Project-Team EXPRESSION

Expressiveness in Human Centered Data/Media

Vannes-Lannion-Lorient

Activity Report

2013

Contents

Contents

1 Team

Head of the team
Pierre-François Marteau, Professor, Université de Bretagne Sud

Administrative assistant
Sylviane Boisadan

Université de Bretagne Sud
Nicolas Béchet, Assistant Professor
Giuseppe Bério, Professor
Sylvie Gibet, Professor
Gildas Ménier, Assistant Professor
Jeanne Villaneau, Assistant Professor
Jean-François Kamp, Assistant Professor

Université de Rennes1
Nelly Barbot, Assistant Professor
Arnaud Delhay, Assistant Professor
Gwénolé Lecorvé, Assistant Professor
Damien Lolive, Assistant Professor

PhD students
Pamela Carreno, Université de Bretagne Sud, ANR INGREDIBLE, first year
Lei Chen, Université de Bretagne Sud/Univ. McGill, ARED, first year
David Guennec, Université Rennes 1, 2nd year
Hai Hieu Vu, Université de Bretagne Sud
Raheel Qader, Université de Rennes 1, 1st year
Guyao Ke, Université de Bretagne Sud, 4th year
Thibault Le-Naour, Université de Bretagne Sud, 3rd year

Master students

Associate members

Vincent Barreaud, Assistant Professor
Farida Said, Assistant Professor
Jonathan Chevelu, IGR, 09/2012, ANR Phorevox
Ludovic Hamon, post-doc, 09/2012, projet SIGN3D
Caroline Larboulette, ATER, 10/2012, Université de Bretagne Sud

2 Overall Objectives

2.1 Overview

Expressivity or expressiveness are terms which are often used in a number of domains. In biology, they relate to genetics and phenotypes, whereas in computer science, expressivity of programming languages refers to the ability to formalize a wide range of concepts. When it comes to human expressivity, we will consider the following reading: expressivity is the way a human being conveys emotion, style or intention. Considering this definition, the Expression team focuses on studying human language data conveyed by different media: gesture, speech and text. Such data exhibit an intrinsic complexity characterized by the intrication of multidimensional and sequential features. Furthermore, these features may not belong to the same representation levels - basically, some features may be symbolic (e.g., words, phonemes, etc.) whereas others are digital (e.g., positions, angles, sound samples) - and sequentiality may result from temporality (e.g., signals).

Within this complexity, human language data embed latent structural patterns on which meaning is constructed and from which expressiveness and communication arise. Apprehending this expressiveness, and more generally variability, in multidimensional time series, sequential data and linguistic structures is the main proposed agenda of Expression. This main purpose comes to study problems for representing and characterizing heterogeneity, variability and expressivity, especially for pattern identification and categorization.

The proposed research project targets the exploration and (re)characterization of data processing models in three mediated contexts:

1. Expressive gesture analysis, synthesis and recognition,
2. Expressive speech analysis and synthesis,
3. Expression in text and language.

2.2 Key Issues

In its current configuration, EXPRESSION addresses the following key issues:

- **Data acquisition**: gesture, speech or text data are characterized by high levels of heterogeneity and variability. Studying such media requires using high quality data sets suited for a well defined and dedicated task. The data acquisition process is thus a crucial

step since it will conditioned the outcomes of the team research, from the characterization of the studied phenomena, to the quality of the data driven model that will be extracted and to the assessment of the developed applications. The production of high quality and focused corpora is thus a main issue for our research communities.

- **Multi-level representations**: we rely on multi-level representations (semantic, phonological, phonetic, signal processing) to organize and apprehend data. The heterogeneity of these representations (from metadata to raw data) prevents us from using standard modeling techniques that rely on homogeneous features. Building new multi-level representations is thus a first research direction. Such representations will provide efficient information access, support for database enrichment through boostrapping and automatic annotation;

- **Data mining and knowledge extraction**: this research addresses data processing (indexing, filtering, retrieving, clustering, classification, recognition) through the development of distances or similarity measures, rule-based or pattern-based models, and machine learning methods. The developed methods will tackle symbolic data levels (semantic, lexical, etc.) or time series data levels (extraction of segmental units or patterns from dedicated databases);

- **Generation**: we are also interested in the automatic generation of high-quality content reproducing human behavior on two modalities (gesture and speech). In particular, to guarantee the adequate expressivity, the variability of the output has to be finely controlled. For gesture, statements and actions can be generated from structural models (composition of gestures in French sign language (LSF) from parameterized linguistic units). For speech, classical approaches are data-driven and rely either on speech segment extraction and combination, or on the use of statistical generation models. In both cases, the methods are based at the same time on data-driven approaches and on cognitive and machine learning control processes (e.g. neuromimetic;

- **Use cases and evaluation**: the objective is to develop intuitive tools and in particular sketched-based interfaces to improve or facilitate data access (using different modes of indexing, access content, development of specific metrics, and graphical interfaces), and to integrate our aforementioned models into these tools. Furthermore, whereas many encountered sub-problems are machine learning tasks that can be automatically evaluated, synthesizing human-like data requires final perceptive (i.e., manual) evaluations. Such evaluations are costly and developing automatic methodologies to simulate them is a major challenge. In particular, one axis of research directly concerns the development of cross-disciplinary evaluation methodologies.

3 Scientific Foundations

3.1 Gesture analysis, synthesis and recognition

Keywords:

.

Gesture communication and expression thanks to advanced technologies such as new sensors, mobile devices, or specialized interactive systems, have brought a new dimension to a broad range of applications never before experienced, such as entertainments, pedagogical and artistic applications, rehabilitation, etc. The study of gestures requires more and more understanding of the different levels of representation that underly their production, from meanings to motion performances characterized by high-dimensional time-series data. This is even more true for skilled and expressive gestures, or for communicative gestures, involving high level semiotic and cognitive representations, and requiring extreme rapidity, accuracy, and physical engagement with the environment.

Many previous works have studied movements and gestures that convey a specific meaning, also called semiotic gestures. In the domain of co-verbal gestures, Kendon [Ken80] is the first author to propose a typology of semiotic acts. McNeil extends this typology with a theory gathering the two forms of expression, speech and action [McN92]. In these studies, both modalities are closely linked, since they share a common cognitive representation. Our research objectives focus more specifically on body movements and their different forms of variations in nonverbal communication or bodily expression. We consider more specifically full-body voluntary movements which draw the user's attention, and express through body language some meaningful intent, such as sign language or theatrical gestures. Generally, these movements are composed of multimodal actions that reveal a certain expressiveness, whether unintentional of deliberate.

Different qualitative aspects of expressiveness have already been highlighted in motion. Some of them rely on the observation of human motion, such as those based on the Laban Movement Analysis theory, in which the expressiveness is essentially contained into the Effort and Shape components [Mal87]. Motion perception through bodily expressions has also given rise to many works in nonverbal communication. In the psychology and neuroscience literature, recent studies have focused in particular on the recognition of emotion in whole

[Ken80] A. Kendon, "Gesticulation and speech Two aspects of the process of utterance", *in : The Relation Between Verbal and Nonverbal Communication*, p. 207–227, 1980.

[McN92] D. McNeill, *Hand and Mind - What Gestures Reveal about Thought*, The University of Chicago Press, Chicago, IL, 1992.

[Mal87] V. Maletik, *Body, Space, Expression : The Development of Rudolf Laban?s Movement and Dance Concepts*, Walter de Gruyter Inc., 1987.

body movements [Wal98,Gal09,THB06,dG06,CG07].

In computation sciences, many researches have been conducted to synthesize expressive or emotional states through the nonverbal behavior of expressive virtual characters. Two major classes of approaches can be distinguished: those that specify explicit behaviors associated with pure synthesis techniques, or those offering data-driven animation techniques. In the first category we find embodied conversational agents (ECAs) that rely on behavioral description languages [KW04], or on sets of expressive control parameters [CCZB00,HMBP05]. More recently, some computational models consider the coordination and adaptation of the virtual agent with a human or with the environment in interacting situations. The models in such cases focus on rule-based approaches derived from social communicative theories [Pel09,Kop10]. In the second category, motion captured data is used with machine learning techniques to capture style in motion and generate new motion with variations in style [BH00,Her03,GMHP04,HPP05]. In these works authors consider a low-level definition of style, in terms of variability observed among several realizations of the same gesture. If some relevant works rely on qualitative

[Wal98] H. Wallbott, "Bodily expression of emotion", *in : European Journal of Social Psychology*, *28*, p. 879–896, 1998.

[Gal09] P. Galaher, "Individual differences in nonverbal behavior: dimensions of style", *in : Journal of Personality and Social Psychology*, *51*, 1, p. 133–145, 2009.

[THB06] L. Torresani, P. Hackney, C. Bregler, "Learning motion style synthesis from perceptual observations", *in : Advances in Neural Information Processing Systems*, p. 1393–1400, 2006.

[dG06] B. de Gelder, "Toward a biological theory of emotional body language", *Biological Theory 1*, 2006, p. 130–132.

[CG07] E. Crane, M. Gross, *Motion Capture and Emotion: Affect Detection in Whole Body Movement*, *Affective Computing and Intelligent Interaction, ACII, Lecture Notes in Computer Science*, Springer Verlag, 2007, In Proc. of ACII,.

[KW04] S. Kopp, I. Wachsmuth, "Synthesizing multimodal utterances for conversational agents", *Journal of Visualization and Computer Animation 15*, 1, 2004, p. 39–52.

[CCZB00] D. Chi, M. Costa, L. Zhao, N. Badler, "The EMOTE model for effort and shape", *in : SIGGRAPH'00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, ACM Press/Addison-Wesley Publishing Co., p. 173–182, 2000.

[HMBP05] B. Hartmann, M. Mancini, S. Buisine, C. Pelachaud, "Design and evaluation of expressive gesture synthesis for embodied conversational agents", *in : AAMAS*, p. 1095–1096, 2005.

[Pel09] C. Pelachaud, *Studies on Gesture Expressivity for a Virtual Agent*, *63*, 1, 2009.

[Kop10] S. Kopp, "Social resonance and embodied coordination in facetoface conversation with artificial interlocutors", *Speech Communication 52*, 6, 2010, p. 587–597.

[BH00] M. Brand, A. Hertzmann, "Style machines", *in : ACM SIGGRAPH 2000*, p. 183–192, 2000.

[Her03] A. Hertzmann, "Machine learning for computer graphics: A manifesto and tutorial", *in : Computer Graphics and Applications, 2003. Proceedings. 11th Pacific Conference on*, IEEE, p. 22–36, 2003.

[GMHP04] K. Grochow, S. L. Martin, A. Hertzmann, Z. Popović, "Style-based inverse kinematics", *ACM Transactions on Graphics 23*, 3, 2004, p. 522–531.

[HPP05] E. Hsu, K. Pulli, J. Popović, "Style translation for human motion", *in : ACM Transactions on Graphics (TOG)*, *24*, 3, ACM, p. 1082–1089, 2005.

or quantitative annotations of motion clips (*e.g.* [AI06,MBS09]), or propose relevant methods to create a repertoire of expressive behaviors (*e.g.* [**?**]), very few approaches deal with both motion-captured data and their implicit semantic and expressive content.

In our approach, we will consider that gesture is defined as expressive, meaningful bodily motion. It combines multiple elements which intrinsically associate *meaning*, *style*, and *expressiveness*. The *meaning* is characterized by a set of signs that can be linguistic elements or significant actions. This is the case when gestures are produced in the context of narrative scenarios, or expressing utterances in sign languages. The *style* includes both the identity of the subject, determined by the morphology of the skeleton, the gender, the personality, and the way the motion is performed, according to some specific task (e.g. moving in a graceful or jerky way). The *expressiveness* characterizes the nuances that are superimposed on motion, guided by the emotional state of the actor, or associated to some willful intent. For example, theatrical performances may contain intentional emphasis that are accompanied by effects on the movement kinematics or dynamics. Most of the time, it is very difficult to separate all these components, and the resulting movements give rise to different physical realizations characterized by some variability that can be observed into the raw motion data and subsequently characterized. For simplicity we will assume later that the notion of expressiveness includes all forms of variability.

Our line of research focuses specifically on the study of variability and variation in motion captured data, linked to different forms of expressiveness, or to the sequencing of semantic actions according to selected scenarios. Motion capture is used for retrieving relevant features that encode the main spatio-temporal characteristics of gestures: low-level features are extracted from the raw data, whereas high-level features reflect structural patterns encoding linguistic aspects of gestures [ACD+09a]. Many data-driven synthesis model have been developed in order to re-use or modify motion capture data and therefore produce new motions with all the realism and nuances present in the examples. We focus in our approach on machine learning methods that capture all the subtleties of human movement and generate more expert gestures while maintaining the style, expressiveness and semantic inherent to human actions [Her03,AI06,HCGM06,PP10]. One of the novelties of our approach is that it is conducted through an

[AI06] O. Arikan, L. Ikemoto, *Computational Studies of Human Motion: Tracking and Motion Synthesis*, Now Publishers Inc, 2006.

[MBS09] M. Müller, A. Baak, H.-P. Seidel, "Efficient and Robust Annotation of Motion Capture Data", *in : Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, p. 17–26, August 2009.

[ACD+09a] C. Awad, N. Courty, K. Duarte, T. L. Naour, S. Gibet, "A Combined Semantic and Motion Capture Database for Real-Time Sign Language Synthesis", *in : IVA*, p. 432–438, 2009.

[Her03] A. Hertzmann, "Machine learning for computer graphics: A manifesto and tutorial", *in : Computer Graphics and Applications, 2003. Proceedings. 11th Pacific Conference on*, IEEE, p. 22–36, 2003.

[HCGM06] A. Heloir, N. Courty, S. Gibet, F. Multon, "Temporal alignment of communicative gesture sequences", *Journal of Visualization and Computer Animation 17*, 3-4, 2006, p. 347–357.

[PP10] T. Pejsa, I. S. Pandzic, "State of the Art in Example-Based Motion Synthesis for Virtual Characters in Interactive Applications", *in : Computer Graphics Forum*, *29*, 1, Wiley Online

analysis / synthesis scheme, corrected and refined through an evaluation loop (*e.g.* [GMD12]). Consequently, the data-driven models, which incorporate constraints derived from observations, should significantly improve the quality and the credibility of the gesture synthesis ; furthermore, the analysis of original or synthesized data through classification or recognition models should refine the prior hypothesis. Finally, evaluation takes place at different levels in the analysis / synthesis loop, and is performed qualitatively or quantitatively through the definition of original use cases.

3.2 Speech processing and synthesis

Keywords: Speech synthesis, unit selection, HMM-based synthesis, prosody, phonology.

Based on a textual input, a Text-To-Speech (TTS) system produces a speech signal that corresponds to a vocalization of the given text [All76,Tay09]. Classically, this process can be decomposed into two steps. The first one realizes a sequence of linguistic treatments on the input text, especially syntactical, phonological and prosodic analysis. These treatments give as output a phoneme sequence enriched by prosodic tags. The second step is then the signal generation from this symbolic information.

In this framework, two concurrent methodological approaches are opposed: corpus-based speech synthesis [Bre92,Dut97], and statistical parametric approach, mainly represented by the HMM-based TTS system called HTS [MTKI96,TZ02,ZTB09]. Corpus-based speech synthesis consists in the juxtaposition of speech segments chosen in a very large speech database in order to obtain the best possible speech quality. On the other hand, HTS, which is more recent, consists in modeling the speech signal by using stochastic models whose parameters are estimated *a priori* on a training corpus. These models are then used in a generative way so as to create a synthetic speech signal from a given parametric description.

Corpus-based speech synthesis is a reference since at least a decade. Examples of systems

Library, p. 202–226, 2010.

[GMD12] S. Gibet, P.-F. Marteau, K. Duarte, "Toward a Motor Theory of Sign Language Perception", *Human-Computer Interaction and Embodied Communication, GW 2011 7206*, 2012, p. 161–172.

[All76] J. Allen, "Synthesis of speech from unrestricted text", *Proceedings of the IEEE 64*, 4, 1976, p. 433–442.

[Tay09] P. Taylor, *Text-to-speech synthesis*, *1*, Cambridge University Press, Cambridge UK, 2009.

[Bre92] A. Breen, "Speech synthesis models: a review", *Electronics & communication engineering journal 4*, 1, 1992, p. 19–31.

[Dut97] T. Dutoit, "High-quality text-to-speech synthesis: An overview", *Journal of Electrical and Electronics Engineering 17*, 1, 1997, p. 25–36.

[MTKI96] T. Masuko, K. Tokuda, T. Kobayashi, S. Imai, "Speech synthesis using HMMs with dynamic features", *in : Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *1*, IEEE, p. 389–392, 1996.

[TZ02] K. Tokuda, H. Zen, "An HMM-based speech synthesis system applied to English", *in : Speech Synthesis, 2002.*, 2002.

[ZTB09] H. Zen, K. Tokuda, A. W. Black, "Statistical parametric speech synthesis", *Speech Communication 51*, 11, 2009, p. 1039–1064.

using this technique are ATR Chatr [BT94], CMU Festival [TBC98], Microsoft Whistler [HAA+96], IBM ViaVoice Text-To-Speech [PBE+06], AT&T Natural Voices [JS02], Loquendo TTS [QDMS01], Microsoft text-to-speech voices, ATR XIMERA [KTN+04], Acapela Voice, Voxygen [Vox13] and IrcamTTS [BSHR05]. This technique relies on systems conceived in the 80s/90s. In particular, diphone speech synthesis has shown to produce results of variable quality but generally with high intelligibility. Corpus-based synthesis can be viewed as an extension of this approach, by allowing multiple instances of the units and also variable length units to be considered during a selection step. The problem then turns into finding the optimal acoustic unit sequence. This selection is generally done via a dynamic programming approach such as the Viterbi algorithm. In particular, common implementation of this algorithm, [TBC98,Don98,BBd02], try to minimize the audible distortions at junctions between units as well as distances to prosodic and phonological targets.

Restituted timber quality, which is judged very near to natural, is the main reason of corpus-based speech synthesis success. Another reason is certainly the overall good intelligibility of the synthesized utterances [MA96]. Nevertheless, the main limitation is the lack of expressiveness. Generally, synthesized voices only have a neutral melody without any con-

[BT94] A. W. Black, P. Taylor, "CHATR: a generic speech synthesis system", *in : Proceedings of the 15th conference on Computational linguistics*, p. 983—-986, 1994.

[TBC98] P. Taylor, A. Black, R. Caley, "The architecture of the Festival speech synthesis system", *in : The Third ESCA Workshop in Speech Synthesis*, Citeseer, p. 147–151, 1998.

[HAA+96] X. Huang, A. Acero, J. Adcock, H.-W. Hon, J. Goldsmith, J. Liu, M. Plumpe, "Whistler: a trainable text-to-speech system", *in : International Conference on Spoken Language Processing*, p. 2387–2390, 1996.

[PBE+06] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, M. A. Picheny, "The IBM expressive text-to-speech synthesis system for American English", *IEEE Transactions on Audio, Speech and Language Processing 14*, 4, July 2006, p. 1099–1108.

[JS02] M. Jilka, A. K. Syrdal, "The AT&T german text-to-speech system: realistic linguistic description", *in : International Conference on Spoken Language Processing*, 2002.

[QDMS01] S. Quazza, L. Donetti, L. Moisa, P. L. Salza, "Actor: a multilingual unit-selection speech synthesis system", *in : ISCA ITRW on Speech Synthesis*, p. 209, 2001.

[KTN+04] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, K. Tokuda, "Ximera: a new tts from atr based on corpus-based technologies", *in : ISCA ITRW on Speech Synthesis*, p. 179–184, 2004.

[Vox13] Voxygen, "Voxygen Online Speech Synthesis System, http://voxygen.fr", 2013, `http://voxygen.fr/`.

[BSHR05] G. Beller, D. Schwarz, T. Hueber, X. Rodet, "A hybrid concatenative synthesis system on the intersection of music and speech", *in : Proceedings of Journées d'Informatique Musicale*, p. 41–45, 2005.

[Don98] E. M. Donovan, R. E.and Eide, "The ibm trainable speech synthesis system", *International Conference on Spoken Language Processing*, 1998.

[BBd02] O. Blouin, Christopheand Rosec, P. C. Bagshaw, C. d'Alessandro, "Concatenation cost calculation and optimisation for unit selection in TTS", *in : In Proceedings of IEEE Workshop on Speech Synthesis*, p. 231–234, 2002.

[MA96] I. Murray, J. Arnott, "Synthesizing emotions in speech: is it time to get excited?", *in : Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, *3*, Ieee, p. 1816–1819, 1996.

trolled affect, emotion, intention or style [Sch01,RSHM09,SCK06]. This is mainly a consequence of the low expressiveness in recorded speech corpora, whose style is often constrained to read speech.

However, expressiveness is an essential component in oral communication. It regroups different speaker and context dependent elements from different abstraction levels which all together enable to highlight **an emotion, an intention or a particular speaking style** [LDM11]. Acoustically, fundamental frequency, intensity and durations of some signal segments are judged to be decisive elements [IAML04,Abe95,Bla07,GR94,IMK+04]. Phonologically, phenomena like phoneme elisions (notably *schwas* in French) or disfluences (e.g., hesitations, repetitions, false starts, etc.) mark different emotional states. At lexical, sentential and more abstract levels, other elements such as the choice of words, syntactic structures, punctuation marks or logical connectors are also important.

States of the art of existing techniques are presented in [Eri05,Sch09,GP13]. These articles state that current systems have important lacks concerning expressiveness. Moreover, they clearly show the need for expressiveness description languages and for more flexibility in TTS systems, especially in corpus-based systems.

Indeed, controlling expressiveness in speech synthesis requires high level languages to precisely and intuitively describe expressiveness that must be conveyed by an utterance. Some works exist, notably concerning corpus annotation [DGWS06], but for the moment, no language

[Sch01] M. Schröder, "Emotional Speech Synthesis : A Review", *in : Proceedings of Eurospeech*, 2001.

[RSHM09] A. R. F. Rebordao, M. a. M. Shaikh, K. Hirose, N. Minematsu, "How to Improve TTS Systems for Emotional Expressivity", *in : Interspeech*, p. 524–527, 2009.

[SCK06] V. Strom, R. Clark, S. King, "Expressive Prosody for Unit-selection Speech Synthesis", *in : Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, 2006.

[LDM11] A. Lacheret-Dujour, M. Morel, "Modéliser la prosodie pour la synthèse à partir du texte : Perspectives sémantico-pragmatiques", *in : Au commencement était le verbe. Syntaxe, sémantique et cognition*, N. Neveu, Franck / Blumenthal, Peter / Le Querler (editor), note 23, Peter Lang, 2011, p. 299–325.

[IAML04] I. Iriondo, F. Alias, J. Melenchon, M. A. Llorca, "Modeling and Synthesizing Emotional Speech for Catalan Text-to-Speech Synthesis", *in : Affective Dialogue Systems*, p. 197–208, 2004.

[Abe95] M. Abe, "Speaking Styles: Statistical Analysis and Synthesis by a Text-to-Speech System", *in : Progress in Speech Synthesis*, J. P. Van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg (editors), Springer Verlag, 1995, ch. 39, p. 495–510.

[Bla07] A. W. Black, "Speech Synthesis for Educational Technology", *in : SLATE*, p. 78–81, 2007.

[GR94] C. Gerard, C. Rigaut, "Patterns prosodiques et intentions des locuteurs : le rôle crucial des variables temporelles dans la parole", *Le Journal de Physique IV 04*, C5, May 1994, p. 505–508.

[IMK+04] Y. Irie, S. Matsubara, N. Kawaguchi, Y. Yamaguchi, Y. Inagaki, "Speech Intention Understanding based on Decision Tree Learning", *in : Interspeech*, p. 2185–2188, 2004.

[Eri05] D. Erickson, "Expressive speech: production, perception and application to speech synthesis", *Acoustical Science and Technology 26*, 4, 2005, p. 317–325.

[Sch09] M. Schröder, "Expressive speech synthesis: Past, present, and possible futures", *in : Affective information processing*, Springer, 2009, p. 111–126.

[GP13] D. Govind, S. R. M. Prasanna, "Expressive speech synthesis: a review", *International Journal of Speech Technology*, 2013, p. 1–24.

[DGWS06] G. Demenko, S. Grocholewski, A. Wagner, M. Szymanski, "Prosody annotation for corpus based speech synthesis", *in : Proceedings of the Eleventh Australasian International Conference*

is sufficient to build up a complete editorial chain. This point constitutes an obstacle towards automatic or semi-automatic creation of high-quality spoken contents.

The number of works on the integration of expressiveness into TTS systems is in constant augmentation these last years. Most of speech synthesis methods have been subject to extension attempts. In particular, we can cite the diphone approach [BNS02], the corpus-based approach [EAB+04,CRK07], or even the parametric approach [WHLW06,TYMK07]. Adding to this, several languages have been used: notably Spanish [ISA07], Polish [DGWS06], Japanese [WHLW06,TYMK07], English [SCK06], and French [AVAR06,LFV+11].

Speech synthesis related domains share some common problems, but generally with an opposed point of view. In speaker processing and automatic speech recognition, acoustic models try to represent the speech signal spectrum so as to deduce a footprint or to erase specificities and move towards a generic model [SNH03,SFK+05]. In speech recognition again, word pronunciations are generally given by large phonetized lexicons. Nevertheless, tools

on Speech Science and Technology, p. 460–465, 2006.

[BNS02] M. Bulut, S. S. Narayanan, A. K. Syrdal, "Expressive speech synthesis using a concatenative synthesizer", *in : Proc. ICSLP*, p. 1265–1268, 2002.

[EAB+04] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, J. Pitrelli, "A corpus-based approach to expressive speech synthesis", *in : Fifth ISCA Workshop on Speech Synthesis*, 2004.

[CRK07] R. A. J. Clark, K. Richmond, S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system", *Speech Communication 49*, 4, 2007, p. 317–330.

[WHLW06] C.-H. Wu, C.-C. Hsia, T.-H. Liu, J.-F. Wang, "Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis", *Audio, Speech, and Language Processing, IEEE Transactions on 14*, 4, 2006, p. 1109–1116.

[TYMK07] N. Takashi, J. Yamagishi, T. Masuko, T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis", *IEICE TRANSACTIONS on Information and Systems 90*, 9, 2007, p. 1406–1413.

[ISA07] I. Iriondo, J. C. Socoró, F. Alías, "Prosody modelling of Spanish for expressive speech synthesis", *in : Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, *4*, IEEE, p. IV–821, 2007.

[DGWS06] G. Demenko, S. Grocholewski, A. Wagner, M. Szymanski, "Prosody annotation for corpus based speech synthesis", *in : Proceedings of the Eleventh Australasian International Conference on Speech Science and Technology*, p. 460–465, 2006.

[SCK06] V. Strom, R. Clark, S. King, "Expressive Prosody for Unit-selection Speech Synthesis", *in : Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, 2006.

[AVAR06] N. Audibert, D. Vincent, V. Aubergé, O. Rosec, "Expressive speech synthesis: Evaluation of a voice quality centered coder on the different acoustic dimensions", *in : Proc. Speech Prosody, 2006*, p. 525–528, 2006.

[LFV+11] P. Lanchantin, S. Farner, C. Veaux, G. Degottex, N. Obin, G. Beller, F. Villavicencio, T. Hueber, D. Schwartz, S. Huber *et al.*, "Vivos Voco: A Survey of Recent Research on Voice Transformations at IRCAM", *in : International Conference on Digital Audio Effects (DAFx)*, p. 277–285, 2011.

[SNH03] D. Sundermann, H. Ney, H. Hoge, "VTLN-based cross-language voice conversion", *in : Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, IEEE, p. 676–681, 2003.

[SFK+05] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, A. Venkataraman, "MLLR transforms as features in speaker recognition", *in : in Proceedings of the 9th European Conference on Speech Communication and Technology*, Citeseer, 2005.

are still needed to automatically associate one or several phonetizations to out-of-vocabulary words [BÓ1,BN08,IFJ11]. Other works aim at modeling disfluences, i.e., errors within the elocution of a sentence, in order to help a recognition system to deal with these irregularities [Shr94,SS96]. By extension, these models are useful to clean a manual or automatic transcription, and make it closer to written text conventions [LSS+06]. Although all these studies share common traits with the expressive speech synthesis problem, they all try to characterize the effects of expressiveness to get rid of them, and not the other way around. Finally, emotion detection is also a subject of interest. In [LTAVD11], the authors are interested in emotion recognition from linguistic clues while the authors of [SMLR05] propose models mixing acoustic and linguistic features to detect emotions in speech signals. In the case of expressive speech synthesis, dependencies highlighted by these works would have to be reversed in order to predict acoustic features from expressiveness input instructions.

In this context, the scientific goal of the team in speech processing is to **take into account expressiveness in speech synthesis systems**. This objective leads us to the research topics detailed in the rest of the document.

3.3 Text processing

Keywords: Textual data, data acquisition, text mining, knowledge acquisition, knowledge data discovery.

We are interested to extract, automatically and semi-automatically, information and knowledge from textual data. The purpose of extracting information and knowledge is to organize and define software components useful for designing and realizing computer-based systems facilitating several activities performed by individuals (possibly when working for companies).

[BÓ1] F. Béchet, "LIA_PHON : un systï¿½me complet de phonï¿½tisation de textes", *Traitement Automatique des Langues (TAL) 42*, 1, 2001, p. 47–67.

[BN08] M. Bisani, H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion", *Speech Communication*, 2008.

[IFJ11] I. Illina, D. Fohr, D. Jouvet, "Multiple Pronunciation Generation using Grapheme-to-Phoneme Conversion based on Conditional Random Fields", *in : Proceedings of the International Conference on Speech and Computer (SPECOM)*, 2011.

[Shr94] E. Shriberg, *Preliminaries to a Theory of Speech Disfluencies*, PdD Thesis, University of California, Berkeley, California, USA, 1994.

[SS96] A. Stolcke, E. Shriberg, "Statistical Language Modeling for Speech Disfluencies", *in : Proc. of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, *1*, p. 405–408, Atlanta, Georgia, USA, may 1996.

[LSS+06] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies", *IEEE Transactions on Audio, Speech, and Language Processing 14*, 5, 2006, p. 1526–1540.

[LTAVD11] M. Le Tallec, J.-Y. Antoine, J. Villaneau, D. Duhaut, "Affective Interaction with a Companion Robot for Hospitalized Children: a Linguistically based Model for Emotion Detection", *in : Proceedings of the 5th Language and Technology Conference (LTC'2011)*, p. 6 pages, Poznan, Pologne, 2011.

[SMLR05] B. Schuller, R. Müller, M. Lang, G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles", *in : Proc. Interspeech*, p. 805–808, 2005.

For instance, 1) making easier, quicker and reliable the way to decide on the base of textual data, 2) making explicit hidden information conveyed by textual data and, as a consequence, 3) enabling understanding of other individuals' behaviors, ideas and so on, 4) making computer-based systems more efficient and effective on the base of available textual data, and finally 5) supporting individuals in following what textual data implicitly suggest.

Knowledge extraction from textual data can be divided into three topics: textual data acquisition and filtering, text mining, and knowledge representation.

Textual data acquisition and filtering: The first step in order to deal with textual data is the acquisition process and filtering. Raw textual data can be automatically or manually obtained, and need some process like filtering to be mined. One of this process is the task of corpora annotation. Manually annotated corpora are a key resource for natural language processing. They are essential for machine learning techniques and they are also used as references for the system evaluations. The question of data reliability is of first importance to assess the quality of manually annotated corpora. The interest for such enriched language resources has reached domains (semantics, pragmatics, affective computing) where the annotation process is highly affected by the coders subjectivity. The reliability of the resulting annotations must be trusted by measures that assess the inter-coders agreement. Currently, the κ-statistic is a prevailing standard but critical works show its limitations [AP08] and alternative measures of reliability have been proposed [**?**]. We conduct some experimental studies to investigate the factors of influence that should affect reliability estimation.

Text mining: Due to the explosion of available textual data, text mining and Information Extraction (IE) from texts have become important topics of study in recent years. Specially, text mining is particularly adapted in order to identify expressiveness in textual data. For instance, tasks like sentiment analysis or opinion mining allow to identify expressiveness. Several kinds of techniques have been developed to mine textual data. We focus on sequential pattern extraction which is particularly adapted to textual data. Sequential mining aims at discovering frequent sub-sequences in large sequence databases. Two important paradigms are proposed to reduce the important number of patterns: using constraints and condensed representations. Constraints allow a user to focus on the most promising knowledge by reducing the number of extracted patterns to those of potential interest. There are now generic approaches to discover patterns and sequential patterns under constraints (e.g., [NLHP,PHW02,PHW07,Bon04]). The strength is that constraint-based pattern mining challenges two major problems in pat-

[AP08] R. Artstein, M. Poesio, "Inter-Coder Agreement for Computational Linguistics", *COMPUTATIONAL LINGUISTICS 34*, 4, 2008, p. 555–596.

[NLHP] R. Ng, L. Lakshmanan, J. Han, A. Pang, "Exploratory mining and pruning optimizations of constrained associations rules", *in : Proceedings of SIGMOD'98*, p. 13–24.

[PHW02] J. Pei, J. Han, W. Wang, "Mining Sequential Patterns with Constraints in Large Databases", ACM Press, p. 18–25, 2002.

[PHW07] J. Pei, J. Han, W. Wang, "Constraint-based sequential pattern mining: the pattern-growth methods", *Journal of Intelligent Information Systems 28*, 2007, p. 133–160.

[Bon04] F. Bonchi, "On closed constrained frequent pattern mining", *in : In Proceedings IEEE Int. Conf. on Data Mining ICDM'04*, Press, p. 35–42, 2004.

tern mining: effectiveness and efficiency. Because the set of frequent sequential patterns can be very large, a complementary method is to use condensed representations. Condensed representations, such as closed sequential patterns [YHA03,WH04], have been proposed in order to eliminate redundancy without loss of information. The main challenge in sequential pattern extraction is to be able to combine constraints and condensed representations as in itemsets paradigm which can be useful in many tasks as to analyze gesture and speech captured data.

Knowledge representation: We use ontologies as main tool enabling explicit and precise representation of information and knowledge about concepts and relationships hidden in available texts. Indeed, textual data provide samples of concepts and relationships (such as words 'my car...' as example of a possible concept 'car'), as well as references to concepts and relationships (such as word 'car' as reference to a possible concept 'mean of transport' or just to 'car'). Finding those concepts and relationships is a prerequisite to further enrich earlier ontology versions by adding new artefacts (for instance, new axioms), not (necessarily) provided in available texts. Ontologies can be built manually by following specialized methodologies [FLGPJ97]. However, with the growing number of available texts, performing the work manually becomes even more difficult, error-prone and expensive. As a consequence, various works have focused on how ontologies can be automatically extracted [CV05,MFP09,NV04,VB08,BM05,PCR10]. However, as also recently highlighted [Gan13], despite the performed work, there is the need to understand much better foundations for bridging the gap between techniques usable for processing and analyzing texts and information for filling ontology contents (basically con-

[YHA03] X. Yan, J. Han, R. Afshar, "CloSpan: Mining Closed Sequential Patterns in Large Databases", *in : SDM*, 2003.

[WH04] J. Wang, J. Han, "BIDE: Efficient Mining of Frequent Closed Sequences", *in : Proceedings of the 20th International Conference on Data Engineering, ICDE '04*, IEEE Computer Society, p. 79–, Washington, DC, USA, 2004, `http://dl.acm.org/citation.cfm?id=977401.978142`.

[FLGPJ97] M. Fernandez-Lopez, A. Gomez-Perez, N. Juristo, "METHONTOLOGY: from Ontological Art towards Ontological Engineering", *in : Proceedings of the AAAI97 Spring Symposium*, p. 33–40, Stanford, USA, March 1997.

[CV05] P. Cimiano, J. Vï¿½lker, "Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery", *in : Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, A. Montoyo, R. Munoz, E. Metais (editors), *Lecture Notes in Computer Science, 3513*, Springer, p. 227–238, Alicante, Spain, 2005.

[MFP09] D. Maynard, A. Funk, W. Peters, "SPRAT: a tool for automatic semantic pattern-based ontology population", *in : IN: INTERNATIONAL CONFERENCE FOR DIGITAL LIBRARIES AND THE SEMANTIC WEB*, 2009.

[NV04] R. Navigli, P. Velardi, "Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites", *Comput. Linguist. 30*, 2, 2004, p. 151–179.

[VB08] J. Volker, E. Blomqvist, "Prototype for learning networked ontologies", *research report number 3.8.1 for NEON Project*, Instit. AIFB, Univ. of Karlsruhe, 2008.

[BM05] P. Buitelaar, B. Magnini, "Ontology Learning from Text: An Overview", *in : In Paul Buitelaar, P., Cimiano, P., Magnini B. (Eds.), Ontology Learning from Text: Methods, Applications and Evaluation*, IOS Press, p. 3–12, 2005.

[PCR10] J. Park, W. Cho, S. Rho, "Evaluating Ontology Extraction Tools Using a Comprehensive Evaluation Framework", *Data Knowl. Eng. 69*, 10, 2010, p. 1043–1061.

[Gan13] A. Gangemi, "A Comparison of Knowledge Extraction Tools for the Semantic Web", *in : ESWC*, p. 351–366, 2013.

cepts, relationships, axioms). Understanding foundations leads to automation improvement and therefore reduction of the required human effort for extracting valuable ontologies.

4 Application Domains

4.1 Expressive gesture

Participants: Pamela Carreño, Lei Chen, Marc Dupont, Sylvie Gibet, Ludovic Hamon, Jean-François Kamp, Caroline Larboulette, Thibaut Le Naour, Pierre-François Marteau, Gildas Ménier.

- Editing motion captured data and animating signing avatars from concatenate synthesis; this application domain covers the design of corpora and sign language indexed databases, and the retrieval of motion capture data to compose new utterances and control a virtual signer [GCDLN11,HGB13,LAGT+13]. This application requires high-quality data simultaneously recorded (body and hand motion, facial expression, gaze direction), as well as efficient and original access to the database, for example using sketched-based gestural interaction associated to motion retrieval techniques.

- Analysis and synthesis of expressive motion data applied to theatrical gestures; this application domain covers the design of expressive scenarios in the domain of theatrical gestures, and the development of learning techniques to generate new expressive gestures.

- Using spatial relationship to analyze and edit motion; this application domain covers the reconstruction of marker trajectories and the animation of mesh constrained by an implicit skeleton and driven by 3D markers trajectories.

4.2 Expressive speech

Participants: Nelly Barbot, Vincent Barreaud, Jonathan Chevelu, Arnaud Delhay, Sébastien Le Maguer, Gwénoé Lecorvé, Damien Lolive.

[GCDLN11] S. Gibet, N. Courty, K. Duarte, T. Le Naour, "The SignCom System for Data-driven Animation of Interactive Virtual Signers : Methodology and Evaluation", *in : Transactions on Interactive Intelligent Systems*, *1*, 1, ACM, 2011.

[HGB13] L. Hamon, S. Gibet, S. Boustila, "Édition interactive d'énoncés en langue des signes française dédiée aux avatars signeurs", *in : TALN - 20ème conférence du Traitement Automatique du Langage Naturel 2013*, Sables d'Olonne, France, June 2013, `http://hal.inria.fr/hal-00911629`.

[LAGT+13] F. Lefebvre-Albaret, S. Gibet, A. Turki, L. Hamon, R. Brun, "Overview of the Sign3D Project High-fidelity 3D recording, indexing and editing of French Sign Language content", *in : Third International Symposium on Sign Language Translation and Avatar Technology (SLTAT) 2013*, Chicago, États-Unis, October 2013. Programme " investissements d'avenir ", Direction Générale de la Compétitivité, de l'Industrie et des Services (DGCIS), ministère du redressement productif,, `http://hal.archives-ouvertes.fr/hal-00914661`.

- Speaker characterization and voice personalization: models that can be adapted to a speaker thus taking into account its mood, personality or origins. Complete process of voice creation taking into account personalization of voice.
- Linguistic corpus design and corpus creation process: this application domain covers both the design of recording scripts and restriction of audio corpora to address specific tasks.
- High-quality multimedia content generation: this application is really meaningful in the framework of speech synthesis as it needs a fine control of expressiveness in order to keep user's attention.

Expressiveness tends to make users accept TTS outputs by producing less impersonal speech. Thus, it plays a fundamental role in a large number of concrete applications. Among all applications, we can mention:

- high-quality audiobook generation;
- online learning and in particular autonomous language learning;
- device personalization for disabled people, for who expressive voice creation is an important need;
- video games.

4.3 Expression in textual data

Participants: Nicolas Béchet, Giuseppe Bério, Pierre-François Marteau, Gildas Ménier, Farida Said, Jeanne Villaneau, Hai Hieu Vu.

Expression in textual data can conduct to many application dealing with knowledge extraction from textual data. We list some of them in the following.

- Use text mining, natural language processing and clustering techniques in order to study and interpret texts in regard to their linguistic and tonal style.
- Building ontologies to support a man-machine dialogue system.
- Automatic extension of French emotional norms. Emotional norms are resources which are used in opinion mining or sentiment analysis.

5 Software

5.1 SMR

Participants: Ludovic Hamon, Sylvie Gibet, Thibaut Le Naour.

We have developed in the team a whole motion-capture-driven synthesis pipeline dedicated to the editing of gestures in French Sign Language (LSF) and the generation of gestures that can be visualized through a 3D virtual signer. This system is able to produce novel utterances from the corpus data by combining motion chunks that have been previously captured on a real signer, and by using these data to animate a virtual signer. As gestures in signed languages are by essence multichannel, i.e. meaningful information is conveyed by multiple body parts acting in parallel, it follows that a sign editing system manipulates motion segments that are decomposed on these channels over time. The editing system is able to accurately and efficiently retrieve the annotated SL items from the database, and to concatenate the corresponding motion chunks spatially (i.e. along the channels), and temporally within an animation system. This last one thus synchronizes and handles at the same time several modalities involved in signed gestures and produce a continuous flow of skeleton postures.

This development has given rise to the achievement of a software library called *SMR_SignCom* dedicated to the indexing, storing, motion modeling, and 3D virtual character control and visualization. The virtual character is able to produce new utterances in French Sign language (LSF), through the combining of semantic items contained in the corpus. The *SMR_SignCom* library is composed of several modules.

- One module ensures the managing of a motion database. It is dedicated to the recording and the processing of movements associated to their annotations.
- One module is dedicated to the modeling of skeleton, the motion recomposing from the extraction and assembling of motion chunks, the animation of a skeleton-based character from several controllers associated to the different body parts (upper-torso, lower-torso, hands, facial expressions, and gaze), and the visualization of the resulting animated motion (skeleton representation).
- One rendering engine dedicated to the visualization of the 3D virtual environment.

5.2 ROOTS

Participants: Nelly Barbot, Vincent Barreaud, Jonathan Chevelu, Arnaud Delhay, Sébastien Le Maguer, Gwénolé Lecorvé, Damien Lolive.

The development of new methods for given speech and natural language processing tasks usually faces, beyond scientific aspects, various technical and practical data management problems. Indeed, the sets of required annotated features and their desired distribution in the training data are rarely the same for two different tasks, and many dedicated systems or expert resources use different file formats, time scales, or alphabets of tags.

In this context, ROOTS, stemming for Rich Object Oriented Transcription System, is an open source toolkit dedicated to annotated sequential data generation, management and processing, especially in the field of speech and language processing. It works as a consistent middleware between dedicated data processing or annotation tools by offering a consistent view of various annotation levels and synchronizing them. Doing so, ROOTS ensures a clear separation between description and treatment. Theoretical aspects of multilevel annotation

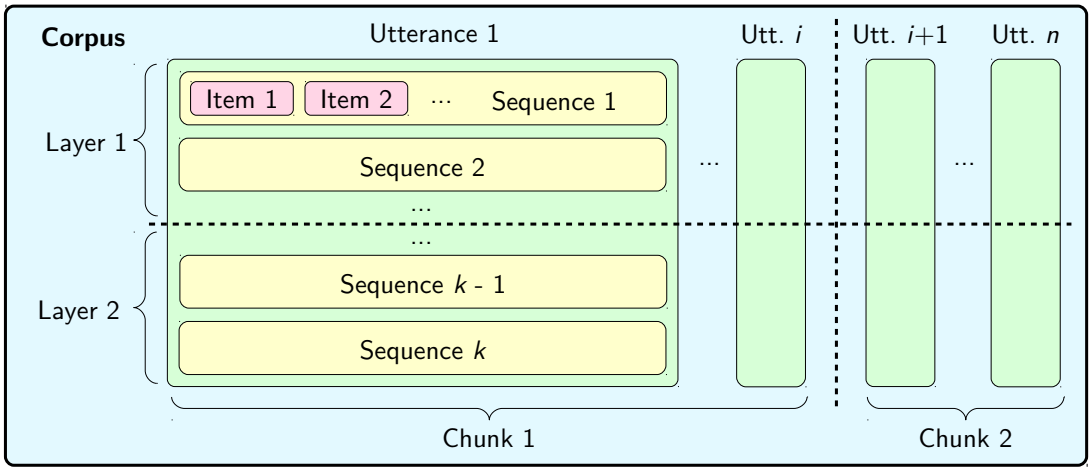

Figure 1: Hierarchical organization of data in ROOTS.

synchronization have previously been published in [BBB+11] while a prototype had been presented and applied to an audiobook annotation task in [BCLML12].

As summarized in Figure 1, data are organized hierarchically in Roots, starting from fine grain information in items and moving to macroscopic representations as corpora. As a fundamental concept, data in ROOTS is modeled as sequences of items. These items can be of many types, e.g., words, graphemes, named entity classes, signal segments, etc., and can thus represent various annotation levels of the same data. Correspondences between items from different sequences are then defined as algebraic relations, leading to a graph where nodes are items and edges are derived from relations. Then, interrelated sequences are gathered into utterances. According to the application domain, utterances can refer to sentences, breath groups, or any relevant unit. A part of the recent work on ROOTS has focused on extending this hierarchization of data to easily handle large collections of data. Hence, the notion of corpus has been defined as a list of utterances or, recursively, as a list of subcorpora (called chunks), for instance to represent a chapter as a list of paragraphs. Besides chunks, corpora can also be partitioned "horizontally" into layers which gather annotations from a same field. The following operations are allowed for each data hierarchization level:

- Item: get/set the content/characteristics; get other items in relation; dump[1].
- Sequence: add/remove/get/update items; dump,

[1]Dump refers to input/output operations in raw text, XML and JSON formats.

[BBB+11] N. BARBOT, V. BARREAUD, O. BOËFFARD, L. CHARONNAT, A. DELHAY, S. LE MAGUER, D. LOLIVE, "Towards a Versatile Multi-Layered Description of Speech Corpora Using Algebraic Relations", *in : Conference of the International Speech Communication Association (Interspeech)*, p. 1501–1504, Florence, Italie, 2011.

[BCLML12] O. BOËFFARD, L. CHARONNAT, S. LE MAGUER, D. LOLIVE, "Towards Fully Automatic Annotation of Audio Books for TTS", *in : Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, European Language Resources Association (ELRA), Istanbul, Turkey, may 2012.

- Relation: get items related to another; link or unlink items; dump,
- Utterance: add/remove/get/update sequences; add/remove/get/update direct or composed relations; dump,
- Corpus: add/remove/get/update an utterance; add/remove chunks/layers; load/save; dump.

ROOTS is made of a core library and of a collection of utility scripts. All functionalities are accessible through a rich API either in C++ or in Perl. Recently, this API has greatly evolved and to ease building ROOTS corpora based on this API (e.g., with the notion of corpus), and accessing information in flexible and intuitive manners. Extra developments have also led to the following improvements: new wrapping scripts for basic corpus processing operations (merge, split, search) have been written and a LaTeX/PGF graphical output mechanism has been added in order to expertise and analyse the content of annotated utterances. This visualization functionality has been developed during the 3-month summer internship of Andrei Zene, a Romanian B.Sc. student.

The toolkit ROOTS is original compared to other related tools. Among them, GATE [CB02] proposes a framework to develop NLP pipelines but does not provide facilities to switch between GATE bundled processing components and external tools. More recently, the NITE XML Toolkit, or NXT, proposes a generic data organization model able to represent large multimodal corpora with a wide range of annotation types [CEHK05,CCB+10]. Whereas NXT considers corpora as databases from which data is accessed through a query language, ROOTS lets the user browse data as he sees fit. In a more general approach, UIMA [FL04,FLG+06] proposes software engineering standards for unstructured data management, including annotation and processing. UIMA is technically too advanced for fast and light prototyping. It is rather devoted to industrial developments. In the end, ROOTS is closer to work done within the TTS system Festival [BTCC02]. This system relies on a formalism called HRG, standing for Heterogenous Relation Graphs, which offers a unique representation of different information

[CB02] D. CUNNINGHAM, H.AND MAYNARD, V. BONTCHEVA, K.AND TABLAN, "GATE: an architecture for development of robust HLT applications", *in: Proceedings of the Annual Meeting of the ACL*, p. 168–175, 2002.

[CEHK05] J. CARLETTA, S. EVERT, U. HEID, J. KILGOUR, "The NITE XML Toolkit: Data Model and Query Language", *Language Resources and Evaluation 39*, 4, 2005, p. 313–334.

[CCB+10] S. CALHOUN, J. CARLETTA, J. M. BRENIER, N. MAYO, D. JURAFSKY, M. STEEDMAN, D. BEAVER, "The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue", *Language Resources and Evaluation 44*, 4, 2010, p. 387–419.

[FL04] D. FERRUCCI, A. LALLY, "UIMA: an architectural approach to unstructured information processing in the corporate research environment", *Natural Language Engineering 10*, 3-4, 2004, p. 327–348.

[FLG+06] D. FERRUCCI, A. LALLY, D. GRUHL, E. EPSTEIN, M. SCHOR, J. W. MURDOCK, A. FRENKIEL, E. W. BROWN, T. HAMPP, Y. DOGANATA *et al.*, "Towards an interoperability standard for text and multi-modal analytics", 2006.

[BTCC02] A. W. BLACK, P. TAYLOR, R. CALEY, R. CLARK, "The Festival speech synthesis system", *research report*, University of Edinburgh, 2002.

levels involved in the TTS system [TBC01]. Still, our tool is different from HRG in the sense that the latter is part of the TTS system Festival whereas ROOTS is completely autonomous. Moreover, ROOTS comes along with a true application programming interface (API), in C++ and Perl for the moment.

As a result of recent improvements, ROOTS is now in use in most of the software developed for speech processing, namely the corpus-based speech synthesizer, corpus generation/analysis tools or the phonetizer. Moreover, ROOTS serves as a basis for corpus generation and information extraction for the ANR Phorevox project. For instance, we have built a corpus containing 1000 free e-books which is planned to be proposed to the community. Finally, ROOTS has been registered in 2013 at the Program Protection Agency (*Agence pour la Protection des Programmes*, APP) and publicly released under the terms of LGLP licence on `http://roots-toolkit.gforge.inria.fr`. A paper has been published in the main international language resource conference to let the community know about this release [10].

5.3 Web based listening test system

Participants: Vincent Barreaud, Arnaud Delhay, Sébastien Le Maguer, Damien Lolive.

The listening test platform is developed by the team, especially to evaluate speech synthesis models. This platform has been developed to propose to the community a ready to use tool to conduct listening tests under various conditions. Our main goals were to make the configuration of the tests as simple and flexible as possible, to simplify the recruiting of the testees and, of course, to keep track of the results using a relational database.

The most widely used listening tests used in the speech processing community are available (AB-BA, ABX, MOS, MUSHRA, etc.).

This software is currently implemented in PHP and integrated in the Symfony2 framework with Doctrine as database manager and Twig templates. This configuration makes the platform accessible from a wide variety of browsers.

The platform is designed to enable researchers to build wide tests available through the web. The main functionalities provided are as follows:

- Users are given roles, which give them privileges,
- Users answer test during a trial which can be interrupted and resumed later,
- Users give information on their listening conditions at each trial beginning,
- Tests are imported from Zip archives that contain a XML configuration file and the stimulii,
- Users can be imported from a XML configuration file.
- A tester can monitor his test and discard results of a testee on the basis of its statistical behavior.

[TBC01] P. TAYLOR, A. W. BLACK, R. CALEY, "Heterogeneous relation graphs as a formalism for representing linguistic information", *Speech communication 33*, 2001, p. 153–174.

- The platform is open-source (under AGPLv3 Licence).

5.4 Automatic segmentation system

Participants: Damien Lolive.

The automatic segmentation system consists of a set of scripts aligning the phonetic transcription of a text with its acoustic signal counterpart. The system is made of two parts: the first one generates a phonetic graph including phonological variants (pauses, liaisons, schwas,...), the second one, based on HMM modeling, performs a Viterbi decoding determining the most likely phonetic sequence and its alignment with the acoustic signal.

To be efficient, the system should be applied to texts that have been manually checked (compliance with the recording, spelling, syntax) and annotated. The annotation stage consists in adding tags indicating excepts in foreign language, non standard pronunciation and noises (breathing, laughter, coughing, sniffing, snorting, etc.). It is also possible to improve the decoding performances by adding a list of phonetization of proper names and foreign pronunciations.

5.5 Corpus-based Text-to-Speech System

Participants: Nelly Barbot, Jonathan Chevelu, Arnaud Delhay, David Guennec, Damien Lolive.

For research purposes we developed a whole text-to-speech system designed to be flexible. The system, implemented in C++, intensively use templates and inheritance, thus providing the following benefits:

- the algorithm used for unit selection can be easily changed. For instance, we implemented both $A^\star$ and Beam-search simply by using subclassing and without changing the heart of the system.
- cost functions can also be changed the same way which provides a simple way to experiment new functions.

Moreover the system implements state of the art technique to achieve good performance while manipulated large speech corpora such as hash tables and pre-selection filters [**?**]. To achieve this, each phone in the corpus is given a binary key which enables A^* to take or reject the unit. Thus, the key contains phonetic, linguistic and prosodic information. Binary masks are used to get access only to the desired information during runtime.

The pre-selection filters are integrated to the hash functions used to access the units in the corpus in order to reduce the number of candidates explored. For the moment, the whole set of filters is the following:

1. Is the unit a Non Speech Sound ?
2. Is the phone in the onset of the syllable?
3. Is the phone in the coda of the syllable?

4. Is the phone in the last syllable of its breath group?

5. Is the current syllable in word end?

6. Is the current syllable in word beginning?

Concretely, the pre-selection filters are relaxed one by one, starting from the end of the list, if no unit corresponding to the current set is found. One drawback is that we can explore candidates far from the target features we want, thus risking to produce artefacts but this backtracking mechanism insures to find a unit and to produce a solution. The priority order of the filters is the one given above.

Finally, high level features are also available to get, for example, the best path or the N-best paths, with a detailed output of the cost values.

Some developments are currently undertaken to provide more features and pre-selection filters and also to improve flexibility of the system to gain a fine control over prosody. This last objective is linked to the main objectives of the team to control expressivity during synthesis.

5.6 Recording Studio

Participants: Nelly Barbot, Vincent Barreaud, Damien Lolive.

A main goal of the Expression project consists in developing high quality voice synthesis. Our research activities use speech corpora as a raw material to train statistical models. A good speech corpus quality relies on a consistent speech flow (the actor does not change his speaking style during a session) recorded in a consistent (and quiet) acoustic environment. In order to expand our research scope, it is often interesting to vary the speech style (dialogs, mood, accent, etc.) as well as the language style. Unfortunately, such corpora are hard to obtain and generally do not meet specific experimental requirements. To deal with these constraints, speech resources need to be recorded and controlled by our own protocols.

5.6.1 Hardware architecture

The funding of this recording studio comes from MOB-ITS (CPER, 2007-2013). The MOB-ITS platform (Mobile and interactive access to data) is a joint project of IRISA teams in Lannion (IUT and ENSSAT). This contract is part of the support to the "Pôle de compétitivité Images & Réseaux".

This recording studio consists in two rooms: an isolation booth and control room.

The isolation booth can fit three persons. It is designed to attenuate the noises of 50dB and is equipped with two recording sets. A recording set consists in a high quality microphone (Neumann U87AI), a high quality closed head set (Beyer DT 880 250ohms), a monitor and a webcam.

The control room is equipped with two audio networks, a video network and computer network. The first audio network is a high quality digital recording line going from the isolation booth microphones to a digital sound card through a preamplifier (Avalon Design AD2022), an equalizer (Neve 8803 Dual Channel) and finally an analogic/digital converter (Lynx Aurora 8). The digital sound is edited with a logical sampling table (Avid Pro Tools).

In addition to the signal issued by the isolation room, the digital audio network can record the signals from an Electro-Gloto Graph (EGG) that capture the glottal activity of the actor. This activity is used to induce the F0 (first formant) trajectory which is the main indicator of the prosody. This activity must be digitalized and recorded along with the audio activity in order to reduce the latency between the two signal.

The second audio network is for control purpose and is fully analogic. It is used by the operator to control the quality of the recorded sound, the consistency of the actor, the accuracy of the transcription. An actor can receive audio feedback of his own voice, disturbing stimulii (music, other voices, their own delayed voice) or directions from the operator through this audio line. This network consists in four Neumann KH 120 loud-speakers (two in the booth, two in the control room), a head set amplifier (ART headamp 6 pro) and an analogic sampling table (Yamaha MG206C). The computer network stores the recording sessions scenarii and prompt the actor.

The video network switches the video output (computers, webcam) to screens installed in the isolation booth (for prompting) and the control room (for monitoring).

5.6.2 Software architecture

Actors in the isolation booth must be prompted to utter speech with various indications (mood, intonation, speed, accent, role, ...). The prompt must be presented on the simplest interface, for instance a lcd screen or a tablet. The latest developments on the recording studio consist in a software implemented on the computer at the end of the digital audio network that record sound files, segment them and link them to the transcription. This software is controlled by the operator who checks that the actors actually uttered the prompted sentence and the quality of the recording. Thus, the operator can possibly reject (in fact, annotate) a file and prompt the actors again with the discarded sentence.

The digital sound card used for recording only offers microsoft drivers. Consequently, this software has been developed with the Windows Audio and Sound API (WASAPI). The main difficulty resides in the simultaneous recording of two distinct channels (audio and EGG) without any jitter between the two signals.

6 New Results

6.1 Data processing and management

Participants: Pierre-François Marteau, Sylvie Gibet, Gildas Ménier.

Time series and sequence similarity : We developed a framework dedicated to the construction of what we call *discrete elastic inner product* allowing one to embed sets of non-uniformly sampled multivariate *time* series or sequences of varying lengths into inner product space structures. This framework is based on a recursive definition that covers the case of multiple embedded *time elastic* dimensions. We prove that such inner products exist in our general framework and show how a simple instance of this inner product class operates on

some prospective applications, while generalizing the Euclidean inner product. Classification experimentations on time series and symbolic sequences datasets demonstrate the benefits that we can expect by embedding time series or sequences into elastic inner spaces rather than into classical Euclidean spaces. These experiments show good accuracy when compared to the euclidean distance or even dynamic programming algorithms while maintaining a linear algorithmic complexity at exploitation stage, although a quadratic indexing phase beforehand is required.

6.2 Expressive Gesture

Participants: Sylvie Gibet, Pamela Carreño, Lei Chen, Marc Dupont, Ludovic Hamon, Jean-François Kamp, Caroline Larboulette, Thibaut Le Naour, Pierre-François Marteau, Gildas Ménier.

6.2.1 High-fidelity 3D recording, indexing and editing of French Sign Language content - Sign3D project

A complete workflow from the movement capture (including all upper body part articulations, facial expression and gaze direction) to their restitution using a 3D virtual signer has been defined and achieved.

Corpus creation As automatic signed information in public areas is one of the most promising applications of virtual signers, we have concentrated our efforts on messages such as "opening hours", "entrance fees" or "perturbation messages". The main challenge was to build the corpus in order to be able to compose other utterances by recombination and to have enough spatio-temporal variability for each sign.

High-fidelity motion capture and annotations The sequences of the *Sign3D* project have been captured thanks to a Vicon T160 optical motion capture system (16 Megapixel cameras) combined with a head-mounted occulometer (MocapLab MLab 50-W). Markers were placed on the whole signer's upper body, including her face and fingers which allows for a complete performance capture (Figure 2 left). After motion capture, the marker set is rigged onto a 3D virtual signer mesh in order to animate both its skeleton and face (Figure 2 right).

During motion capture, a reference video (a frontal view of the signer) has also been recorded. Then deaf signers have annotated videos with the Elan software [CS08]. Sentences are segmented into signs, labeled by a string conveying its meaning. Other meta-data have also been added to the segments describing handshape, facial expression, body posture, or any other feature that may be relevant to choose a good sign variant when creating new utterances.

Motion database and retrieval The corpus annotation allows a mapping between the meanings of the signs (and their distinctive features) and their performance presented as

[CS08] O. Crasborn, H. Sloetjes, "Enhanced ELAN functionality for sign language corpora.", *in : In: Proceedings of LREC 2008, Sixth International Conference on Language Resources and Evaluation.*, 2008.

Figure 2: Motion capture session: the real signer from the recorded video (left) and the 3D mesh of the corresponding virtual signer (right)

motion captured data, using a method close to [ACD+09b]. Once the new sentence is composed, the solution has to select the *good* sign sequence, which means retrieving in the database the good variant of each sign. At this stage, it is important to point out that the goal is to find at least one way to express an information (*i.e.*: opening hours of the museum: 8am - 17pm[2]), but the way of signing is not specified.

From the play list to the animation The previous step results in one (or several) list(s) of segments that have to be concatenated into a new SL sentence. If several lists are correct from a syntactical point of view, the selected one will have to be the one that optimizes the transitions between motion segments. Then, an animation engine computes the combination of motion chunks.

The first concepts of this editing system have been presented in [12], and the results of the platform in [16].

6.2.2 Using spatial relationships for analysis and editing of motion

The animation of virtual characters driven by data is one of the key topics in computer graphics. In this context, a motion is classically defined by a list of skeletons over time, each of them being described by a vector of positions and rotations. The 3D mesh is then controlled by the skeletons by a rigging step between the skeleton and the mesh. In his PhD work, Thibaut le Naour has proposed to study other representations of the motion through a set of spatial relationships [2]. Two approaches have been proposed: the first one considers the motion in the metric space and the second one characterizes each posture by a differential representation using the Laplacian operator. Two relevant studies that illustrate the possibilities offered by such representations and methods are presented below.

[2]We make an analogy with a written language but the program naturally uses a language model dedicated to sign language.

[ACD+09b] C. Awad, N. Courty, K. Duarte, T. L. Naour, S. Gibet, "A Combined Semantic and Motion Capture Database for Real-Time Sign Language Synthesis", *in: IVA*, p. 432–438, 2009.

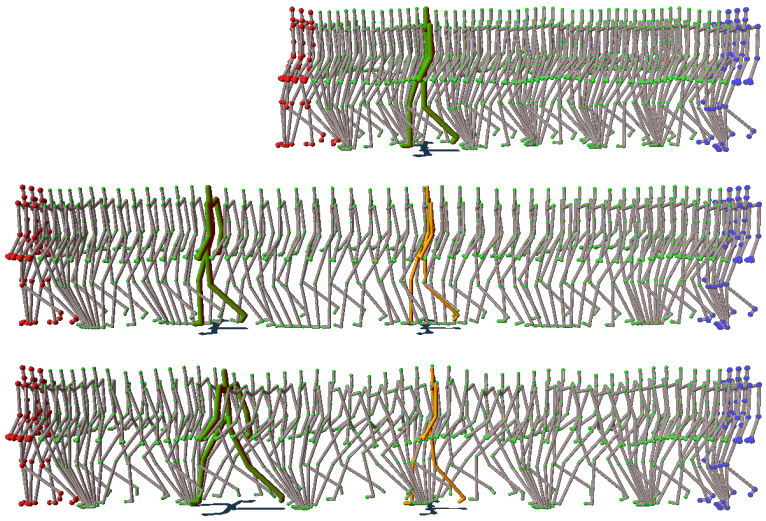

Figure 3: Effects of the weights of the Laplacian matrix on the walk motion deformation.

3D+t motion Laplacian for motion editing

Motion editing requires the preservation of spatial and temporal information of the motion. During editing, this information should be preserved at best. A new representation of the motion based on the Laplacian expression of a 3D+t graph is proposed: the set of connected graphs given by the skeleton over time. Through this Laplacian representation of the motion, an application which allows an easy and interactive editing, correction or retargeting of a motion is presented. The new created motion is the result of the combination of two minimizations, both linear and non-linear: the first one penalizes the difference of energy between the Laplacian coordinates from an animation to the desired one. The second one preserves the length of segments. Using several examples, the benefits of the method and in particularly the preservation of the spatiotemporal properties of the motion in an interactive context is demonstrated in [4]. An example of the Laplacian editing possibilities is given in Figure 3 for walking deformations.

Reconstructing markers' trajectories from motion captured data When capturing motion, we observe at the end of the acquisition process noisy data and missing trajectories that can be due to occlusions or markers' inversions during the recording process. This happens for example when capturing rapid and accurate hand movements, giving rise to many contacts and occlusions. Reconstituting these incomplete data is a difficult problem. A new method for reconstructing markers' trajectories has been developed, base on a new motion representation. This one relies on a graph encoding the spatial relations between markers. Assuming that there is a spatial and temporal correlation between the displacement of a marker and its neighbors, the use of the discrete Laplacian operator to reconstruct the missing data is proposed. Each posture of the motion is described by a graph whose vertices are characterized by differential information and some edges are associated with distance constraints. Differential information is used to preserve the spatial relationships between markers, while distance constraints enable to preserve the length of some edges. This method

provides realistic reconstruction of trajectories, even when there are strong signal disruptions [15].

6.2.3 Synthesis of human motion by machine learning methods: a review

With the increasing number of virtual applications requiring human life-like characters, the computer animation domain has been challenged to provide algorithms and programs capable to define the behaviors of virtual characters through space and time. However, as many of these applications have real humans as final users and humans are easily prone to perceive the most subtle errors in human-like motions, computationally correct animations are not sufficient. In order to keep the user engaged with the application, natural and realistic looking animations must be produced. One of the best ways to animate virtual humans with human-looking motions is to record the movements of human actors. Yet, capturing good quality data is an expensive and time-consuming process. Under these circumstances, it is desirable to re-use y available data instead of capturing new examples everytime a new application is developed. However, modyfying the data to re-use it in different scenarios has proven to be a difficult task.

One appealing solution to these limitations is to combine human motion capture data with machine learning techniques. By doing so, it is possible to benefit from all the knowledge and experience of a research area which purpose is to study data and to construct models that generalize well outside of the examples. Through machine learning principles and techniques, it is possible to model the process that produced the registered motions and use this model to generate new data. The new data is consistent with the examples, with the user's expectations and constraints, and with the context in which it will be used. Furthermore, machine learning methods can capture all the subtleties of human movement and generate more expert gestures while maintaining the style, expressiveness and semantic inherent to human actions. In [9], we presented a state-of-the-art of the learning-based methods used in character animation.

6.2.4 Character Animation, Perception and Simulation

Skinning, also called Linear Blend Skinning (LBS) or Skeleton Subspace Deformations (SSD) is the standard technique for character animation to date. The character is composed of an internal skeleton and a polygonal mesh tied together through a rigging process. Although very efficient, this technique suffers several drawbacks including the well-known collapsing joint and candy wrapper effects. In addition, due to the nature of the algorithm effect itself it is not directly possible to simulate muscle bulging, dynamics or dynamic wrinkles of skin or clothes.

We have proposed a novel method for deforming the skin of 3D characters in real-time using an underlying set of muscles [5]. We use a geometric model based on parametric curves to generate the various shapes of the muscles. Our new model includes tension under isometric and isotonic contractions, a volume preservation approximation as well as a visually accurate sliding movement of the skin over the muscles. The deformation of the skin is done in two steps: first, a skeleton subspace deformation is computed due to the bones movement; then, vertices displacements are added due to the deformation of the underlying muscles.

We have tested our algorithm with a GPU implementation. The basis of the parametric primitives that serve for the muscle shape definition is stored in a cache. For a given frame, the shape of each muscle as well as its associated skin displacement are defined by only the splines control points and the muscle's new length. The data structure to be sent to the GPU is thus small, avoiding the data transfer bottleneck between the CPU and the GPU. Our technique is suitable for applications where accurate skin deformation is desired as well as video games or virtual environments where fast computation is necessary.

However, more accurate does not mean perceived as more realistic. The progress in character animation and rendering that has been made in the last twenty years has brought us to the uncanny valley. To climb out of there, the current trend is to include perception into the rending/animation algorithms. To derive new perceptually based algorithms, perceptual studies are conducted to find out what effects are perceived by the human eye/brain, and in which conditions.

When creating a real-time character animation (e.g. for video games), fine grain details such as cloth wrinkles are very often omitted due to the limitation of available resources. However, we believe that they are crucial to increase realism, hence user immersion. We have conducted perceptual studies of dynamic wrinkles on clothes to show that they indeed have an important impact on realism [6]. The models, animations and textures we used are game realistic and the wrinkling algorithm is fast enough to be used in a game. We have found that parameters influencing the perception of dynamic wrinkles include movement type and speed, viewing angle, and texture colors. Our results should be useful for animators or game designers to help them add wrinkles only where they are necessary, and activate them only when they have an impact on realism.

Characters necessarily evolve in an environment and interact with objects. We have worked on a novel physically based algorithm that simulates the deformation of paper when it burns [13]. We use a particle system to represent the fire and a mass-spring system coupled to a heat propagation solver to deform the polygonal mesh representing the paper sheet. When burnout, the paper becomes non-elastic and fractures automatically occur where the stress is important. By tuning the physical parameters of size, grammage, density, dimensional stability, specific heat and thermal conductivity, we are able to simulate the crumpling and burning of various types of paper.

6.3 Expressive Speech

Participants: Nelly Barbot, Vincent Barreaud, Jonathan Chevelu, Arnaud Delhay, David Guennec, Gwénolé Lecorvé, Sébastien Le Maguer, Damien Lolive, Claude Simon.

Within this topic, we have developed a software that is an original contribution that allows to represent in a joint manner several description levels of a speech utterance. ROOTS[BBB+11] is particularly well adapted to the description of speech corpora, is a priori independent from

[BBB+11] N. Barbot, V. Barreaud, O. Boëffard, L. Charonnat, A. Delhay, S. Le Maguer, D. Lolive, "Towards a Versatile Multi-Layered Description of Speech Corpora Using Algebraic Relations", *in : Conference of the International Speech Communication Association (Interspeech)*, p. 1501–1504, Florence, Italie, 2011.

the language, and can be easily extended to new descriptions. Two significant notions support the software architecture: first, sequences guarantee a time-based order relation, and second, relations allow to connect the set of sequences. One originality of ROOTS resides in an algebraic modeling of relations between sequences (technically, using a relation is just accessing a matrix structure). Such a model allows to compose very efficiently (both in terms of spatial and time complexity) new relations that can be missing from the description. Thus, for example, it is possible to create a relation from a word sequence towards an allophone sequence only by using two intermediate relations which are word/syllable and syllable/allophone relations. During the software specification, focus has been put on a serialization process towards JSON. Each object is then responsible for its own data. This solution enables us to think about the use of ROOTS in applications with a data flow architecture where processes feed a unified and coherent structure.

The first prototype has been completely rewritten in C++ in order to improve his efficiency and speed. This API is now widely used in the team and is growing as new functionalities are added depending of our new needs. For example, in the context of the ANR project Phorevox, special needs for corpus reduction appeared, equally in the context of the development of the TTS system for which we need to extract information from a speech corpora to build an optimized data structure. A use case has been published for the construction of a corpora in French from an audiobook in [BCLM+12].

Concerning the extension of ROOTS to new data, some progress has been done concerning the representation of named entities and prosodic information. For the latter, we have added the possibility to add prosodic structure of an utterance as a sequence of prosodic phrases (tree structures).

In particular, this year we have extended ROOTS to allow an easy manipulation of large corpus by introducing the concepts of corpus and layers of information. Some improvements have also been made concerning the speed of the library allowing to treat large data amounts within a reasonable time (creating a corpus from 1000 ebooks takes 2 days knowing that most of this time is devoted to disk access).

Finally, the registration of the library to the french APP is under progress and the software is available on the following url: `http://roots-toolkit.gforge.inria.fr`.

6.3.1 Optimal corpus design

Set covering algorithms are efficient tools for solving an optimal linguistic corpus reduction. This year we have extended this approach to use it in the frame of the ANR project PHOREVOX. The main objective was twofold:

- build a large corpus of free ebooks automatically annotated and stored in the ROOTS format

[BCLM+12] O. BOËFFARD, L. CHARONNAT, S. LE MAGUER, D. LOLIVE, G. VIDAL, "Vers une annotation automatique de corpus audio pour la synthèse de parole (Towards Fully Automatic Annotation of Audio Books for Text-To-Speech (TTS) Synthesis) [in French]", *in : Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, ATALA/AFCP, p. 731–738, Grenoble, France, 2012.

- extend the corpus reduction approach to include features linked to the task to achieve such as complex linguistic features (POS tags, linguistic context, etc.)

The final application targeted is to automatically generate exercices for children and teachers who use the PHOREVOX platform. Exemples of exercices are dictations, phonological oppositions exercices, and rhymes.

Moreover, the optimal corpus design goal is to design corpora that are well-adapted to a given task or speaker. Consequently, the end of the process is the generation of a speech corpus usable to build a high quality voice with a text-to-speech system. This quality depends on different aspects such as speaker's voice, recording quality, annotation accuracy and also phonological richness. In order to limit the recording and post-processing costs as well as to ensure voice homogeneity, the record script has to be of a reasonable size while guaranteeing phonological covering. Designing a rich corpus with a minimal size has been widely studied and the most used strategy consists in reducing a huge text corpus according to set covering constraints before the recording phase.

We have started a study to evaluate the impact of the corpus reduction on the quality of the synthesized voice. The goal is to determine the set of events to cover to improve the voice generation process and the best compromise in case of partial covering. The methodology involves speech corpora already recorded on which we apply the reduction process. At the end, the output of the text-to-speech system based on a reduced corpus is evaluated using objective measures and listening tests. The quality loss between a whole corpus and a reduced one following three strategies is investigated: the first one produces a covering of units described by the features used during the selection stage in the speech synthesis system. The second one simply uses diphone covering that is completed up to the size of the first one by a random of syntagms. Finally, the third one is a simple diphone covering used as a baseline system. The used reduction algorithm is *ASA*. Without surprise, systems using whole corpora obtain the best marks while systems using minimal corpora are judged the worst. The other systems are between those boundaries and are indistinguishable. Those results are confirmed by objective measures. The experiments tend to show two results that explain difficulties in the standard approach of corpus reduction. On the one hand, the quality loss induced by corpus reduction can, in some extend, be attributed to the lack of long units. On the other hand, covering every rare units leads to inefficient solutions compared to random selection.

6.3.2 Pronunciation modeling

Pronunciation modeling consists in characterizing and predicting the phonetic form of a given graphemic, i.e., orthographic, input utterance. Such tasks can be either addressed through

dictionary lookup [dCP98,Wei98], rule-based [TP91,Bec01] or statistical methods [BN08,Cla09,LDM09, IFJ11].

This year, the team EXPRESSION has started preliminary work in this field in order to address phonetic variabilities of speech implied by accents, speaking styles and emotions. To do so, 2 statistical modeling approaches have been investigated, namely joint ngrams [BN08] and conditional random fields [IFJ11] (CRFs), in order to be able to apply standard adaptation and rescoring techniques. A baseline phonetization system has been developed based on the French pronunctiation lexicon MHATLex [PDC00]. This system is based on CRFs because this type of model allows dependencies between phonemes and many acoustic and linguistic features. Moreover, the system relies on the formalism of weighted finite state transducers in order to enable working on phonetization lattices, whereas many traditional systems simply output lists of pronunciation hypotheses. More generally, the model has been designed independently of the language, easily extensible to various features.

Based on the developed phonetizer, the Ph.D. thesis of Raheel Qader will start on January, 1st 2014. This work will mainly focus on the pronunciation variant modeling for the sake of expressive synthesis speech. Moreover, work will also be carried out on speaking style pronunciation adaptation in French. Whereas the main application on this activity is focused towards speech synthesis (French and English), extensions to automatic speech recognition should also be thought in the future.

6.3.3 Optimal speech unit selection for text-to-speech systems

This work has been undertaken principally during the Master's Internship of David Guennec and is now going on during his PhD, started in october 2012. With this work, we have developed a unit-selection Text-To-Speech synthesis system. Since the quantity of information used by such a system is very high (several hours of speech), the problem of finding the best sequence

[dCP98] M. de Calmès, G. Pérennou, "BDLex: a lexicon for spoken and written French", *in : Proceedings of 1st International Conference on Langage Resources & Evaluation*, p. 1129–1136, 1998.

[Wei98] R. Weide, "The CMU pronunciation dictionary, release 0.6", Carnegie Mellon University, 1998.

[TP91] J. Tiboni, G. Perennon, "Phonotypical transcription through the GEPH expert system", *in : Second European Conference on Speech Communication and Technology*, 1991.

[Bec01] F. Bechet, "LIAPHON - Un système complet de phonetisation de textes", *Traitement Automatique des Langues (T.A.L.) edition Hermes 42*, 1, 2001.

[BN08] M. Bisani, H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion", *Speech Communication*, 2008.

[Cla09] V. Claveau, "Letter-to-Phoneme Conversion by Inference of Rewriting Rules", *in : Proc. the Conf. of the Intl Speech Communication Association (Interspeech)*, p. 1299–1302, Brighton, UK, Sept. 2009.

[LDM09] A. Laurent, P. Delï¿½glise, S. Meignier, "Grapheme to phoneme conversion using an SMT system", *in : Interspeech 2009, 10th Annual Conference of the International Speech Communication Association*, p. 708–711, Brighton, Angleterre, Septembre 2009.

[IFJ11] I. Illina, D. Fohr, D. Jouvet, "Multiple Pronunciation Generation using Grapheme-to-Phoneme Conversion based on Conditional Random Fields", *in : Proceedings of the International Conference on Speech and Computer (SPECOM)*, 2011.

[PDC00] G. Pérennou, M. De Calmes, "MHATLex: Lexical Resources for Modelling the French Pronunciation.", *in : LREC*, 2000.

of units of speech can be very time-consuming. Consequently, a number of optimizations have been done.

First, the data structure used internally by the system to represent the speech units is optimized in such a way that memory accesses is limited (for example, a unit have information about its context, its linguistic or acoustic properties, etc.). Moreover, the way information is accessed in the corpora makes use of computationally efficient structures.

As the problem of finding the best sequence of units can be expressed as a graph, we have implemented a $A^\star$ algorithm to perform it. For us, this algorithm is a way of relaxing the Markovian hypothesis which is traditionally made in state of the art systems. We have also implemented a cost function based on acoustic and prosodic features. For the moment the structure of this cost function has been manually fixed. This year we have work on the evaluation of the behavior of the system by exploring the ranking of the paths given by the A* algorithm as well as several cost functions. The results, that have been submitted and are under review, show that the ranking given by the algorithm is discriminant enough between the different paths. Consequently, their is no clear difference between the first path and the 100th.

Future work will try to assess if the cost function is discriminant enough by trying several strategies exploring the following paths and the extreme case of reverting the cost function to find the worse path. Moreover, to have results comparable with the literature, we plan to implement the classical Viterbi approach and compare it with $A^\star$ in terms of efficiency and performance.

Finally, the perceptive evaluations concerning the cost function, in the best case, show that concatenations quality is most of the time of good quality, which is not the case for prosody. Another part of our future investigations will concern the introduction of a prosodic model and prosodic features into the cost function to answer the following question: what is relevant to improve the prosody naturalness and gain more expressive control for the system output ?

6.3.4 Experimental evaluation of a statistical speech synthesis system

This work is covered within the framework of the PhD thesis of Sébastien Le Maguer.

During the past fifteen years, statistical speech synthesis has taken an important place in the speech synthesis methods. The most representative system is HTS which uses HMM to model speech, represented by parameters extracted from annotated signal. To train models, the system needs corpora which are composed by large amount of data and a complete description of each speech segment. This description contains information from different levels (acoustic, phonology, linguistic). These descriptive features are positional informations (like the position of the syllable in the word), prosodic informations (like the accentuation of the syllable) or linguistic informations (like the part-of-speech of the current word). A complete list of these descriptive features can be found in the 2000 Tokuda's paper[TZB00].

In order to create such a corpus for the french language, we consider that it is necessary to assess the influence of the descriptive features on the speech parameter generation stage.

[TZB00] K. Tokuda, H. Zen, A. W. Black, "An hmm-based speech synthesis system applied to english", *in : Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, p. 227–230, 2000.

Therefore, we have first finalized experiments in order to evaluate the impact of these features on the spectrum parameter generation. The method relies on GMM to model the generated spectrum space and enables to assess the likelihood of the reference data according to this space. Additionaly to this objective evaluation, listening tests have been conducted[LMBB12].

Since this GMM based methodology is global evaluation approach, we have also considered a pair-wised frame distance methodology which allows a finer analysis. We have complemented the previous works by applying these methodologies to the spectral and F0 coefficients generated by HTS and those stemming from the analysis-by-synthesis processing. The results have been published in [14]. Finally, the duration and aperiodicity modeling have also been evaluated in a similar way. All these results have been detailled in the PhD dissertation of S. Le Maguer, who has defended his PhD thesis in july 2013 [1].

6.4 Miscellaneous

Participants: Nelly Barbot.

In collaboration with Laurent Miclet (Dyliss IRISA) and Henri Prade (IRIT), we have defined and studied analogical proportions in the general setting of lattices, and more particularly of distributive lattices. Analogical proportions are statements involving four entities, of the form 'A is to B as C is to D'. They play an important role in analogical reasoning. Their formalization has received much attention from different researchers in the last decade, in particular in a propositional logic setting. Analogical proportions can turn out to be an original approach to classification in machine learning and their use shows promising in the task of language translation.

We have studied the decomposition of analogical proportions in canonical proportions, as well as the resolution of analogical proportion equations, which plays a crucial role in reasoning. These works have been published in [8, 7]. These works are supported by the CNRS PEPS APA 2013.

6.5 Expression in textual data

Participants: Nicolas Béchet, Giuseppe Bério, Pierre-FranÃ§ois Marteau, Jeanne Villaneau.

6.5.1 Text mining

co-clustering and co-classification of bilingual data : The quantitative notion of comparability (based on a bilingual dictionary) between two sets of bilingual documents leads to consider the bipartite graph induced by the comparability relation. When quantitative similarity relations exist in each of the considered linguistic spaces, one may ask if and how the

[LMBB12] S. Le Maguer, N. Barbot, O. Boëffard, "Evaluation segmentale du système de synthèse HTS pour le français (Segmental evaluation of HTS) [in French]", *in : Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, ATALA/AFCP, p. 569–576, Grenoble, France, 2012.

Figure 4: Improve the quality of textual data for the purpose of ontology building

comparability mapping affects local similarities. We have initiated the study of the conceptual and practical consequences of such a similarity-comparability connection by developing an algorithm (Hit-ComSim)[17, 20] based on the concept of similarity induced by the topology of the comparability graph. Doing so, we end-up improving jointly the accuracy of classification and clustering algorithms performed in each of the two linguistic spaces, as well as the mapping of comparable clusters that are obtained. First experiments carried on data collected on news wires RSS feeds show clear improvements in clustering and classification accuracy in both considered linguistic spaces (English and French).

We have initiated the study of the conceptual and practical consequences of such a connection similarity-comparability by developing an algorithm (Hit-ComSim) based on the principle of similarity induced by the topology of the comparability graph.

Sequential pattern extraction : We focus on the use of sequential patterns with textual documents. Sequential patterns can be useful in many tasks like log analysis, recommendation, event detection, stylistic analysis. In recent work we have focused on the use of sequential patterns on geographical data and biomedical texts. This work is based on algorithms allowing to extract sequential patterns with multiple kinds of constraints like numeric or linguistic constraints. We develop this new axis since September 2013 and many work as still in progress and unpublished like algorithmic problem with sequential pattern extraction, stylistic and thematic characterisation of texts using sequential patterns, link geographical data resulting from texts with satellites images.

6.5.2 Knowledge representation

For bridging the gap between techniques usable for processing and analyzing texts and information for filling ontology contents (basically concepts, relationships, axioms) we have undertaken an approach working both *a priori* and *a posteriori* (Figure 4). *A priori* way of working concerns how to improve the quality of textual data for the purpose of ontology building. *A posteriori* way of working concerns how to detect and possibly solve known problems in automatically or semi-automatically built ontologies by using available tools.

We follow the approach alongside two ways. According to the first way, ontology building is manually driven by human starting from available texts. However, a predefined meta-ontology framework constraints humans whenever they introduce concepts and relationships. According to the second way, existing tools for automatically or semi-automatically building ontologies from texts are analyzed, compared and selected according to their "performance" within a test-bed platform [3]; then, possible problems and improvements of resulting ontologies are

investigated (some preliminary results can be found in [11]).

6.5.3 Text processing

Corpus creation and annotation : We participated in the development of the first French spoken corpus annotated in coreference, with the LI and LLL teams of Orleans and Tours [18]. The corpus is form now on freely available (Creative Commons BY-NC-SA). This research allowed to complete previous experimental studies on the reliability of the inter-coder agreement metrics (publication EACL 2014).

Text summarization : An other research related to automatic multitext summarization was initiated and some preliminary questions investigated: system evaluations, automatic detection of the concepts in a given domain by using web data (wikipedia) [19], measuring similarity between sentences (publication TSD 2014).

7 Contracts and Grants with Industry

7.1 SIGN3D

Participants: Sylvie Gibet, Ludovic Hamon

.

The *SIGN3D* project is funded by the French Ministry of Industry (DGCIS: Direction Générale de la Compétitivité de l'Industrie et des Services, Program "Investissements d'avenir": usages, services and innovative contents). Three partners are participating to this project: two companies (Mocaplab: team leader, and Websourd), and one academic laboratory (IRISA). The subject concerns the editing of motion in French sign language using high definition gesture databases. The project aims at creating a range of innovative tools for the recording and the editing of captured motion of French Sign Language (LSF) content. The challenge is to design a complete workflow from the movement capture (including body and hand movements, facial expressions and gaze direction) to the restitution using concatenate synthesis applied on a 3D virtual signer. The project officially started in January 2012 and will end in October 2014.
`http://sign3d.websourd.org/`

7.2 INGREDIBLE

Participants: Pamela Carreño, Sylvie Gibet, Ludovic Hamon, Caroline Larboulette, Pierre-François Marteau.

The *INGREDIBLE* project project is funded by the French Research Agency (program ANR CONTINT. The partners are: LabSTICC (team leader), LIMSI-CNRS, IRISA, Virtualys, Final users (DEREZO, Brest; STAPS lab., Orsay.

The goal of the INGREDIBLE project is to propose a set of scientific innovations in the domain of human/virtual agent interaction. The project aims to model and animate an autonomous virtual character whose bodily affective behavior is linked to the behavior

of a human actor. The application domain studied in Expression concerns the analysis and synthesis of expressive gestures in interactive theatrical scenarios.

The project started in September 2012. Pamela Carreño, who is a PhD student working in the project, has proposed a methodology for building a corpus of full-body theatrical gestures on the basis of magician tricks enriched with emotional content. The constructed corpus and expressive sequences of actions have been validated through several perceptual studies focusing on the complexity of the produced movements as well as the recognizability of the produced emotions.

7.3 PHOREVOX

Participants: Nelly Barbot, Jonathan Chevelu, Damien Lolive.

Expression is leader of a ANR CONTINT project named Phorevox and accepted on the 16th december, 2011. Phorevox aims to promote the use of high quality speech synthesis to help in learning french. The consortium is made of IRISA/Expression, Voxygen (France Telecom spin off), CREAD (didactic and tests), LLF Paris VII (linguistics) and Zeugmo (web platform).

The project has official begun the 1st of june 2012. During the first semester of the project, the main means deployed for the project has concerned the organisation of the project as well as the deployment of collaborative tools. We also have recruited a research engineer for the whole length of the project. His role is to work on the problematics concerning the IRISA contribution, to assure the project coordination and the technical support on collaborative tools. One of the first consortium actions was to build up tools simplifying exchanges between the group members and to improve project visibility.

Practically, some difficulties to recruit an engineer have postponed the real technical start of the project. This delay will be taken into account to postpone the end of the project too (6 months accepted by ANR).

At the end of November, a mid-term presentation has been done in front of the ANR committee. During this presentation, we have shown a demonstration video of the prototype with several exercices (dictation, phonological opposition, word segmentation) including an adapted speech synthesis voice.

We are now carrying on the development of the project with some experiments planned in real conditions to assess platform ergonomy and to have the first feedbacks concerning the suitability of the approach to help children learn phonology.

8 Other Grants and Activities

8.1 International Collaborations

- **Horizon 2020 project proposal:** The team collaborates in writting a European project proposal with Orange Labs since October 2013. This proposal will be submitted as part of the ICT 2014 call of the Horizon 2020 Framework Programme of the European Commission. Precisely, the consortium gathers the following partners (and countries):

Athens International Airport (Greece), Athens University of Economics and Business (Greece), CEA (France), Fondazione Bruno Kessler (Italy), GeoMobile (Germany), Idiap Research Institute (Switzerland), and Orange Labs (France). The addressed topic will be "multimodal and natural computer interaction" and will focus on smart conversational virtual assistants in international transport hubs and complex environments with an emphasis on multimodality and multilinguality. The EXPRESSION team would address the speech synthesis issues within this project.

9 Dissemination

9.1 Involvement in the Scientific Community

- Pierre-François Marteau is a member of various program committee for international (ISDA, SOCPAR, SOCO, IISPA) conferences and served as a reviewer in international journals (IEEE TPAM, IEEE TNN, IEEE TKDE, MIT JMLR, Elsvier JPR). He serves as an expert for French Ministry of Research (CIR/JEI) and ANRT (CIFRE). He was member of a thesis committee at Nantes University, LINA. He is member of the Strategic Orientation Committee at IRISA and member of the scientific committee at Université de Bretagne Sud.

- Sylvie Gibet serves as a reviewer for the IEEE Transactions on Affective Computing, for the journal on Multimodal User Interfaces, and for the journal IEEE Transactions on Visualization and Computer Graphics. She has served as a reviewer for the first international conferences TiGer (Tilburg Gesture Research Meeting combining gesture and speech in interaction), the Third International Symposium on Sign Language Translation and Avatar Technology (SLTAT), and the international conference GRAPP. She has given a lecture at McGill University in October 2013. She has been the advisor of Thibaut Le Naour who defended his PhD thesis in December 2013.

- Arnaud Delhay is an elected member of the 'Commission Recherche' (Research comittee) of the IUT of Lannion.

- Damien Lolive is an elected member of the 'Conseil Scientifique' (Scientific council) of ENSSAT, Lannion. He also serves as a reviewer for the IEEE Transactions on Speech and Language processing and for the International conference of the International Speech Communication Association (Interspeech).

- Gwénolé Lecorvé serves as a reviewer for the IEEE Signal Processing Letters and for the International conference of the International Speech Communication Association (Interspeech).

9.2 Teaching

- Sylvie Gibet teaches the following computer science courses at the faculty of sciences, Université de Bretagne Sud: functional programming and algorithmic in License level, and multimedia digital processing and computer animation in Master level (master WMR).

- Jean-François Kamp teaches human-computer interaction, programming at the computer science department of IUT Vannes. He is responsible for student internships.
- Gwénolé Lecorvé teaches the following computer science courses at École National Supérieure des Sciences Appliquées et de Technologie (ENSSAT): operating systems in Licence level, distributed algorithmics, and artificial intelligence in Master level.
- Damien Lolive teaches the following computer science courses at École National Supérieure des Sciences Appliquées et de Technologie (ENSSAT): object-oriented programming in Licence Level, and XML, compilers architecture and formal languages theory in Master Level.
- Arnaud Delhay teaches databases and web programming at IUT of Lannion in Licence level and theory of computational complexity and web server programming at École National Supérieure des Sciences Appliquées et de Technologie (ENSSAT) in Master Level.
- Pierre-François Marteau teaches programming, logics, and information retrieval courses in Licence and Master levels, at Ecole Nationale Supérieure de Bretagne Sud.
- Gildas Ménier teaches various computer sciences courses at the faculty of sciences, Université de Bretagne Sud
- Jeanne Villaneau teaches various computer sciences courses at Ecole Nationale Supérieure de Bretagne Sud.

9.3 Conferences, workshops and meetings, invitations

Nelly Barbot has been a member, as co-director, of the jury of the PhD thesis of S. Le Maguer [1]: *Évaluation expérimentale d'un système statistique de synthèse de la parole, HTS, pour la langue française*, Thèse de l'université de Rennes 1, defended the 5th of july 2013.

9.4 Graduate Student and Student intern

The team hosted the internship of Andrei Zene from the Technical University of Cluj-Napoca, Romania, during 3 months. Andrei has worked on the software tool ROOTS and especially on the display function to produce a readable and user-friendly way of displaying information extracted from multi-sequence structured data.

10 Bibliography

Doctoral dissertations and "Habilitation" theses

[1] S. LE MAGUER, *Évaluation expérimentale d'un système statistique de synthèse de la parole, HTS, pour la langue française*, Thesis, Université de Rennes 1, July 2013, `http://tel.archives-ouvertes.fr/tel-00913565`.

[2] T. Le Naour, *Utilisation des relations spatiales pour l'analyse et l'édition de mouvement*, PhD Thesis, 2013.

Articles in referred journals and book chapters

[3] T. Gherasim, M. Harzallah, G. Berio, P. Kuntz, "A comparative analysis of some approaches for automatic construction of ontologies from textual resources", *in : Advances In Knowledge Discovery and Management*, Springer, 2013, p. 177–201.

[4] T. Le Naour, N. Courty, S. Gibet, "Spatio-temporal coupling with the 3D+t motion Laplacian", *Computer Animation and Virtual Worlds*, 2013, p. CAV1518, `http://hal.archives-ouvertes.fr/hal-00861779`.

[5] J. Ramos, C. Larboulette, "A Muscle Model for Enhanced Character Skinning", *Journal of WSCG 21*, 2, jul 2013, p. 107–116, `http://zabador.free.fr/Publications/2013/RL13`.

Publications in Conferences and Workshops

[6] F. J. Alcon Palazon, D. Travieso, C. Larboulette, "Influence of Dynamic Wrinkles on the Perceived Realism of Real-Time Character Animation", *in : Proceedings of the International Conference on Computer Graphics, Visualization and Computer Vision (WSCG)*, M. M.Oliveira, V. Skala (editors), *Annual Conference Series, Full Papers*, Eurographics, Vaclav Skala â Union Agency, p. 95–103, c/o University of West Bohemia, Univerzitni 8, CZ 306 14 Plzen, Czech Republic, jun 2013, `http://zabador.free.fr/Publications/2013/ATL13`.

[7] N. Barbot, L. Miclet, H. Prade, "Analogical proportions and the factorization of information in distributive lattices", *in : 10th International Conference on Concept Lattices and Their Applications (CLA)*, La Rochelle, France, October 2013, `http://hal.inria.fr/hal-00908005`.

[8] N. Barbot, L. Miclet, H. Prade, "Proportions analogiques et factorisation de l'information dans les treillis distributifs", *in : Journées d'Intelligence Artificielle Fondamentale (JIAF)*, Aix en Provence, France, June 2013, `http://hal.inria.fr/hal-00908020`.

[9] P. Carreno-Medrano, S. Gibet, P.-F. Marteau, "Synthèse de mouvements humains par des méthodes basées apprentissage : un état de l'art", *in : AFIG 2013*, Limoges, France, November 2013, `http://hal.archives-ouvertes.fr/hal-00913056`.

[10] J. Chevelu, G. Lecorvé, D. Lolive, "ROOTS: a toolkit for easy, fast and consistent processing of large sequential annotated data collections", *in : Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 2014, `http://hal.inria.fr/hal-00974628`.

[11] R. Chulyadyo, M. Harzallah, G. Berio, "Core Ontology based Approach for Treating the Flatness of Automatically Built Ontology", *in : Proceedings of KEOD*, p. 316:323, Portugal, 2013.

[12] L. Hamon, S. Gibet, S. Boustila, "Édition interactive d'énoncés en langue des signes française dédiée aux avatars signeurs", *in : TALN - 20ème conférence du Traitement Automatique du Langage Naturel 2013*, Sables d'Olonne, France, June 2013, `http://hal.inria.fr/hal-00911629`.

[13] C. Larboulette, P. Quesada Barriuso, O. Dumas, "Burning Paper: Simulation at the Fiber's Level", *in : Proceedings of the ACM SIGGRAPH Conference on Motion in Games, ISBN: 978-1-4503-2546-2*, ACM New York, NY, USA Â©2013, ACM Press, p. 25–30, nov 2013. http://dx.doi.org/10.1145/2522628.2522906, `http://zabador.free.fr/Publications/2013/LQD13`.

[14] S. Le Maguer, N. Barbot, O. Boëffard, "Evaluation of contextual descriptors for HMM-based speech synthesis in French", *in : ISCA Speech Synthesis Workshop (SSW8)*, Barcelone, Spain, 2013, http://hal.inria.fr/hal-00987809.

[15] T. Le Naour, N. Courty, S. Gibet, "Utilisation des relations spatiales pour la reconstruction de trajectoires de marqueurs issues de la capture de mouvement", *in : AFIG 2013*, Limoges, France, November 2013.

[16] F. Lefebvre-Albaret, S. Gibet, A. Turki, L. Hamon, R. Brun, "Overview of the Sign3D Project High-fidelity 3D recording, indexing and editing of French Sign Language content", *in : Third International Symposium on Sign Language Translation and Avatar Technology (SLTAT) 2013*, Chicago, États-Unis, October 2013, http://hal.archives-ouvertes.fr/hal-00914661.

[17] P.-F. Marteau, G. Ménier, "Similarités induites par mesure de comparabilité : signification et utilité pour le clustering et l'alignement de textes comparables", *in : Actes 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 515–522, Les Sables d'Olonne, France, June 2013, http://hal.archives-ouvertes.fr/hal-00873811.

[18] J. Muzerelle, A. Lefeuvre, J.-Y. Antoine, E. Schang, D. Maurel, J. Villaneau, I. Eshkol, "ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement", *in : Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 555–563, Les Sables d'Olonne, France, 2013.

[19] H. H. Vu, J. Villaneau, F. Saïd, *in : Actes des JournÃ©es Internationales de Linguistique de Corpus 2013 (JILC 2013)*, Lorient, France, 2013. to appear.

Miscellaneous

[20] G. Ke, P.-F. Marteau, G. Ménier, "Improving the clustering or categorization of bi-lingual data by means of comparability mapping", October 2013, http://hal.archives-ouvertes.fr/hal-00958730.