008, vol. 123, p. 4482-4497, http://hal.inria.fr/inria-00336049/en/.

[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", 2007, vol. 21, n^{o} 3, p. 443-457.

[4] 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/.

[5] C. LATIRI, K. SMAÏLI, C. LAVECCHIA, D. LANGLOIS. *Mining monolingual and bilingual corpora*, in "Intelligent Data Analysis", November 2010, vol. 14, n^{o} 6, p. 663-682, http://hal.inria.fr/inria-00545493/en.

[6] C. LAVECCHIA, K. SMAÏLI, D. LANGLOIS, J.-P. HATON. *Using inter-lingual triggers for Machine translation*, in "Eighth conference INTERSPEECH 2007", Antwerp/Belgium, 08 2007, http://hal.inria.fr/inria-00155791/en/.

[7] 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)", 2005, vol. 118 (1), p. 444–460, http://hal.archives-ouvertes.fr/hal-00008682/en/.

[8] S. RAYBAUD, D. LANGLOIS, K. SMAÏLI. *"This sentence is wrong." Detecting errors in machine-translated sentences.*, in "Machine Translation", August 2011, vol. 25, n^{o} 1, p. p. 1–34 [*DOI :* 10.1007/s10590-011-9094-9], http://hal.inria.fr/hal-00606350/en.

Publications of the year

Articles in International Peer-Reviewed Journal

[9] S. RAYBAUD, D. LANGLOIS, K. SMAÏLI. *"This sentence is wrong." Detecting errors in machine-translated sentences.*, in "Machine Translation", August 2011, vol. 25, n^{o} 1, p. p. 1–34 [*DOI :* 10.1007/s10590-011-9094-9], http://hal.inria.fr/hal-00606350/en.

[10] J. RAZIK, O. MELLA, D. FOHR, J.-P. HATON. *Frame-Synchronous and Local Confidence Measures for Automatic Speech recognition*, in "International Journal of Pattern Recognition and Artificial Intelligence", 2011, vol. 25, n^{o} 2, p. 1-26 [*DOI :* 10.1142/S0218001411008543], http://hal.inria.fr/hal-00579092/en.

[11] A. TOUTIOS, S. OUNI, Y. LAPRIE. *Estimating the control parameters of an articulatory model from electromagnetic articulograph data*, in "The Journal of the Acoustical Society of America", May 2011, vol. 129, n^{o} 5, p. 3245-3257 [*DOI :* 10.1121/1.3569714], http://hal.inria.fr/inria-00578733/en.

Articles in National Peer-Reviewed Journal

[12] F. TANTINI, A. TERLUTTE, F. TORRE. *Combinaisons d'automates et de boules de mots pour la classification de séquences*, in "Revue d Intelligence Artificielle", June 2011, vol. 25, n^{o} 3, p. 411-434 [*DOI :* 10.3166/RIA.25.411-434], http://hal.inria.fr/hal-00643057/en.

International Conferences with Proceedings

[13] C. ANDERSON, C. CERISARA, C. GARDENT. *Vers la détection des dislocations à gauche dans les transcriptions automatiques du Français parlé / Towards automatic recognition of left dislocation in transcriptions of Spoken French*, in "Traitement Automatique des Langues Naturelles - TALN'2011", Montpellier, France, June 2011, 6, http://hal.inria.fr/hal-00600510/en.

[14] A. BONNEAU, B. WROBEL-DAUTCOURT. *Efficiency of five labial correlates for /i/ and /y/ in adverse contexts*, in "The ninth International Seminar on Speech Production - ISSP'11", Montreal, Canada, 2011, http://hal.inria.fr/inria-00579160/en.

[15] J. BUSSET, Y. LAPRIE. *Adaptation of cepstral coefficients for acoustic-to-articulatory inversion*, in "International Seminar on Speech Production 2011 - ISSP'11", Montréal, Canada, June 2011, http://hal.inria.fr/inria-00599108/en.

[16] C. CERISARA, C. GARDENT. *The JSafran platform for semi-automatic speech processing*, in "12th Annual Conference of the International Speech Communication Association - Interspeech 2011", Florence, Italie, August 2011, 4, http://hal.inria.fr/hal-00600520/en.

[17] C. CERISARA, P. KRAL, C. GARDENT. *Commas recovery with syntactic features in French and in Czech*, in "12thAnnual Conference of the International Speech Communication Association - Interspeech 2011", Florence, Italie, August 2011, 4, http://hal.inria.fr/hal-00600528/en.

[18] S. DEMANGE, S. OUNI. *Acoustic-to-articulatory inversion using an episodic memory*, in "International Conference on Acoustics, Speech and Signal Processing (ICASSP)", Prague, Tchèque, République, IEEE - Signal Processing Society, May 2011, http://hal.inria.fr/inria-00578740/en.

[19] S. DEMANGE, S. OUNI. *Continuous episodic memory based speech recognition using articulatory dynamics*, in "12th Annual Conference of the International Speech Communication Association - Interspeech 2011", Florence, Italie, August 2011, http://hal.inria.fr/inria-00602414/en.

[20] C. GILLOT, C. CERISARA. *Similarity language model*, in "12thAnnual Conference of the International Speech Communication Association - Interspeech 2011", Florence, Italie, August 2011, 4, http://hal.inria.fr/hal-00600531/en.

[21] I. ILLINA, D. FOHR, D. JOUVET. *Grapheme-to-Phoneme Conversion using Conditional Random Fields*, in "12th Annual Conference of the International Speech Communication Association - Interspeech 2011", Florence, Italie, International Speech Communication Association (ISCA) et The Italian Regional SIG - AISV (Italian Speech Communication Association), August 2011, http://hal.inria.fr/inria-00614981/en.

[22] I. ILLINA, D. FOHR, D. JOUVET. *Multiple Pronunciation Generation using Grapheme-to-Phoneme Conversion based on Conditional Random Fields*, in "XIV International Conference "Speech and Computer" (SPECOM'2011)", Kazan, Russie, Fédération De, September 2011, http://hal.inria.fr/inria-00616325/en.

[23] S. INGMAR, S. OUNI. *Investigating articulatory differences between upright and supine posture using 3D EMA*, in "9th International Seminar on Speech Production - ISSP'11", Montreal, Canada, June 2011, http://hal.inria.fr/inria-00602427/en.

[24] S. INGMAR, S. OUNI. *Towards an articulatory tongue model using 3D EMA*, in "9th International Seminar on Speech Production - ISSP'11", Montreal, Canada, June 2011, http://hal.inria.fr/inria-00602423/en.

[25] D. JOUVET, D. FOHR, I. ILLINA. *About handling boundary uncertainty in a speaking rate dependent modeling approach*, in "12th Annual Conference of the International Speech Communication Association - Interspeech 2011", Florence, Italie, International Speech Communication Association (ISCA) et The Italian Regional SIG - AISV (Italian Speech Communication Association), August 2011, http://hal.inria.fr/inria-00614781/en.

[26] D. JOUVET, D. FOHR, I. ILLINA. *Building a Pronunciation Lexicon for a Speech Transcription System from Wiktionary Pronunciations only*, in "XIV International Conference "Speech and Computer" (SPECOM'2011)", Kazan, Russie, Fédération De, September 2011, http://hal.inria.fr/inria-00616330/en.

[27] D. JOUVET, L. MESBAHI, A. BONNEAU, D. FOHR, I. ILLINA, Y. LAPRIE. *Impact of Pronunciation Variant Frequency on Automatic Non-Native Speech Segmentation*, in "5th Language & Technology Conference - LTC'11", Poznan, Pologne, November 2011, p. 145-148, http://hal.inria.fr/hal-00639118/en.

[28] Y. LAPRIE, J. BUSSET. *A curvilinear tongue articulatory model*, in "International Seminar on Speech Production 2011 - ISSP'11", Montréal, Canada, June 2011, http://hal.inria.fr/inria-00599109/en.

[29] Y. LAPRIE, J. BUSSET. *Construction and evaluation of an articulatory model of the vocal tract*, in "19th European Signal Processing Conference - EUSIPCO 2011", Barcelona, Espagne, August 2011, http://hal.inria.fr/inria-00599130/en.

[30] C. LATIRI, K. SMAÏLI, C. LAVECCHIA, C. NASRI, D. LANGLOIS. *Phrase-based machine translation based on text mining and statistical language modeling techniques*, in "12th International Conference on Intelligent Text Processing and Computational Linguistics - CICLing2011", Tokyo, Japon, 2011, http://hal.inria.fr/inria-00579335/en.

[31] L. MESBAHI, D. JOUVET, A. BONNEAU, D. FOHR, I. ILLINA, Y. LAPRIE. *Reliability of non-native speech automatic segmentation for prosodic feedback*, in "Workshop on Speech and Language Technology in Education - SLaTE 2011", Venise, Italie, ISCA, August 2011, http://hal.inria.fr/inria-00614930/en.

[32] U. MUSTI, V. COLOTTE, A. TOUTIOS, S. OUNI. *Introducing Visual Target Cost within an Acoustic-Visual Unit-Selection Speech Synthesizer*, in "International Conference on Auditory-Visual Speech Processing - AVSP2011", Volterra, Italie, August 2011, http://hal.inria.fr/inria-00602403/en.

[33] S. OUNI. *Tongue Gestures Awareness and Pronunciation Training*, in "12th Annual Conference of the International Speech Communication Association - Interspeech 2011", Florence, Italie, August 2011, (accepted), http://hal.inria.fr/inria-00602418/en.

[34] S. RAYBAUD, D. LANGLOIS, K. SMAÏLI. *Broadcast news speech-to-text translation experiments*, in "The Thirteenth Machine Translation Summit", Xiamen, Chine, September 2011, p. 378-381, http://hal.inria.fr/hal-00628101/en.

[35] R. SOCK, F. HIRSCH, Y. LAPRIE, P. PERRIER, B. VAXELAIRE, G. BROCK, F. BOUAROUROU, C. FAUTH, V. FERBACH-HECKER, L. MA, J. BUSSET, J. STURM. *An X-ray database, tools and procedures for the study of speech production*, in "9th International Seminar on Speech Production (ISSP 2011)", Montréal, Canada, June 2011, p. 41-48, http://hal.inria.fr/hal-00610297/en.

[36] F. STOUTEN, I. ILLINA, D. FOHR. *Clustering repeated Out-Of-Vocabulary word tokens in order to model them for broadcast news transcription*, in "The XIVth International Conference Speech and Computer - SPECOM'2011", Kazan, Russie, Fédération De, September 2011, p. 73-80, http://hal.inria.fr/hal-00639123/en.

[37] A. TOUTIOS, U. MUSTI, S. OUNI, V. COLOTTE. *Weight Optimization for Bimodal Unit-Selection Talking Head Synthesis*, in "12thAnnual Conference of the International Speech Communication Association - Interspeech 2011", Florence, Italie, August 2011, (accepted), http://hal.inria.fr/inria-00602407/en.

[38] A. TOUTIOS, S. OUNI. *Predicting Tongue Positions from Acoustics and Facial Features*, in "12th Annual Conference of the International Speech Communication Association - Interspeech 2011", Florence, Italie, August 2011, http://hal.inria.fr/inria-00602412/en.

Conferences without Proceedings

[39] M. DARGNAT, A. BONNEAU, K. BARTKOVA, V. COLOTTE. *"Non-conclusive" Slopes in French: First Results*, in "Interface Discours et prosodie 2011", Manchester, Royaume-Uni, University of Salford, September 2011, http://hal.inria.fr/hal-00642721/en.

[40] M. DARGNAT, A. BONNEAU, V. COLOTTE, K. BARTKOVA. *Intra- and Inter-clausal Continuation Slopes in French: First Results*, in "Experimental and Theoretical Advances in Prosody 2", Montréal, Canada, September 2011, http://hal.inria.fr/hal-00642735/en.

Scientific Books (or Scientific Book chapters)

[41] A. BONNEAU, V. COLOTTE. *Automatic Feedback for L2 Prosody Learning*, in "Speech and Language Technologies", I. IPSIC (editor), Intech, June 2011, p. 55-70, http://hal.inria.fr/inria-00579255/en.

[42] M. EMBARKI, S. OUNI, M. YEOU, C. GUILLEMINOT, S. AL MAQTARI. *Acoustic and EMA study of pharyngealization : Coarticulatory effects as index of stylistic and regional distinction*, in "Instrumental Studies in Arabic Phonetics", Current Issues in Linguistic Theory, Benjamins, 2011, p. 1-56, http://hal.inria.fr/hal-00348775/en.

[43] S. SIDHOM, JEAN-PAUL. HATON, M. GHENIMA. *Information Systems and Economic Intelligence : 4th. International Conference SIIE 2011. (Proceedings). February 17-19, 2011 (Marrakech, Morocco)*, IGA Morocco, IGA Morocco, February 2011, vol. 1, http://hal.inria.fr/inria-00580162/en.

References in notes

[44] 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", 1995, vol. 3, n^{o} 4, p. 85–89.

[45] F. BAHJA, J. DI MARTINO, E. H. IBN ELHAJ, D. ABOUTAJDINE. *An improvement of the eCATE algorithm for F0 detection*, in "10th International Symposium on Communications and Information Technologies - ISCIT 2010", Japon Tokyo, 2010, http://hal.inria.fr/inria-00545441/en.

[46] F. BAHJA, J. DI MARTINO, E. H. IBN ELHAJ. *Real-Time Pitch Tracking using the eCate Algorithm*, in "5th International Symposium on I/V Communications over fixed and Mobile Networks - ISIVC 2010", Maroc Rabat, 2010, http://hal.inria.fr/inria-00545435/en.

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

[48] R. CLARK, K. RICHMOND, S. KING. *Festival 2 - Build your own general purpose unit selection speech synhtesiser*, in "ISCA 5th Speech Synthesis Workshop", Pittsburgh, 2004, p. 201–206.

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

[50] 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/.

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

[52] C. GARDENT, C. CERISARA, C. ANDERSON. *Building and Exploiting a Dependency Treebank for French Radio Broadcast*, in "TLT9 – the ninth international workshop on Treebanks and Linguistic Theories", Estonie Tartu, November 2010, http://hal.inria.fr/inria-00537147/en.

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

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

[55] A. KIPFFER-PIQUARD. *Prédiction de la réussite ou de l'échec spécifiques en lecture au cycle 2. Suivi d'une population "à risque" et d'une population contrôle de la moyenne section de maternelle à la deuxième année de scolarisation primaire*, ARNT - Lille, 2006, http://hal.inria.fr/inria-00185312/en/.

[56] A. KIPFFER-PIQUARD. *Prédiction dès la maternelle de la réussite et de l'échec spécifique à l'apprentissage de la lecture en fin de cycle 2*, in "Les troubles du développement chez l'enfant", Amiens France, L'HARMATTAN, 2007, http://hal.inria.fr/inria-00184601/en/.

[57] P. KOEHN, H. HOANG, A. BIRCH, C. CALLISON-BURCH, M. FEDERICO, N. BERTOLDI, B. COWAN, W. SHEN, C. MORAN, R. ZENS, C. DYER, O. BOJAR, A. CONSTANTIN, E. HERBST. *Moses: Open Source Toolkit for Statistical Machine Translation*, in "Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session", June 2007.

[58] P. KOEHN. *Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models*, in "6th Conference Of The Association For Machine Translation In The Americas", Washington, DC, USA, 2004, p. 115-224.

[59] P. KRAL, C. CERISARA. *Dialogue act recognition approaches*, in "Computing And Informatics", 2010, vol. 29, n^{O} 2, p. 227–250, http://hal.inria.fr/inria-00431396/en.

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

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

[62] C. LAVECCHIA, K. SMAÏLI, D. LANGLOIS, J.-P. HATON. *Using inter-lingual triggers for Machine translation*, in "Eighth conference INTERSPEECH 2007", Antwerp/Belgium, 08 2007, http://hal.inria.fr/inria-00155791/en/.

[63] 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.

[64] J. D. MARTINO, Y. LAPRIE. *An Efficient F0 Determination Algorithm based on the Implicit Calculation of the Autocorrelation of the Temporal Excitation Signal*, in "6th European Conference on Speech Communication and Technology EUROSPEECH", 1999.

[65] F. J. OCH, H. NEY. *Improved statistical alignment models*, in "ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics", Morristown, NJ, USA, Association for Computational Linguistics, 2000, p. 440–447.

[66] L. SPRENGER-CHAROLLES, C. BOGLIOTTI, A. PIQUARD-KIPFFER, G. LELOUP. *Stabilité dans le temps des deficits en et hors lecture chez des adolescents dyslexiques (données longitudinales)*, in "ANAE", 2009, vol. 103, p. 243-253.

[67] L. SPRENGER-CHAROLLES, P. COLÉ, A. PIQUARD-KIPFFER, G. LELOUP. *EVALEC, Batterie informatisée d'évaluation diagnostique des troubles spécifiques d'apprentissage de la lecture.*, 2010, http://hal.inria.fr/inria-00545950/en.

[68] L. SPRENGER-CHAROLLES, P. COLÉ, A. KIPFFER-PIQUARD, F. PINTON, C. BILLARD. *Reliability and prevalence of an atypical development of phonological skills in French-speaking dyslexics*, in "Reading and writing", 2009, vol. 22, p. 811-842.

[69] L. SPRENGER-CHAROLLES, P. COLÉ, D. BÉCHENNEC, A. KIPFFER-PIQUARD. *French normative data on reading and related skills from EVALEC, a new computerized battery of tests (end Grade 1, Grade 2, Grade 3, and Grade 4)*, in "Revue Européenne de Psychologie Appliquée", 2005, p. 157-186, http://hal.inria.fr/inria-00184979/en/.

[70] A. WERGHI, J. DI MARTINO, S. BEN JEBARA. *On the Use of an Iterative Estimation of Continuous Probabilistic Transforms for Voice Conversion*, in "5th International Symposium on I/V Communications over fixed and Mobile Networks - ISIVC 2010", Maroc Rabat, 2010, http://hal.inria.fr/inria-00545428/en.