IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2012

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

Creation of the Project-Team: May 01, 2001 .

1. Members

Research Scientists
Yves Laprie [Team Leader, Senior Researcher, CNRS, HdR]
Anne Bonneau [Senior Researcher, CNRS]
Dominique Fohr [Senior Researcher, CNRS]
Denis Jouvet [Senior Researcher, Inria, HdR]

Faculty Members
Vincent Colotte [Associate Professor, Lorraine University]
Joseph di Martino [Associate Professor, Lorraine University]
Irina Illina [Associate Professor, I.U.T. Charlemagne, Lorraine University, HdR]
David Langlois [Associate Professor, IUFM, Lorraine University]
Agnès Piquard-Kipffer [Associate Professor, IUFM, Lorraine University]
Odile Mella [Associate Professor, Lorraine University]
Slim Ouni [Associate Professor, I.U.T. Charlemagne, Lorraine University]
Kamel Smaïli [Professor, Lorraine University, HdR]

External Collaborators
Jean-Paul Haton [Professor emeritus, Lorraine University, Institut Universitaire de France]
Marie-Christine Haton [Professor emeritus, Lorraine University, HdR]

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

PhD Students
Christian Gillot [MENRT grant, thesis defended in Dec 2012]
Sylvain Raybaud [MENRT grant, defended in Dec 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]
Motaz Saad [French ambassy grant since 1st November 2012]
Cyrine Nasri [since september 2011]

Post-Doctoral Fellow
Ingmar Steiner [Inria Cordi Grant until April 2012]

Administrative Assistant
Hélène Zganic [Inria]

2. Overall Objectives

2.1. Introduction

PAROLE is a joint project to Inria, CNRS, Lorraine 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.

2.2. Highlights of the Year

The movie “Je peux voir les mots que tu dis” (ADT Handicom) won the award for the best documentary at the “festival universitaire pédagogique” in Lyon, April 2012

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 6.1.6.2), 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 [45], [46] which indicates that phonemic discrimination at the beginning of kindergarten is strongly linked to success 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 [52].

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 [42] 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 [37] advocates for the theory of expansion movements: the movement tends to be anticipated when no phonological constraint is imposed on labiality. Cohen and Massaro [40] 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 [39] 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 [41] 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 [41]. 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 [44]. 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 [38] 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 [51]. 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. [54] 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 [48] or MOSES [47] 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 [43] 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 Snoori 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 [49]. 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. JSnoori

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

JSnoori is written in Java and uses signal processing algorithms developed within WinSnoori software with the double objective of being a platform independent signal visualization and manipulation tool, and also for designing exercises for learning the prosody of a foreign language. JSnoori thus focused the calculation of F0, the forced alignment of non native English uttered by French speakers and the correction of prosody parameters (F0, rhythm and energy). Since phonetic segmentations and annotations play a central role in the derivation of diagnosis concerning the realization of prosody by learners, several tools have been incorporated to segment and annotate speech. In particular, a complete phonetic keyboard is available, several kinds of annotation can be used (phonemes, syllables and words) and forced alignment can exploit variants to cope with non native accents. In addition, JSnoori offers real time F0 calculation which can be useful from a pedagogical point of view.

5.3. Xarticulator

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

Xarticulators software is intended to delineate contours of speech articulators in X-ray images, construct articulatory models and synthesize speech from X-ray films. This software provide tools to track contours automatically, semi-automatically or by hand, to make the visibility of contours easier, to add anatomical landmarks to speech articulators and to synchronize images together with the sound.

It also enables the construction of adaptable linear articulatory models from the X-ray images.

This year we particularly worked on the possibility of synthesizing speech from X-ray images. We thus designed an algorithm to compute the centerline of the vocal tract in order to segment the vocal tract into elementary tubes approximating the propagation of a one-dimensional wave. In addition we also added the possibility of processing digitized manual delineation results made on sheet of papers when no software was available

5.4. 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 corpora [50]. 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].

5.5. 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.

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

- 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.

5.6. ANTS

contacts : Dominique Fohr (fohr@loria.fr) and Denis Jouvet (denis.jouvet@inria.fr).

The aim of the Automatic News Transcription System (ANTS) is to transcribe radio or TV shows. ANTS is composed of several stages. The first processing steps aim at splitting the audio stream into homogeneous segments of a manageable size and at identifying the segment characteristics in order to allow the use of specific algorithms or models according to the nature of the segment. This includes broad-band/narrow-band speech segmentation, speech/music classification, speaker segmentation and clustering, detection of silences/breathing segments and generally speaker gender classification.

Each segment is then decoded using a large vocabulary continuous speech recognition engine, either the Julius engine or the Sphinx engine. The Julius engine 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. The Sphinx engine processes the speech input segment in a single forward pass using a trigram language model.

Further processing passes are usually run in order to apply unsupervised adaptation processes on the feature computations (VTLN: vocal tract length normalization) and/or on the model parameters (MLLR: maximum likelihood linear regression), or to use speaker adaptive training (SAT) based models. Moreover decoding results of both systems can be efficiently combined for improved decoding performance.

The latest version which relies on a perl script exploits the multiple CPUs available on a computer to reduce the processing time, and runs on both a stand alone linux machine and on the cluster.

5.7. 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 improvements 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

5.8. CoALT

contacts : Dominique Fohr (dominique.fohr@loria.fr) and Odile Mella (odile.mella@loria.fr).

CoALT (Comparing Automatic Labelling Tools) compares two automatic labellers or two speech-text alignment tools, ranks them and displays statistics about their differences. The main feature of our software is that a user can define its own criteria for evaluating and comparing two speech- text alignment tools. With CoALT, a user can give more importance to either phoneme labels or phoneme boundaries because the CoALT elastic comparison algorithm takes into account time boundaries. Moreover, by providing a set of phonetic rules, a user can define the allowed discrepancies between the automatic labelling result and the hand-labelling one.

5.9. 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

5.10. 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).

5.11. VisArtico

contact : Slim Ouni (Slim.Ouni@loria.fr).

VisArtico is intended to visualize articulatory data acquired using an articulograph [30], [29]. It is intended for researchers that need to visualize data acquired from the articulograph with no excessive processing. It is well adapted to the data acquired using the AG500 and AG501 (developed by Carstens Medizinelektronik GmbH), and the articulograph NDI Wave, developed by Northern Digital Inc.

The software allows displaying the positions of the sensors that are simultaneously animated with the speech signal. It is possible to display the tongue contour and the lips contour. The software helps to find the midsagittal plane of the speaker and find the palate contour. In addition, VisArtico allows labeling phonetically the articulatory data.

All this information is very useful to researchers working in the field of speech production, as phoneticians for instance. VisArtico provides several possible views: (1) temporal view, (2) 3D spatial view and (3) 2D midsagittal view. In the temporal view, it is possible to display different articulatory trajectories in addition to the acoustic signal and eventually labels. The midsagittal view can display the tongue contour, the jaw, the lips and the palate.

VisArtico provides several tools to help to improve the quality of interpreting the data. It is cross-platform software as it is developed in JAVA and does not need any external library or framework to be additionally installed. It was tested and worked on Windows, Mac OS, and Linux. It should work on any system having JAVA installed. VisArtico is freely distributed via a dedicated website http://visartico.loria.fr.

6. New Results

6.1. Speech Analysis and Synthesis

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

Signal processing, phonetics, health, perception,articulatory models, speech production, learning language, hearing help, speech analysis, acoustic cues, speech synthesis

6.1.1. Acoustic-to-articulatory inversion

6.1.1.1. Annotation of X-ray films and construction of articulatory models

Two databases have been annotated this year: one composed of 15 short sentences representing more than 1000 X-ray images and a second about CVCVs which has already been annotated by hand on sheets of papers. In the latter case we adapted tools of Xarticul software in order to enable a fast processing of these annotations.

Since images of the first database have been digitized from old films there are several spurious jumps and we thus developed tools to remove them during the construction of articulatory models. The big difference with previous databases processed is the presence of more consonants.

The articulatory model is supplemented by a clipping algorithm in order to take into account contacts between tongue and palate.

6.1.1.2. Articulatory copy synthesis

Acoustic features and articulatory gestures have always been studied separately. Articulatory synthesis could offer a nice solution to study both domains simultaneously. We thus explored how X-ray images could be used to synthesize speech. The first step consisted of connecting the 2D geometry given by mediosagittal images of the vocal tract with the acoustic simulation. Last year we thus developed an algorithm to compute the centerline of the vocal tract, i.e. a line which is approximately perpendicular to the wave front. The centerline is then used to segment the vocal tract into elementary tubes whose acoustic equivalents are fed into the acoustic simulation.

The frequency simulation enables the impact of local modifications of the vocal tract geometry to be evaluated easily. This is useful to investigate the contribution of the sagittal to area transformation in the synthetic speech spectrum. However, the sequence of area functions alone does not suffice to synthesize speech since consonants involve very fine temporal details (closure of the vocal tract and then release of the constriction for stops and fricatives for instance) which additionally have to be synchronized with the temporal evolution of the glottis area. Scenarii have thus been designed for VCV sequences and more generally for any consonant clusters. The idea consists of choosing relevant X-ray images near the VCV to be synthesized. These images can be duplicated just before the closure of the vocal tract, modified to simulate the constriction release for a stop...

This procedure has been applied successfully to copy sentences and VCV for four X-ray films of the DOCVACIM database http://www2i.misha.fr/flora/jsp/index.jsp. The next objective will be to develop a complete articulatory synthesis system.

6.1.1.3. Inversion from cepstral coefficients

The two main difficulties of inversion from cepstral coefficients are: (i) the comparison of cepstral vectors from natural speech and cepstral vectors generated by the articulatory synthesizer and (ii) the access to the articulatory codebook.

Last year we developed a bilinear frequency warping optimized to compensate for the articulatory model mismatch. However, the spectral tilt was not taken into account. We thus combined it with affine adaptation of the very first cepstral coefficients in order to take into account the spectral tilt. It turns out that the new adaptation enables a more relevant comparison of cepstral vectors since the geometric precision of the best solution is less than 1mm.

The second difficulty consists of exploring the articulatory codebook efficiently. Indeed, only a small number of hypercuboids could correspond to the input cepstral vector. The issue is to eliminate all cuboids, which cannot give rise to the input cepstral vector. This is easy when using formants as input data since all cuboids can be indexed easily with extreme values of formants. But this becomes impossible with cepstral vectors because the effect of the excitation source cannot be removed completely from cepstral coefficients. We thus use spectral peaks to access the codebook. However, there exist some spurious spectral peaks, and at the same time some peaks can be absent. We thus designed a lax matching between spectral peaks, which enables the comparison of a series of spectral peaks of the original speech with peaks calculated on synthetic speech. This matching algorithm allows the exploration to focus on 5% of the codebook instead of 40% when using only the peak corresponding to F2 is used.

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.

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.

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. It has also a good precision. In fact, we performed recently an comparative study to assess the precision of the articulograph AG500 in comparison to a concurrent articulograph NDI Wave. In this study, we found that both systems presented similar results. We showed also that the accuracy is relatively independent of the sensor velocity, but decreases with the distance from magnetic center of the system [31].

To make the best use of the articulograph, we developed an original visualization software, VisArtico, which allows displaying the data acquired by an articulograph. It is possible to display the tongue contour and the lips contour animated simultaneously with acoustics. The software helps to find the midsagittal plane of the speaker and find the palate contour. In addition, VisArtico allows labeling phonetically the articulatory data[30].

We continuousely work on the usage this platform to acquire articulatory data that were used for articulatory-to-acoustic inversion but also to study the co-variation of speech clarity and coarticulatory patterns in Arabic [18]. The results revealed evident relationship between speech clarity and coarticulation: more coarticulation in formal speech and in strong prosodic position.

6.1.3. 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). To this end, 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.

The visual target cost includes visual and articulatory information. We implemented and evaluated two techniques: (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. This year, we continued working on improving the quality of the synthesis. This was done by continuously testing new strategies of weight tuning and improving our selection technique [26].

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

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

Our computerized tool EVALEC was published [56] 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) [58]. 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 [57], and by the mean of longitudinal studies of children from age 5 to age 17 [55]. 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.

6.1.4.2. 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 Development) 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.

6.1.5. Enhancement of esophageal voice

6.1.5.1. 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. 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. Recently we developed a new algorithm based on a wavalet transform applied to the cepstrum excitation. The resuts obtained were satisfactory. This work has been published in the ICMCS 2012 conference [14].

6.1.5.2. 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. Recently,we realized that one of the most important problems in speech conversion is the prediction of the excitation. In order to solve this problem we developed a new strategy based on the prediction of the ceptrum excitation pulses. This interesting result has been published in the SIIE 2012 conference [13].

6.1.5.3. Signal reconstruction from short-time Fourier transform magnitude spectra

Joseph Di Martino and Laurent Pierron developed in 2010 an algorithm for real-time signal reconstruction from short-time Fourier magnitude spectra. Such an algorithm has been designed in order to enable voice conversion techniques we are developing in Nancy for pathological voice repair. Recently Mouhcine Chami, an assistant-professor of the INPT institute at Rabat (Morocco) proposed a hardware implementation of this algorithm using FPGAs. This implementation has been publised in the SIIE 2012 conference [17].

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

6.1.6.1. Language learning (feedback on prosody)

A corpus, made up of 8 English sentences and 40 English isolated words has been recorded. Thirty three speakers pronounced the corpus under different conditions : without any audio feedback (first condition), with audio feedback (second condition, experiment realized one week after the first one). In order to test the permanence of the improvement due to feedback, a set of words and all the sentences were then pronounced without feedback (third condition, experiment realized after the second one). An English teacher helped us in the composition of the corpus and recorded it. Parts of this corpus have already been used to test the automatic speech alignment methods developed under the framework of ALLEGRO and implemented in jsnoori (ADT). The feedback will be progressively transferred from Winsnoori to Jsnoori.

6.1.6.2. Production of prosodic contour

The study of French contours (various types of continuations, end of sentences ...) confirmed the existence of patterns which are typical of French prosody. In order to determine the impact of French (the native language) on a second language pronunciation (English), a series of prosodic contours extracted from English sentences uttered by French speakers have been compared to French prosodic countours. To that purpose, French speakers recorded similar sentences in French and in English. Analysis of results is in progress. First results tend to show the impact of the native language ([15] and [10]).

6.1.7. 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 [53] 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. This algorithm is under a submission process in an international journal.

6.2. Automatic Speech Recognition

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

telecommunications, stochastic models, acoustic models, language models, automatic speech recognition, training, robustness

6.2.1. Core recognition

6.2.1.1. Broadcast News Transcription

A complete speech transcription system, named ANTS (see section 5.6), was initially developed in the framework of the Technolangue evaluation campaign ESTER for French broadcast news transcription. This year, in the context of the ETAPE evaluation campaign about transcription of radio and TV debates, the speech transcription system was improved. Large amounts of text data have been collected over the web. This new collected web data, plus new text and speech resources have made possible the creation and training of new acoustic models and new language models. Moreover new processing steps have been included in the transcription system, leading to much better performance than with the initial system. Several system variants have been developed, and for the ETAPE evaluation campaign, their results have been combined.

Extensions of the ANTS system have been studied, including the possibility to use the sphinx recognizers, and unsupervised adaptation processes. Training scripts for building acoustic models for the Sphinx recognizers are now available and take benefit of parallel computations on 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 transcription 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, in the context of the ETAPE evaluation campaign, which deals with the transcription of radio and TV shows, mainly debates, the Julius-based and Sphinx-based transcription systems have been improved. Several system variants have been developed (relying on different features, and/or different normalization schemes, different processing steps, and different unsupervised adaptation processes); and, combining the output of the various systems led to significantly improved performance.

The recently proposed approach to grapheme-to-phoneme conversion based on a probabilistic method: Conditional Random Fields (CRF) was investigated further. CRF gives a long term prediction, and assumes a relaxed state independence condition. The proposed system was validated in a speech recognition context. Our approach compared favorably with the performance of the state-of-the-art Joint-Multigram Models (JMM) for the quality of the pronunciations, and it was also shown that combining the pronunciation variants generated by both the CRF-based and the JMM-based apporaches improves performance [21].

Concerning grapheme-to-phoneme conversion, a special attention was paid to infering the pronunciation variants of proper names [34], and the usage of additional information corresponding to the language origin of the proper name was investigated.

6.2.1.2. Non-native speakers

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

Our approach is based on acoustic model transformation and pronunciation modeling for multiple non-native accents. For acoustic model transformation, two approaches are evaluated: MAP and model re-estimation. For pronunciation modeling, confusion rules (alternate pronunciations) are automatically extracted from a small non-native speech corpus. We presents [9] a novel approach to introduce confusion rules in the recognition system which are automatically learned through pronunciation modelling. The modified HMM of a foreign spoken language phoneme includes its canonical pronunciation along with all the alternate non-native pronunciations, so that spoken language phonemes pronounced correctly by a non-native speaker could be recognized. We evaluate our approaches on the European project HIWIRE non-native corpus which contains English sentences pronunced by French, Italian, Greek and Spanish speakers. Two cases are studied: the native language of the test speaker is either known or unknown. Our approach gives better recognition results than the classical acoustic adaptation of HMM when the foreign origin of the speaker is known. We obtain 22% WER reduction compared to the reference system.

6.2.1.3. Language Model

Christian Gillot has defended his Ph.D. thesis on the 17th September 2012. In his thesis, he proposes a new approach to estimate the language model probabilities for an automatic speech recognition system. The most commonly used language models in the state of the art are based on n-grams smoothed with Kneser-Ney method. Such models make use of occurrence counts of words sequences up to a maximum length (typically 5 words). These counts are computed on a huge training corpus. Christian's Ph.D. thesis starts by an empirical study of the errors of a state-of-the-art speech recognition system in French, which shows that there are many regular language phenomena that are out of reach of the n-gram models. This thesis thus explores a dual approach of the prevailing statistical paradigm by using memory models that process efficiently specific phenomena, in synergy with the n-gram models which efficiently capture the main trends in the corpus. The notion of similarity between long n-grams is studied in order to identify the relevant contexts to take into account in a first similarity language model. The data extracted from the corpus is combined via a Gaussian kernel to compute a new score. The integration of this non-probabilistic model improves the performance of a recognition system. A second model is then introduced, which is probabilistic and thus allows for a better integration of the similarity approach with the existing models. This second model improves the performance on texts in terms of perplexity. Some future works are further described, where the memory-based paradigm is transposed from the estimation of the n-gram probability up to the language model itself. The principle is to combine individual models together, where each model represents a specific syntactic structure, and also to combine these specific models with a standard n-gram model. The objective is to let specific models compensate for some weaknesses of n-gram models, which cannot capture sparse and rare phenomena, nor patterns that do not occur at all in the the training corpus. This approach hence opens new interesting perspectives in particular for domain adaptation.

6.2.1.4. Speech recognition for interaction in virtual worlds

Automatic speech recognition was investigated for vocal interaction in virtual worlds, in the context of serious games in the EMOSPEECH project. 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 manually corrected and used on addition of available broadcast news corpus. Different language models have then been created using different vocabulary sizes. The acoustic models were 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. A client-server speech recognition demonstrator has been developed. The client runs on an iPad; it records the speech input, sends it to the server, waits for the speech recognition answer, and finally displays the results. The server runs on a PC, relies on the sphinx4 decoder for decoding the received speech signal, and then sends the results to the iPad client.

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. Detailed modeling

Detailed acoustic modeling was further 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, 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 [22]. Combining maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP) adaptation techniques leads to better speech recognition performance, and makes it possible to use more classes [35].

The approach was later improved by introducing a classification process which relies on phonetic acoustic models and the Kullback Leibler divergence measure to build maximally dissimilar clusters. This approach lead to better recognition results than the likelihood based classification approach used in previous experiments [20].

These class-based speech recognition systems were combined with more traditional gender-based system in the ETAPE campaign for the evaluation of speech transcription systems on French radio and TV shows.

6.2.2.2. 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) and has been published in the ICMCS 2012 conference [23].

6.2.3. Speech/text alignment

6.2.3.1. Evaluation of speech/text alignment tools

Speech-text alignment tools are frequently used in speech technology and research: for instance, for training or assessing of speech recognition systems, the extraction of speech units in speech synthesis or in foreign language learning. We designed the software CoALT (Comparing Automatic Labelling Tools) for comparing two automatic labellers or two speech-text alignment tools, ranking them, and displaying statistics about their differences.

The main feature of CoALT is that a user can define its own criteria for evaluating and comparing the speech-text alignment tools since the required quality for labelling depends on the targeted application. Beyond ranking, our tool provides useful statistics for each labeller and above all about their differences and can emphasize the drawbacks and advantages of each labeller. We have applied our software for the French and English languages [19] but it can be used for another language by simply defining the list of the phonetic symbols and optionally a set of phonetic rules.

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.

However, the first step in automatic computer assisted language learning is to check that the pronunced word or utterance corresponds to the expected sentence, otherwise, if the user has not pronunced the correct words it is useless to proceed further with a detailed analysis of the pronunciation to check for possible misspronunciations. In order to decide if the pronunced utterance corresponds to the expected word or sentence, a force phonetic alignment of the sentence is compared to free decoding of the same sentence.

Several comparison features are then defined, such as the number of matching phonemes, the percentage of frames having the save category label, ..., as well as the likelihood ratio. A classifier is then used to decide whether text and speech utterance match or not [36], [28].

These non-native phonetic alignments processes developed in the framework of the ALLEGRO project are currently under implementation in the JSNOORI software, and the processing should be completed by the developpement of automatic feedback procedures.

6.3. Speech-to-Speech Translation and Langage Modeling

Participants: Kamel Smaïli, David Langlois, Sylvain Raybaud, Motaz Saad, Denis Jouvet, Cyrine Nasri.

machine translation, statistical models

Sylvain Raybaud has just defended his thesis untitled “De l’utilisation de mesures de confiance en traduction automatique : évaluation, post-édition et application à la traduction de la parole.”. His contributions are the following: study and evaluation of confidence measures for Machine Translation, an original algorithm to automatically build an artificial corpus with errors for training the confidence measures, development of an entire speech-to-text translation system.

In the scope of Confidence Measures, we participated to the World Machine Translation evaluation campaign (WMT2012 http://www.statmt.org/wmt12/quality-estimation-task.html). More precisely, we proposed a Quality Estimation system to the Quality Estimation shared task. The goal was to predict the quality of translations generated by an automatic system. Each translated sentence is given a score between 1 and 5. The score is obtained using several numerical or boolean features calculated according to the source and target sentences. We perform a linear regression of the feature space against scores in the range [1:5]. To this end, we use a Support Vector Machine. We experiment with two kernels: linear and radial basis function. In our system we use the features from the shared task baseline system and our own features (based on the work from the Sylvain Raybaud’s thesis). This leads to 66 features. To deal with this large number of features, we propose an in-house feature selection algorithm. Our system came 5th among 19 systems. This work was publish in [24]. In the continuation of this research, we contributed to the development of a Quality Estimation tool (quest: https://github.com/lspecia/quest). For that, David Langlois was invited by Lucia Specia at University of Sheffield, Computer Sciences department, Natural Language Processing group. We added our own features into quest. This tool is dedicated to be available for the research community.

Another objective of our research work, with the Cyrine Nasri’s Phd thesis, is to retrieve bilingual phrases for machine translation. As in fact, current statistical machine translation systems usually build an initial word-to-word alignment before learning phrase translation pairs. This operation needs many matching between different single words of both considered languages. We propose a new approach for phrase-based machine translation which does not need any word alignments. It is based on inter-lingual triggers determined by Multivariate Mutual Information. This algorithm segments sentences into phrases and finds their alignments simultaneously. Inspite of the youth of this method, experiments showed that the results are competitive but needs some more efforts in order to overcome the one of state-of-the-art methods.

Another aspect of the research of the group is to work on under resourced language related to Arabic. In fact, in several countries through the Arabic world, only few people speak the modern standard Arabic language. People speak something which is inspired from Arabic but could be very different from the modern standard Arabic. This one is reserved for the official broadcast news, official discourses and so on. The study of dialect is more difficult than any other natural language because it should be noted that this language is not written. A preliminary work has been done knowing that our final objective is to propose a machine translation between the different Arabic dialects and modern standrad Arabic. This issue is very difficult and challenging because no corpus does exist, vernaculars are different even within the same country, etc.

Last, Motaz Saad has started his thesis in November 2011. His objective is to work on opinion analysis in multilingual documents from internet. During this year, he retrieved comparable corpus from the web, and proposed a method to align these corpora at document level. He proposed algorithms to measure the degree of comparability between documents. He submitted his work to the International Conference on Corpus Linguistics (CICL2013).

In the framework of the ETAPE evaluation campaign a new machine learning based process was developed to select the most relevant lexicon to be used for the transcription of the speech data (radio and TV shows). The approach relies on a neural network trained to distinguish between words that are relevant for the task and those that are not. After training, the neural network (NN) is applied to each possible word (extracted from a very large text corpus). Then the words that have the largest NN output score are selected for creating the speech recognition lexicon. Such an approach can handle counts of occurences of the words in various data subsets, as well as other complementary informations, and thus offer more perspectives than the traditional unigram-based selection procedures.

7. Bilateral 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 Contrat Plan Etat-Région (CPER) contract. The CPER MISN TALC, for which Christophe Cerisara is co-responsible, with Claire Gardent, have the objective 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 scientific collaborative operations about text-to-speech alignement, in collaboration with the ATILF laboratory. Automatic alignement procedures are available, and a first version of speech data prosodic structuration has been developed.

7.3. National Contracts

7.3.1. ADT JSnoori

JSnoori ADT (2011-2012) is dedicated to porting main functions of WinSnoori in Java and the integration of new facilities targeting language learning. The main objective is to offer functions enabling the development of feedback for foreign language learning and more precisely the mastery of prosody.

This year the architecture has been changed to comply to the MVC (Model View Controller) model. This makes the management of interactions easier and this clearly separates speech processing algorithms from interactions. In addition forced alignment facilities and phonetic edition tools have been integrated for French and English. They enable the segmentation of sentences uttered by learners, and the annotation with international phonetic alphabet (IPS).

Preliminary versions of diagnosis and feedback of prosody have been incorporated for English (see 6.1.6.1).

7.3.2. 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 from cepstral data as input. We particularly worked on the comparison of cepstral vectors calculated on natural speech and those obtained via the articulatory to acoustic mapping. Bilinear frequency warping was combined with affine adaptation of cepstral coefficients. These two adaptation strategies enable a very good recovery of vocal tract shapes from natural speech. The second topic studied is the access to the codebook. Two pruning strategies, a simple one using the spectral peak corresponding to F2 and a more elaborated one exploiting lax dynamic programming applied on spectral peaks enable a very efficient access to the articulatory codebook used for inversion.

7.3.3. 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.

Currently, we are mainly evaluating the system using both subjective and objective perceptual evaluation.

7.4. International Contracts

7.4.1. 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.

7.4.2. 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 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 6.1.5.

8. Partnerships and Cooperations

8.1. European Initiatives

8.1.1. Collaborations in European Programs, except FP7

8.1.1.1. Allegro

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.

In the first part of the project we have investigated the phonetic segmentation of non-native speech and analyzed the precision of the phoneme boundaries as boundaries are critical for making duration-based diagnoses in computer assisted learning of the prosody of a foreign language. The experiments have shown that it is critical to include non-native pronunciation variants in the pronunciation lexicon used for forced alignment. However it is better to avoid introducing unusual variants. The best performance was achieved by introducing variants that were seen at least two times on some development non-native data set. A detailed analysis of the boundary precision was also carried out. It was observed that a good precision was achieved for boundaries between some classes of phonemes (as for example between plosives and vowels, fricatives and vowels, and so on). Hence such information should be taken into account either in choosing the words when designing the exercises, and/or in the diagnosis process.

During this year, a special attention was paid to checking the consistency of the recorded speech signal with the expected text. The goal behind that, is to detect speech utterances that do not match with the expected text because of learner's inattention (not pronouncing the expected words) or acquisition problems (truncation of the speech acquisition - the beginning or the end of the sentence is missing - or background noise troubles). In case of mismatch, no further processing is to be carried on; on the opposite, when the speech utterance matches the expected text, prosodic features will be analyzed in details in order to provide a prosodic diagnosis of the pronunciation and the adequate feedback. In order to detect a possible mismatch, several criteria are computed based on the comparison of the phonetic segmentation resulting from a forced alignment with the phonetic segmentation obtained with a phonetic-loop or with a word-loop grammar; these criteria are then combined by a classifier to decide if the speech utterance and the expected text matches or not (cf. section 6.2.3.2).

The automatic phonetic segmentation has been included in the JSNOORI software (cf. section 5.2), as well as other extensions specific to handling exercises for learning the prosody of a foreign language.

The detection of the fundamental frequency (F0) is a key aspect of tools developed for learning prosody of a foreign language. Errors in F0 detection compromise the diagnosis set about the learner's utterance and the modifications of the prosody as well. Since no method alone can be sufficiently robust we thus investigated the combination of three methods, Yin, the method proposed by de Cheveigné et al., an autocorrelation method and a spectral comb method already developed withinh JSnoori. The three methods were redeveloped in Matlab and combined with a neural network approach.

8.1.1.2. Emospeech

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.

With respect to the development of a speech recognition solution, a prototype was developed in the framework of a serious game, in collaboration with the Talaris team. The speech-based prototype, which relies on the Sphinx4 speech recognition engine, has made possible the collection of speech material, that has later been transcribed. Specialized lexicons have been developed by combining the task-specific vocabulary extracted from the documentation of the serious game, from the speech data collected using the prototype, and from the text data collected by the Talaris team using a text-based prototype, with the most frequent words selecting in broadcast new corpus. Acoustic models have also been adapted using collected speech material.

Parallel to this work, a client/server speech recognizer system has been developed. The client was developed to run on an iPad terminal. Its role mainly consists in recording the speech signal, sending it to the server, waiting for the speech recognition answer, and finally displaying the speech recognition results. The server, runs on a PC, and performs the actual speech recognition task.

9. Dissemination

9.1. Scientific Animation

- The members of the team frequently review articles and papers for Journal of Phonetics, JASA, Acta Acoustica, Computer Speech and Language, Speech communication, TAL, IEEE Journal of Selected Topics in Signal Processing, IEEE Transaction of Information Theory, IEEE Signal Processing Letters, 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, J.P. Haton)
  - JEP, (I. Illina)
- Member of organisational conference committees :
  - AVSP 2013, (S. Ouni)
- Coordinator of French part of ERASMUS Intensif Programme : Learning Computer Programming in Virtual Environment (I. Illina)
- Reviewer of projects in the European FET (Future and Emerging Technologies) (Y. Laprie).
- Member of “Conseil scientifique IRCOM Consortium corpus oraux et multimodaux” (D. Fohr)
- Member of “Association Française pour la Communication Parlée” (French Association for Oral Communication) board (D. Langlois)
- 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)
- Member of the Scientific Committee of Télécom- Bretagne, Brest (Jean-Paul Haton)
- The members of the team have been invited as lecturer:
  - 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.
  - David Langlois has presented the Machine Translation domain to students in University of Tunis, in 2012 november.
  - Anne Bonneau has presented “speech sounds” at the 5e congrès jeunes chercheurs "Chamboule tes sens" Nancy, March 2012.
  - Jean-Paul Haton : “Introduction à l’intelligence artificielle”, Ecole Doctorale SIMEM, Caen, 24/04/2012
  - Jean-Paul Haton : “Intelligence artificielle et langue : état des lieux”, Journée “IA et TALN” AFIA-ATALA, Paris, 12/03/2012
  - Jean-Paul Haton : “Automatic Speech Recognition ; past, present and future”, Invited Keynote Speech, EUSIPCO, Bucharest, 28/08/2012

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

The professors and associate professors of the team are teaching at Lorraine University.

Teaching activities in relation with the team research domains:

- Analyse, Traitement et reconnaissance de la parole, 30HETD, M1, Université de Lorraine (V. Colotte and D. Langlois),
- Natural Langue Processing (in English), 24HETD,
- M2 Erasmus Mundus, université de Lorraine (K. Smaï),
- école d’orthophonie, faculté de médecine, Université de Lorraine (Agnes Piquard-Kipffer and Anne Bonneau),
- école d’audioprothèse, facaulté de pharmacie, Université de Lorraine (Anne Bonneau),

In addition to courses, we highlight the following activities:

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

9.2.2. Supervision

Phd : Sylvain Raybaud, “De l’utilisation de mesures de confiance en traduction automatique : évaluation, post-édition et application à la traduction de la parole”, Lorraine University, Dec 2012, David Langlois and Kamel Smaïli

Phd : Christian Gillot, “Modèles de langue exploitant la similarité structurelle entre séquences pour la reconnaissance de la parole”, Lorraine University, Dec 2012, Christophe Cerisara and Jean-Paul Haton

PhD in progress : Utpala Musti, Acoustic-Visual Synthesis using bimodal selection, September 2009, S. Ouni and V. Colotte

PhD in progress : Julie Busset, Inversion acoustique articulatoire à partir de coefficients cepstraux, to be defended in February 2013

PhD in progress : Motaz Saad, “Etude de comparabilité de corpus multilingues et analyse émotionnelle de leurs contenus”, Kamel Smaïli and David Langlois

PhD in progress : Fadoua Bahja, Détection du Fondamental de la Parole : Application à la Voix Pathologique, from Mai 2009, Joseph Di Martino and Elhassane Ibn Elhaj.

PhD in progress : Othman Lachhab, Reconnaissance de la Parole Continue : Application au Rehaussement de la Voix Pathologique, from November 2010, Joseph Di Martino and Elhassane Ibn Elhaj.

PhD in progress : Cyrine Nasri, “Une alternative à l’alignement standard des séquences de traduction”, Kamel Smaïli

PhD in progress : Arseniy Gorin, Handling trajectories and speaker consistency in automatic speech recognition, October 2011, D. Jouvet.

PhD in progress : Alex Mesnil, “Modliésation bayésienne hiérarchique et parcimonieuse pour le langage naturel”, Emmanuel Vincent and Kamel Smaïli

PhD in progress : Dung Tran, Uncertainty handling for noise-robust automatic speech recogniton, December 2012, E. Vincent and D. Jouvet.

PhD in progress : Luiza Orosanu, Speech recognition for communication help for deaf or hard of hearing people, December 2012, D. Jouvet.

9.2.3. Juries

Participation in PhD thesis Jury for Fethi Bougares (Maine University, November 2012), D. Jouvet, reviewer.

Participation in PhD thesis Jury for Camille Fauth (Strasbourg University, December 2012), Y. Laprie, reviewer.

Participation in Selection committee at ENSSAT Rennes University, may 2012 (V. Colotte, O. Mella, Y. Laprie).

9.3. Popularization

the movie “Je peux voir les mots que tu dis” has been selected for the “13ème Festival du Film de Chercheur”

Participation in TV broadcast news (FR3) about voice forensic examinations (Y. Laprie).

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 Journals

[9] G. Bouselmi, D. Fohr, I. Illina. *Multilingual Recognition of Non-Native Speech using Acoustic Model Transformation and Pronunciation Modeling*, in "International Journal of Speech Technology", June 2012, vol. 15, n^{o} 2, p. 203 - 213, http://hal.archives-ouvertes.fr/hal-00764626.

[10] M. Dargnat, V. Colotte, K. Bartkova, A. Bonneau. *Continuations intra- et interphrastiques du français : premiers résultats expérimentaux*, in "SHS Web of Conferences", July 2012, vol. 1, p. 1471-1485, Cette revue reprend les articles sélectionnés par le comité de lecture du 3e Congrès Mondial de Linguistique Française. [*DOI :* 10.1051/shsconf/20120100142], http://hal.inria.fr/hal-00764639.

[11] A. Piquard-Kipffer. *Prédire dès l'âge de 5 ans le niveau de lecture de fin de cycle 2. Suivi de 85 enfants de langue maternelle française de 4 à 8 ans.*, in "L'information grammaticale", March 2012, n^{o} 133, p. 20-26, http://hal.inria.fr/hal-00681684.

[12] I. Steiner, K. Richmond, I. Marshall, C. Gray. *The Magnetic Resonance Imaging subset of the mngu0 articulatory corpus*, in "Journal of the Acoustical Society of America", February 2012, vol. 131, n^{o} 2, p. 106-111, Author version contains correctly encoded (Unicode) fonts and attached multimedia content. [*DOI :* 10.1121/1.3675459], http://hal.inria.fr/hal-00661082.

International Conferences with Proceedings

[13] F. Bahja, J. Di Martino, E. H. Ibn Elhaj, D. Aboutajdine. *Prediction of Cepstral Excitation Pulses for Voice Conversion*, in "5th. International Conference on Information Systems and Economic Intelligence - SIIE2012", Djerba, Tunisie, 2012, http://hal.inria.fr/hal-00761776.

[14] F. Bahja, E. H. Ibn Elhaj, J. Di Martino. *On the Use of Wavelets and Cepstrum Excitation for Pitch Determination in Real-Time*, in "3rd International Conference on Multimedia Computing and Systems - ICMCS'12", Tangier, Maroc, 2012, http://hal.inria.fr/hal-00761819.

[15] K. Bartkova, A. Bonneau, V. Colotte, M. Dargnat. *Productions of "continuation contours" by French speakers in L1 (French) and L2 (English)*, in "Speech Prosody", Shangai, Chine, May 2012, vol. 1, p. 426-429, http://hal.inria.fr/hal-00763919.

[16] J. Busset, M. Cadot. *Démêler les actions des articulateurs en jeu lors de la production de parole avec le logiciel C.H.I.C. : Analyse de séquences de radiographies de la tête.*, in "ASI6", Caen, France, 2012, p. 284-298, http://hal.inria.fr/hal-00759054.

[17] M. Chami, J. Di Martino, L. Pierron, E. H. Ibn Elhaj. *Real-Time Signal Reconstruction from Short-Time Fourier Transform Magnitude Spectra Using FPGAs*, in "5th. International Conference on Information Systems and Economic Intelligence - SIIE2012", Djerba, Tunisie, 2012, http://hal.inria.fr/hal-00761783.

[18] M. Embarki, S. Ouni, F. Salam. *Clarté de la parole et effets coarticulatoires en arabe standard et dialectal*, in "Actes de la conférence conjointe JEP-TALN-RECITAL 2012", Grnoble, France, ATALA-AFCP (editor), June 2012, vol. 1:JEP, p. 209-216, http://hal.inria.fr/hal-00762581.

[19] D. Fohr, O. Mella. *CoALT: A Software for Comparing Automatic Labelling Tools*, in "Language Resources and Evaluation LREC 2012", Istanbul, Turquie, May 2012, p. 325-328, http://hal.inria.fr/hal-00761781.

[20] A. Gorin, D. Jouvet. *Class-based speech recognition using a maximum dissimilarity criterion and a tolerance classification margin*, in "SLT 2012 - 4th IEEE Workshop on Spoken Language Technology", Miami, États-Unis, December 2012, http://hal.inria.fr/hal-00753454.

[21] D. Jouvet, D. Fohr, I. Illina. *Evaluating grapheme-to-phoneme converters in automatic speech recognition context*, in "ICASSP - 2012 - IEEE International Conference on Acoustics, Speech and Signal Processing

(ICASSP)", Kyoto, Japon, March 2012, p. 4821 - 4824 [*DOI :* 10.1109/ICASSP.2012.6288998], http://hal.inria.fr/hal-00753364.

[22] D. Jouvet, N. Vinuesa. *Classification margin for improved class-based speech recognition performance*, in "ICASSP - 2012 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", Kyoto, Japon, March 2012, p. 4285 - 4288 [*DOI :* 10.1109/ICASSP.2012.6288866], http://hal.inria.fr/hal-00753345.

[23] O. Lachhab, J. Di Martino, E. H. Ibn Elhaj, A. Hammouch. *Real Time Context-Independent Phone Recognition Using a Simplified Statistical Training Algorithm*, in "3rd International Conference on Multimedia Computing and Systems - ICMCS'12", Tangier, Maroc, 2012, http://hal.inria.fr/hal-00761816.

[24] D. Langlois, S. Raybaud, K. Smaïli. *LORIA System for the WMT12 Quality Estimation Shared Task*, in "The Seventh Workshop on Statistical Machine Translation - NAACL 2012", Montréal, Canada, Association for Computational Linguistics, June 2012, p. 114–119, http://hal.inria.fr/hal-00726372.

[25] K. Meftouh, N. Bouchemal, K. Smaïli. *A Study of a Non-Resourced Language: The Case of one of the Algerian Dialects*, in "The third International Workshop on Spoken Languages Technologies for Under-resourced Languages - SLTU'12", Cape-town, Afrique Du Sud, May 2012, –, http://hal.inria.fr/hal-00727042.

[26] U. Musti, C. Lavecchia, V. Colotte, S. Ouni, B. Wrobel-Dautcourt, M.-O. Berger. *ViSAC : Acoustic-Visual Speech Synthesis: The system and its evaluation*, in "FAA: The ACM 3rd International Symposium on Facial Analysis and Animation", Vienne, Autriche, September 2012, -, http://hal.inria.fr/hal-00762568.

[27] C. Nasri, K. Smaïli, C. Latiri, Y. Slimani. *A new method for learning Phrase Based Machine Translation with Multivariate Mutual Information*, in "The 8th International Conference on Natural Language Processing and Knowledge Engineering - NLP-KE'12", HuangShan, Chine, September 2012, ..., http://hal.inria.fr/hal-00727044.

[28] L. Orosanu, D. Jouvet, D. Fohr, I. Illina, A. Bonneau. *Combining criteria for the detection of incorrect entries of non-native speech in the context of foreign language learning*, in "SLT 2012 - 4th IEEE Workshop on Spoken Language Technology", Miami, États-Unis, December 2012, http://hal.inria.fr/hal-00753458.

[29] S. Ouni, L. Mangeonjean. *VisArtico : visualiser les données articulatoires obtenues par un articulographe*, in "Actes de la conférence conjointe JEP-TALN-RECITAL 2012", Grenoble, France, June 2012, vol. 1:JEP, p. 129-135, http://hal.inria.fr/hal-00762575.

[30] S. Ouni, L. Mangeonjean, I. Steiner. *VisArtico: a visualization tool for articulatory data*, in "13th Annual Conference of the International Speech Communication Association - InterSpeech 2012", Portland, OR, États-Unis, September 2012, http://hal.inria.fr/hal-00730733.

[31] C. Savariaux, P. Badin, S. Ouni, B. Wrobel-Dautcourt. *Étude comparée de la précision de mesure des systèmes d'articulographie électromagnétique 3D : Wave et AG500*, in "29e Journées d'Études sur la Parole (JEP-TALN-RECITAL'2012)", Grenoble, France, ATALA-AFCP (editor), 2012, p. 513-520, http://hal.inria.fr/hal-00724682.

[32] I. Steiner, S. Ouni. *Artimate: an articulatory animation framework for audiovisual speech synthesis*, in "Workshop on Innovation and Applications in Speech Technology", Dublin, Irlande, J. Kane (editor), UCD, TCD, March 2012, http://hal.inria.fr/hal-00678964.

[33] I. Steiner, K. Richmond, S. Ouni. *Using multimodal speech production data to evaluate articulatory animation for audiovisual speech synthesis*, in "3rd International Symposium on Facial Analysis and Animation - FAA 2012", Vienna, Autriche, September 2012, http://hal.inria.fr/hal-00734464.

National Conferences with Proceeding

[34] I. Illina, D. Fohr, D. Jouvet. *Génération des prononciations de noms propres à l'aide des champs aéatoires conditionnels*, in "JEP-TALN-RECITAL 2012", Grenoble, France, June 2012, http://hal.inria.fr/hal-00753381.

[35] D. Jouvet, A. Gorin, N. Vinuesa. *Exploitation d'une marge de tolérance de classification pour améliorer l'apprentissage de modèles acoustiques de classes en reconnaissance de la parole*, in "JEP-TALN-RECITAL 2012", Grenoble, France, June 2012, p. 763-770, http://hal.inria.fr/hal-00753394.

[36] L. Orosanu, D. Jouvet, D. Fohr, I. Illina, A. Bonneau. *Détection de transcriptions incorrectes de parole non-native dans le cadre de l'apprentissage de langues étrangères*, in "JEP-TALN-RECITAL 2012", Grenoble, France, June 2012, http://hal.inria.fr/hal-00753387.

References in notes

[37] 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.

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

[39] 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.

[40] M. Cohen, D. Massaro. *Modeling coarticulation in synthetic visual speech*, 1993.

[41] 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/.

[42] 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.

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

[44] 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, 392, I.: Computing Methodologies/I.2: ARTIFICIAL INTELLIGENCE, I.: Computing Methodologies/I.5: PATTERN RECOGNITION, http://hal.inria.fr/inria-00105908/en/.

[45] 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, 277, http://hal.inria.fr/inria-00185312/en/.

[46] 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/.

[47] 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.

[48] 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.

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

[50] 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/.

[51] 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/.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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/.