Project-Team EXPRESSION

Expressiveness in Human Centered Data/Media

Vannes, Lannion & Lorient

Activity Report

2016

Contents

1 Team

1.1 Composition

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

Administrative assistant
Sylviane Boisadan, Université Bretagne Sud
Angélique Le Pennec, Université de Rennes 1

Permanent members
Nelly Barbot, Associate professor, Université de Rennes 1
Nicolas Béchet, Associate professor, Université Bretagne Sud
Giuseppe Bério, Professor, Université Bretagne Sud
Jonathan Chevelu, Associate professor, Université de Rennes 1
Arnaud Delhay-Lorrain, Associate professor, Université de Rennes 1
Sylvie Gibet, Professor, Université Bretagne Sud
Caroline Larboulette, Associate professor, Université Bretagne Sud
Gwénolé Lecorvé, Associate professor, Université de Rennes 1
Damien Lolive, Associate professor, Université de Rennes 1
Gildas Ménier, Associate professor, Université Bretagne Sud
Jeanne Villaneau, Associate professor, Université Bretagne Sud

Associate members
Vincent Barreaud, Associate professor, Université de Rennes 1
Elisabeth Delais-Roussarie, Senior researcher, CNRS/LLF
Jean-François Kamp, Associate professor, Université Bretagne Sud
Farida Said, Associate professor, Université Bretagne Sud

Non-permanent members
Nehla Ghouaiel, ATER, Université Bretagne Sud
Rémi Kessler, Post-doctoral researcher, Université Bretagne Sud (since December 2016)
Saeid Soheily khah, Post-doctoral researcher, Université Bretagne Sud (since November 2016)
Marie Tahon, Post-doctoral researcher, Université de Rennes 1
Gaëlle Vidal, Engineer, Université de Rennes 1 (until December 2016)

PhD students
Sandy Aoun, Université de Rennes 1, ARED/CG22, 1st year
Yonatan Carranza Alarcón, Université Bretagne Sud, 1st year
Pamela Carreno, Université Bretagne Sud, ANR INGREDIBLE, 3rd year
Lei Chen, Université Bretagne Sud/Univ. McGill, ARED, 3rd year
Marc Dupont, Université Bretagne Sud, Thèse CIFRE Thales, 2nd year
Cédric Fayet, Université de Rennes 1, 2nd year
David Guennec, Université Rennes 1, ATER, 4th year, defended on September 2016
Lucie Naert, Université Bretagne Sud, CDE, 1st year
Nicolas Bloyet, Université Bretagne Sud, Thèse CIFRE Seed, 1st year

Stefania Pecóre, Université Bretagne Sud, 2nd year
Raheel Qader, Université de Rennes 1, MENESR, 3rd year
Clément Reverdy, Université Bretagne Sud, CDE+Labo, 3rd year
Aghilas Sini, Université de Rennes 1, LABEX EFL/ANR SynPaFlex, 1st year
Hai Hieu Vu, Université Bretagne Sud, 5th year, defended on January 2016

Master students
Sandy Aoun, University of Lebanon

1.2 Evolution of the staff

The number of PhD students is slightly increasing. Sandy Aoun (Université de Rennes 1 (Lannion)) and Aghilas Sini (Université de Rennes 1 (Lannion)) have been hired as PhD students at IRISA, in December 2016. Lucie Naert and Nicolas Bloyet (Université Bretagne Sud (Vannes)) have been hired as PhD student at IRISA in October 2016 .

The contract of Gaëlle Vidal, recruited to work on the ANR SynPaFlex project, finished by the end of November.

The PhD defense of Vu Hai Hieu was held on 29th Januray 2016. Vu Hai Hieu holds since March 2016 a Postdoc position at CNRS-ATILF, Université of Lorraine.

Pamela Carreno has defended her PhD the 25th November 2016. She holds currently a Postdoc position at EXPRESSION (ATER teaching position at IUT of Vannes).

Jeanne Villaneau has defended her 'habilitation à diriger des recherches' the 9th of March 2016 in Vannes. She holds currently an emeritus professor position at Université Bretagne Sud.

2 Overall Objectives

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 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.

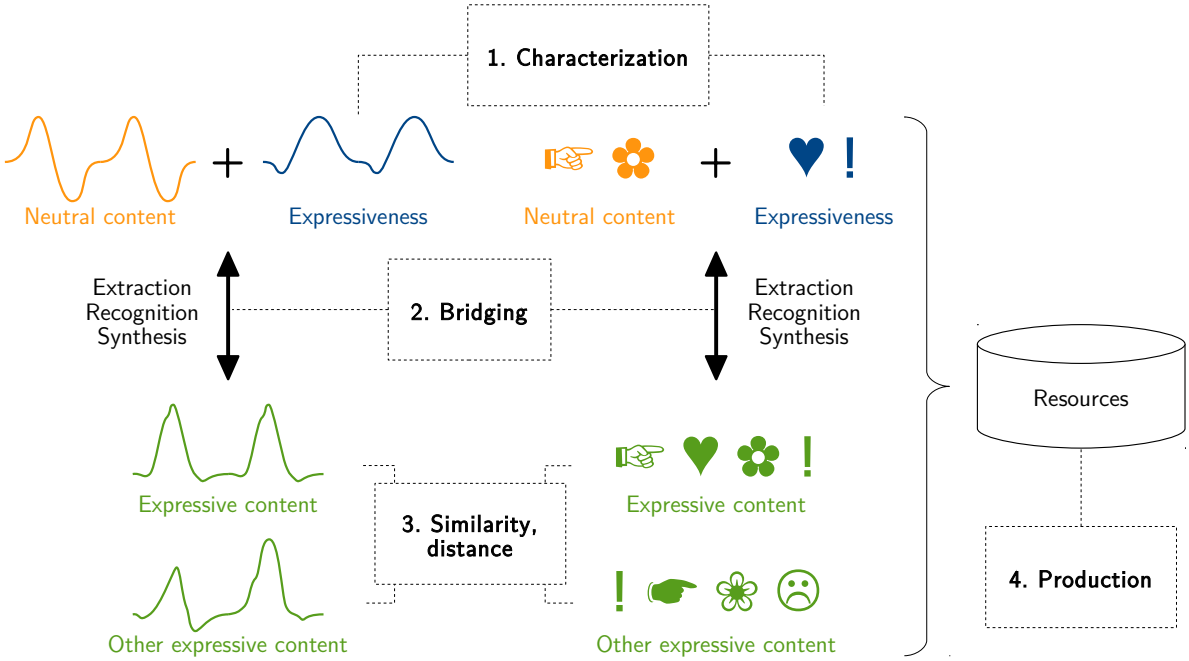

Figure 1: Overview of the main challenges considering both on continuous numerical (left) and discrete symbolic (right) data.

2.1 Main challenges addressed by the team

Four main challenges will be addressed by the team.

C1: The characterization of the expressiveness as defined above in human produced data (gesture, speech, text) is the first of our challenges. This characterization is challenging jointly the extraction, generation, or recognition processes. The aim is to develop models for manipulating or controlling expressiveness inside human or synthetic data utterances.

C2: Our second challenge aims at studying to what extent innovative methods, tools and results obtained for a given media or for a given pair of modality can be adapted and made cross-domain. More precisely, building comprehensive bridges between discrete/symbolic levels (meta data, semantic, syntactic, annotations) and mostly continuous levels (physical signals) evolving with time is greatly stimulating and nearly not explored in the different scientific communities.

C3: The third challenge is to address the characterization and exploitation of data-driven embeddings[1] (metric or similarity space embeddings) in order to ease post-processing of data, in particular to reduce the algorithmic complexity and meet the real-time or big-data challenges. The characterization of similarity in such embeddings is a key issue as well as the indexing, retrieval, or extraction of sub-sets of data relevant to user's defined tasks and needs, in particular the characterization of expressiveness and variability.

C4: The fourth challenge is to contribute to the production of resources that are required, in particular to develop, train and evaluate machine learning (statistical or rule-based) models for human

[1] Given two metric or similarity spaces (X, d) and (X', d'), a map $f : (X, d) \to (X', d')$ is called an embedding.

language data processing. These resources are mainly corpora (built from speech, text and gesture time series), dictionaries, and semantic structures such as ontologies.

All the addressed challenges are tackled through the development of models, methods, resources and software tools dedicated to represent and manage gesture, speech or textual data. Thus we consider a complete processing chain that includes the creation of resources (corpus, thesaurus, semantic network, ontology, etc.), the labeling, indexing and retrieval, analysis and characterization of phenomena via classification and extraction of patterns (mostly sequential).

These challenges also target multi-level aspects, from digital tokens to semantic patterns, taking into account the complexity, the heterogeneity, the multi-dimensionality, the volume, and the nature of our temporal or sequential data.

We are aiming at addressing these challenges in terms of development and exploitation of machine learning and pattern discovery methods for clustering, classification, interactive control, recognition, and production of content (speech signals, texts or gestures), based on different levels of representation (captured or collected data but also knowledge that is specific to the media or the considered application). Finally, both objective and subjective (perceptive) evaluations of these models are a key issue of the research directions taken by the EXPRESSION team.

2.2 Main research focus

Five thematic lines of research are identified to carry out this research.

RF1: Data acquisition – Gesture, speech or text data are characterized by high levels of heterogeneity and variability. Studying such media requires high quality data sets appropriate to a well defined and dedicated task. The data acquisition process is thus a crucial step since it will condition the outcomes of the team research, from the characterization of the studied phenomena, to the quality of the data driven models 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. This research focus addresses mainly the fourth challenge;

RF2: 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 main research direction. Such representations will provide efficient information access, support for database enrichment through bootstrapping and automatic annotation. This research focus contributes mainly to the second, third and fourth challenges;

RF3: 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). This research focus contributes mainly to the first and third challenges.

RF4: 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 adequate expressiveness, 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 parametrized 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). This research focus contributes mainly to the first and fourth challenges since generation can be seen also as a bootstrapping method. As parallels can be possibly drawn between expressive speech and expressive movement synthesis, the focus also contributes to the second challenge;

RF5: Use cases and evaluation – The objective is to develop intuitive tools and in particular sketch-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. As such, this focus contributes to the first challenge and has a direct impact on the fourth challenge. Furthermore, whereas many encountered sub-problems are machine learning tasks that can be automatically evaluated, synthesizing human-like data requires final perceptive (i.e., human) 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. This research focus contributes also to the second challenge;

3 Scientific Foundations

3.1 Expressive gesture analysis, synthesis and recognition

Thanks to advanced technologies such as new sensors, mobile devices, or specialized interactive systems, gesture communication and expression 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 underlying 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 or 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].

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

Motion perception through bodily expressions has also given rise to many work in nonverbal communication. In the psychology and neuroscience literature, recent studies have focused in particular on the recognition of emotion in whole body movements[THB06,dG06,CG07].

In computational sciences, many studies 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 studies rely on qualitative or quantitative annotations of motion clips[AFO03,MBS09], or propose relevant methods to create a repertoire of expressive behaviors[RBC98], 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

[THB06] L. Torresani, P. Hackney, C. Bregler, "Learning motion style synthesis from perceptual observations", *in : Advances in Neural Information Processing Systems*, 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*, *4738*, 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", ACM Press/Addison-Wesley Publishing Co., *in : SIGGRAPH'00: Proc. of the 27th annual conference on Computer graphics and interactive techniques*, 2000.

[HMBP05] B. Hartmann, M. Mancini, S. Buisine, C. Pelachaud, "Design and evaluation of expressive gesture synthesis for embodied conversational agents", *in : AAMAS*, 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*, 2000.

[Her03] A. Hertzmann, "Machine learning for computer graphics: A manifesto and tutorial", *in : Computer Graphics and Applications, 2003. Proc.. 11th Pacific Conference on*, 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, 2005.

[AFO03] O. Arikan, D. A. Forsyth, J. F. O'Brien, "Motion Synthesis from Annotations", *ACM Transactions on Graphics 22*, 3, July 2003, p. 402–08.

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

[RBC98] C. Rose, B. Bodenheimer, M. F. Cohen, "Verbs and Adverbs: Multidimensional Motion Interpolation Using Radial Basis Functions", *IEEE Computer Graphics and Applications 18*, 1998, p. 32–40.

the case when gestures are produced in the context of narrative scenarios, or expressive 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 any kind of variability.

Hence 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+09]. 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 analysis / synthesis scheme, corrected and refined through an evaluation loop (*e.g.*, [7]). Consequently, data-driven models, which incorporate constraints derived from observations, should significantly improve the quality and credibility of the gesture synthesis; furthermore, the analysis of the original or synthesized data by techniques of automatic segmentation, classification, or recognition models should improve the generation process, for example by refining the annotation and cutting movements into significant items. 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 Expressive speech analysis and synthesis

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 in 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.

[ACD+09] 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*, 2009.

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

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

[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, 2010.

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

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

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. 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 controlled 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[GR94,Abe95, IAML04,IMK+04,Bla07]. 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.

[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", IEEE, *in : Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *1*, 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.

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

[Sch01] M. Schröder, "Emotional Speech Synthesis : A Review", *in : Proc. 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*, 2009.

[SCK06] V. Strom, R. Clark, S. King, "Expressive Prosody for Unit-selection Speech Synthesis", *in : Proc. 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.

[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.

[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.

[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*, 2004.

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

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

The state of the art is 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 work exists, notably concerning corpus annotation[DGWS06], but for the moment, no language is sufficient to build up a complete editorial chain. This point constitutes an obstacle towards automatic or semi-automatic creation of high-quality spoken content.

The amount of work on the integration of expressiveness into TTS systems is in constant augmentation these last years. Most speech synthesis methods have been subject to extension attempts. In particular, we can cite the diphone approach[BNS02], the corpus-based approach[CRK07], or even the parametric approach[TYMK07]. Adding to this, several languages have been used: notably Spanish[ISA07], Polish[DGWS06], Japanese[TYMK07], English[SCK06], and French[AVAR06,LFV+11]. On our side, current activities in speech synthesis are conducted on French and English. Although other languages could be added to our system, there is currently no real scientific interest, unless a multilingual environment is required.

Beyond speech synthesis, some problems implied by the human expressiveness can also be found in other domains, but generally with an opposed point of view. In speaker processing and automatic speech recognition (ASR), 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 ASR again, the problem of word pronunciations is also important, especially when facing out-of-vocabulary words, i.e.words neither part of the training data nor of hand-crafted phonetized lexicons. Grapheme-to-phoneme converters are then needed to automatically associate one or several phonetizations to these

[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 : Proc. of the Eleventh Australasian International Conference on Speech Science and Technology*, 2006.

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

[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.

[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 : Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *4*, 2007.

[SCK06] V. Strom, R. Clark, S. King, "Expressive Prosody for Unit-selection Speech Synthesis", *in : Proc. 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*, 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)*, 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*, 2003.

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

words[BÓ1,BN08,IFJ11]. These tools are also used in TTS, needs in TTS and ASR are different. In ASR, the recall over generated pronunciations is maximized, that is the objective is to cover all possible pronunciations of a word to make sure that it will be recognized correctly. At the opposite in TTS, the precision is favored since only one phonetization will be uttered by the system in the end. Thus, extra work on pronunciation scoring and selection is necessary in TTS to improve generic grapheme-to-phoneme models. Other work aims at modeling disfluencies, i.e.errors within the elocution of a sentence, in order to help an ASR 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. Hence, synthesizing expressive speech requires to extend existing disfluency models to fit a generative process. Finally, emotion detection is also a subject of interest. In[LTAVD11], the authors are interested in emotion recognition from linguistic cues 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 based on given input classes of expressiveness.

In this context, the scientific goal of the team in speech processing is to take into account expressiveness in speech synthesis systems.

3.3 Expressiveness in textual data

The usage of textual data is dramatically growing: indeed, individuals and organizations communicate and express themselves by using texts, often through Internet both publicly and privately. Textual data may be fully unstructured (a free text) or may be found inside predefined structures (such as web pages, standardized reports, semi-formal models). In the context this research project, textual data may also be considered as transcripts of gesture and speech scenarios. The main research objective is to be able to *identify, characterize, and transfer expressiveness in texts*. In the case of textual data, the definition of expressiveness formulated as: expressiveness is defined as any variation in text that, while keeping the content semantics, conveys other types of interesting and meaningful information such as style, morphology, and so on. This more specific definition, using a more adapted terminology, is consistent with Figure 1 (section 2.1, page 6) where "neutral content" is here meant as the "semantic content" and "expressive content" as "other types of interesting and meaningful information". Expressiveness,

[BÓ1] F. Béchet, "LIA_PHON : un système complet de phonétisation 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 : Proc. 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", Atlanta, Georgia, USA, *in : Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *1*, 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", Poznan, Pologne, *in : Proc. of the 5th Language and Technology Conference (LTC'2011)*, 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*, 2005.

as defined above, is quite important for at least two key aspects:

1. Deriving, inferring and extracting implicit information;
2. Characterizing concrete ways for expressing the same semantic content with variations (style, sentiments, etc.)

We consider that for achieving the main research objective, the text axis needs to be based on text acquisition, text mining, and knowledge generation. Text acquisition is required to enrich texts for better specifying both the content semantics and any additional, possibly hidden or contextual, information. Text mining is required for finding, possibly targeted, information within one or several texts. Finally, knowledge generation is required for structuring content semantics and meaning of other types of information for further usage. This further usage generically refers to design and implement computer-based systems facilitating several activities performed by individuals. For instance, (i) making easier, quicker and reliable any choice based on textual data (ii) making explicit hidden information conveyed by textual data and, as a consequence, (iii) enabling understanding of individuals' behaviours, ideas and so on, (iv) making computer-based systems more efficient and effective on the base of available textual data, and finally (v) supporting individuals in following what textual data implicitly suggest. Additionally, in the context of this research project, further usages can be pointed for improving systems supporting gesture and speech scenarios, as mentioned at the beginning of this section.

Accordingly, *all* the three following topics needs to be studied to implement the definition of expressiveness: text acquisition, text mining, and knowledge generation.

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 processes 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 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 work show its limitations [AP08] and alternative measures of reliability have been proposed [Kri04]. We conduct some experimental studies to investigate the factors of influence that should affect reliability estimation. This challenge deals with the general challenge C1.

Text mining: Due to the explosion of available textual data, text mining and information extraction (IE) from texts have become important topics in recent years. Text mining is particularly adapted 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. Sequential pattern extraction 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

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

[Kri04] K. Krippendorff, "Reliability in Content Analysis: Some Common Misconceptions and Recommendations.", *Human Communication Research 30*, 3, 2004, p. 411–433.

(e.g., [NLHP98,PHW02,PHW07,Bon04]). Constraint-based pattern mining challenges two major problems in pattern 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.This challenge spans over the general challenges C1 and C3.

Knowledge generation: Knowledge generation consists in organizing the information which can be manually or automatically extracted from texts and in representing it a compact way. This representation can vary according to the adopted abstraction level. When studying texts as sequences of words, this representation can be referred to as a language model, while knowledge will rather be represented as ontologies when considering more abstract, higher level, views of texts.

Language models aims at deriving and weighted short linguistic rules from texts, typically using statistical approaches, in order to approximate their shallow structure. These rules are useful to compare texts [SC99] or to help applications in choosing the most likely utterances among a large set of candidates. For instance, language models are used in machine translation [MBC+06], paraphrase generation [QBD04] or ASR [RJ93] to ensure against ungrammatical output texts. The most widely spread language modeling technique is the n-gram approach [Jel76], but major advances have been achieved recently, leading to outperform this venerable approach. Especially, methods based on neural

[NLHP98] R. Ng, L. Lakshmanan, J. Han, A. Pang, "Exploratory mining and pruning optimizations of constrained associations rules", *in : Proc. of SIGMOD'98*, 1998.

[PHW02] J. Pei, J. Han, W. Wang, "Mining Sequential Patterns with Constraints in Large Databases", ACM Press, 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", Press, *in : In Proc. IEEE Int. Conf. on Data Mining ICDM'04*, 2004.

[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", IEEE Computer Society, Washington, DC, USA, *in : Proc. of the 20th International Conference on Data Engineering*, *ICDE '04*, 2004, `http://dl.acm.org/citation.cfm?id=977401.978142`.

[SC99] F. Song, W. B. Croft, "A general language model for information retrieval", *in : Proc. of the eighth international conference on Information and knowledge management*, 1999.

[MBC+06] J. B. Marino, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. Fonollosa, M. R. Costa-Jussà, "N-gram-based machine translation", *Computational Linguistics 32*, 4, 2006, p. 527–549.

[QBD04] C. Quirk, C. Brockett, W. B. Dolan, "Monolingual Machine Translation for Paraphrase Generation.", *in : EMNLP*, 2004.

[RJ93] L. Rabiner, B. Juang, *Fundamentals of Speech Recognition*, Prentice hall Englewood Cliffs, New Jersey, 1993.

[Jel76] F. Jelinek, "Continuous Speech Recognition by Statistical Methods", *Proc. of the IEEE 64*, 4, Apr. 1976, p. 532–556.

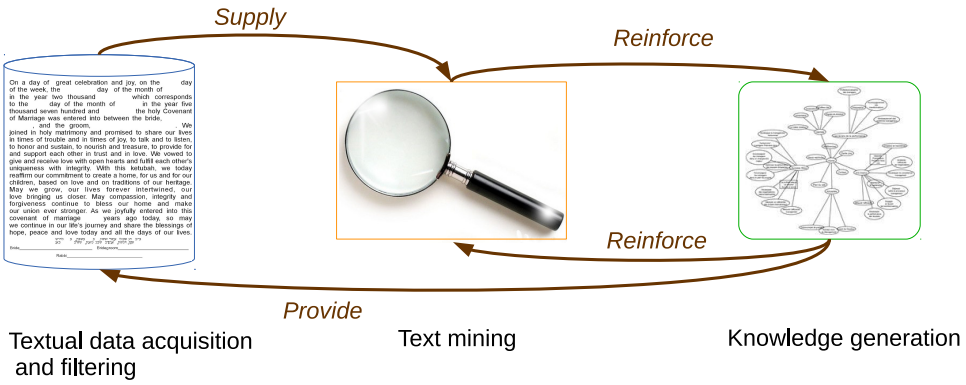

Figure 2: Possible interactions within the text axis.

networks exhibit very good performances [SG02,BDVJ03,MKB+10,MDK+11,Mik12]. As these models are still new, they need to be further studied and extended. In the scope of the team EXPRESSION, these models should be used to model expressive texts.

Second, ontologies are a 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. However, as also recently highlighted [Gan13], despite the work performed, there is still the need to understand much better how to bridge the gap between; on the one side, techniques usable for processing and analyzing texts and, on the other side, information for filling in ontology content (basically concepts, relationships, axioms). Understanding foundations leads to automation improvement and therefore reduction of the required human effort for extracting valuable ontologies.

Hence, a major challenge for the team is to propose solutions to integrate textual expressive information into these knowledge representations. This challenge is part of the general challenge C3.

Interaction between acquisition, mining, and knowledge generation: A key point concerns interactions between the three aforementioned domains. Interactions are required to improve, by

[SG02] H. SCHWENK, J.-L. GAUVAIN, "Connectionist Language Modeling for Large Vocabulary Continuous Speech Recognition", *in : Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *1*, 2002.

[BDVJ03] Y. BENGIO, R. DUCHARME, P. VINCENT, C. JAUVIN, "A Neural Probabilistic Language Model", *Journal of Machine Learning Research 3*, 2, February 2003, p. 1137–1155.

[MKB+10] T. MIKOLOV, M. KARAFIAT, L. BURGET, J. CERNOCKY, S. KHUDANPUR, "Recurrent Neural Network Based Language Model", *in : Proc. of the Conf. of the Intl Speech Communication Association (Interspeech)*, 2010.

[MDK+11] T. MIKOLOV, A. DEORAS, S. KOMBRINK, L. BURGET, J. CERNOCKY, "Empirical Evaluation and Combination of Advanced Language Modeling Techniques", *in : Proc. of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, 2011.

[Mik12] T. MIKOLOV, *Statistical language models based on neural networks*, PdD Thesis, Brno University of Technology, 2012.

[Gan13] A. GANGEMI, "A Comparison of Knowledge Extraction Tools for the Semantic Web", *in : ESWC*, 2013.

reusing methods and techniques proper to each of them, solutions proposed for solving individual challenges. Therefore, interactions should be elicited, planned and coordinated as part of the research activity. Figure 2 shows possible interactions (arrows in the figure) between the text processing domains of interest. For instance, some concrete examples of these interactions are:

- Text mining can be helpful for early steps of knowledge generation by highlighting interesting segments and phenomena to be studied in texts; similarly, generated knowledge can be helpful for constraining text mining;
- Ontology learning (a task of knowledge generation) Knowledge representations can be improved by using mined patterns in order to generate semantic relations, concepts and instances within an ontology or a statistical model;
- Knowledge generation is needed to characterize the content semantics; annotations, resulting from text acquisition, are required to introduce additional information for further characterizing variations according to any knowledge generation process.

Bringing interoperability between proposed methods and developing such interactions is one of the challenges of the text research axis and contributes to the general challenge C2.

4 Application Domains

Many applications domains can be considered for the three modalities. In this section, we only select a few of them:

- Sign Language Translation and Avatar technology; This application domain covers in particular the design of corpora and sign language indexed databases, the development of analysis / synthesis software to control sign language virtual characters [6], and the design of innovative interfaces to manipulate the data. This kind of application may require the recording of high-quality data (body and hand motion, facial expression, gaze direction), or real-time interactive devices to communicate more efficiently and intuitively with the application. Sign language video books can be a targeted application.
- Interactive Multimedia Technology using Gesture; Controlling expressively by gesture the behavior of simulated objects is an emerging research field which can lead to numerous applications: games using gesture as input or virtual assistants as output, virtual theater, or more generally performative art controlled by gesture.
- High quality expressive speech generation is one the major domain with numerous concrete applications like high-quality audiobook generation, online learning, device personalization for disabled people, or video games. In all these cases, expressiveness tends to make users accept TTS outputs by producing less impersonal speech. To be precise, the three following applicative functionalities need to be developed:
  - 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.

Finally, some more text-focused applications domains can be mentioned:

- Under-resourced language analysis will be made possible for instance by developing new tools (like POS Tagger, syntactic parser) for unusual languages (as Latin or Sanskrit), based on sequential pattern extraction.
- Video games, plagiarism detection, recommendation system are instances of applications where extraction and transfer of different expressive forms within textual data (like language registry or state of mind) using patterns or rules will be very useful.
- Opinion mining and sentiment analysis will benefit from new corpora of French emotional norms, i.e. dictionaries which give the polarity of each entry. Building such resources is very expensive and automatic processes have to be tested to extend manually built norms [VB11].
- Human-machine dialogue systems (potentially including TTS) will be improved by integrating expressive models and features, enabling for instance text modulation in order to fit users' profiles.
- Information retrieval and automatic summarization systems are applications where semi-automatically built ontologies will provide a better understanding of texts.

5 New Results

5.1 New Results by Key Issues

In accordance with the Team Project, the main outcomes for 2016 are listed into the following key issues items defined above for the team:

Data acquisition:

1. Hand Gesture: An experimental data set has been built in the scope of Nehla Ghouaiel post-doctorate to develop a benchmark dedicated to upper body movement recognition in stream. We have selected an excerpt of the Naval Air Training and Operating Procedures Standardization (NATOPS) dataset which includes six of the twenty-four body-hand gestures used when handling aircraft on the deck of an aircraft carrier. The dataset includes automatically tracked 3D body postures and hand shapes using a Kinect sensor. The body feature includes 3D joint positions for left/right elbows and wrists, and is represented as a 12D input feature vector. The hand feature includes probability estimates of five predefined hand shapes: opened/closed-palm,thumb-up/down, and "no hand". This data set has been used to develop a challenge in the scope of the AALTD workshop hosted in the ECML-PKDD 2016 conference (https://aaltd16.irisa.fr/challenge/).
2. Facial Expression: Facial data have been captured through a Qualisys mocap system. The mocap data has been recorded at a frequency rate of 200 Hz, with a marker set adapted to sign language facial expressions. The corpus is composed of emotional and clausal sign language expressions, with six basic emotions (joy, sadness, disgust, fear, surprise, anger). Two non professional signing actors (one male and one female) participated to the experience. Each actor performed one sequence of isolated expressions, several sequences of expressions, and small sentences expressing a specific emotion. For each emotion, three levels of intensity have also been considered.

Multi-level representations *No new result this year regarding this key issue.*

[VB11] N. Vincze, Y. Bestgen, "An automatic procedure for extending lexical norms by means of the analysis of word co-occurrences in texts", *TAL 52*, 3, 2011, p. 191–216.

Data Mining and Knowledge Extraction

1. On-line supervised spotting and classification of sub-sequences can be performed by comparing some distance between the stream and previously learned time series. However, learning a few incorrect time series can trigger disproportionately many false alarms. We have developed a fast technique to prune bad instances away and automatically select appropriate distance thresholds. Our main contribution is to turn the ill-defined spotting problem into a collection of single well-defined binary classification problems, by segmenting the stream and by ranking subsets of instances on those segments very quickly. We further demonstrate our technique is effective and robust in the context of an online gesture recognition application.

2. Extracting Descriptors and Classifying Percussive Gestures. Timpani gestures have been characterized by temporal kinematic features (position, velocity, acceleration, Jerk), containing most information responsible for the sound-producing actions. In order to evaluate the feature sets, a classification approach has been conducted under three main attack categories (*Legato*, *Accent* and $Vertical\ Accent$) and sub-categories (dynamics, striking position). Two studies have been carried out to evaluate the performances of six subjects according to their professor's performance: intra-subject and inter-subjects classification [19]. Results are presented in terms of a quantitative ranking of students, using professional gestures as training set, and their gestures as test set.

Generation

1. Generating Expressive Bodily Motions from Expressive End-Effector Trajectories. Recent results in the affective computing sciences point towards the importance of virtual characters capable of conveying affect through their movements. However, in spite of all advances made on the synthesis of expressive motions, almost all of the existing approaches focus on the translation of stylistic content rather than on the generation of new expressive motions. Based on studies that show the importance of end-effector trajectories in the perception and recognition of affect, we have proposed a new approach for the automatic generation of affective motions. In this approach, expressive content is embedded in a low-dimensional manifold built from the observation of end-effector trajectories. These trajectories are taken from an expressive motion capture database. Body motions are then reconstructed by a multi-chain Inverse Kinematics controller. The similarity between the expressive content of MoCap and synthesized motions is quantitatively assessed through information theory measures [18].

2. Facial Expression Synthesis and Annotation. We have created a blendshape-based system for facial animation of $3D$ signing avatars from motion captured data (mocap). The blendshape coefficients are computed along time, by optimizing two energy costs expressing respectively geometrical and deformation measures between mocap markers and 3D positions on the avatar's mesh. Such an approach allows to generate from mocap data the facial animation of the avatar. The same blendshape representation is then used with a hidden Markov model to automatically segment the facial data and annotate this data with the selected emotions [33].

3. Simulation from Motion Descriptors. In 2015, we have achieved a review of computable descriptors of human motion. We have classified them into two categories: low-level descriptors that compute quantities directly from the raw motion data ; and high-level descriptors that use low-level ones to compute boolean, single value or continuous quantities that can be interpreted, automatically or manually, to qualify the meaning, style or expressiveness of a motion. Most of high-level expressive descriptors are computed from Laban Effort components. We have provided formulas inspired from the state of the art that can be applied to 3D motion capture data [26].

In 2016, we have used those descriptors of motion to drive the parameters of a physically based simulation of a non-anthropomorphic entity (a stylized tree with three branches) [27]. Through a perceptual study we have proven that high-level motion descriptors can encode and transfer emotions to a very different entity through an indirect mapping of descriptors to parameters of a simulation.

4. A probabilistic framework for pronunciation generation. Traditional utterance phonetization methods concatenate pronunciations of uncontextualized constituent words. This approach is too weak for some languages, like French, where transitions between words imply pronunciation modifications. Moreover, it makes it difficult to consider global pronunciation strategies, for instance to model a specific speaker or a specific accent. To overcome these problems, we have proposed a new original phonetization approach for French to generate pronunciation variants of utterances [9]. This approach offers a statistical and highly adaptive framework by relying on conditional random fields and weighted finite state transducers. The approach is evaluated on a corpus of isolated words and a corpus of spoken utterances. A communication towards the french speech and language processing community has been done this year [28].

5. Probabilistic pronunciation modeling for spontaneous speech. Pronunciation adaptation consists in predicting pronunciation variants of words and utterances based on their standard pronunciation and a target style. This is a key issue in text-to-speech as those variants bring expressiveness to synthetic speech, especially when considering a spontaneous style. We have proposed a new pronunciation adaptation method which adapts standard pronunciations to the style of individual speakers in a context of spontaneous speech [12]. Its originality and strength are to solely rely on linguistic features and to consider a probabilistic machine learning framework, namely conditional random fields, to produce the adapted pronunciations. Features are first selected in a series of experiments, then combined to produce the final adaptation method. Backend experiments on the Buckeye conversational English speech corpus show that adapted pronunciations significantly better reflect spontaneous speech than standard ones, and that even better could be achieved if considering alternative predictions. Further work has been done this year in this topic leading to [32].

6. Pronunciation adaptation to improve TTS quality. Text-to-speech (TTS) systems are built on speech corpora which are labeled with carefully checked and segmented phonemes. However, phoneme sequences generated by automatic grapheme-to-phoneme converters during synthesis are usually inconsistent with those from the corpus, thus leading to poor quality synthetic speech signals. To solve this problem , our idea is to train corpus-specific phoneme-to-phoneme conditional random fields with a large set of linguistic, phonological, articulatory and acoustic-prosodic features. Features are first selected in cross-validation condition, then combined to produce the final best feature set. Pronunciation models are evaluated in terms of phoneme error rate and through perceptual tests. Experiments carried out on a French speech corpus show an improvement in the quality of speech synthesis when pronunciation models are included in the phonetization process [34]. Appart from improving TTS quality, the presented pronunciation adaptation method also brings interesting perspectives in terms of expressive speech synthesis. Complementary experiments have also been conducted to investigate (i) the choice of optimal features among acoustic, articulatory, phonological and linguistic ones, and (ii) the selection of a minimal data size to train the CRF. Results reported in [35] show that small training corpora can be used without much degrading performance. These results also bring interesting perspectives for more complex adaptation scenarios towards expressive speech synthesis.

7. A phonological penalty to improve TTS quality. Unit selection speech synthesis systems generally rely on target and concatenation costs for selecting the best unit sequence. The role of the

concatenation cost is to insure that joining two voice segments will not cause any acoustic artifact to appear. For this task, acoustic distances (MFCC, F0) are typically used but in many cases, this is not enough to prevent concatenation artifacts. Among other strategies, the improvement of corpus covering by favouring units that naturally support well the joining process (vocalic sandwiches) seems to be effective on TTS. In this work, we investigate if vocalic sandwiches can be used directly in the unit selection engine when the corpus was not created using that principle. First, the sandwich approach is directly transposed in the unit selection engine with a penalty that greatly favours concatenation on sandwich boundaries. Second, a derived fuzzy version is proposed to relax the penalty based on the concatenation cost, with respect to the cost distribution. We show that the sandwich approach, very efficient at the corpus creation step, seems to be inefficient when directly transposed in the unit selection engine. However, we observe that the fuzzy approach enhances synthesis quality, especially on sentences with high concatenation costs. This work has been published in [24, 25] .

8. Adapting prosodic models for expressive speech generation. Rhythmic patterns observed in natural and synthesized speech are compared for three literary forms (rhymes, poems, and fairy tales). The aim of the comparison is to evaluate how rhythm could be improved in synthesized speech, which could allow its adaptation to specific styles or genres. This study is based on the analysis of a corpus of six rhymes, four poems and two extracts from fairy tales. All texts were recorded by three speakers and were generated with two distinct synthesized voices. The comparison of the rhythmic patterns observed is done by analyzing duration in relation to prosodic structure in the various data sets. This approach allows to show that rhythmic differences between synthesized and natural speech are mostly due to the marking of prosodic structure. This work, published in [22, 21], has been done in collaboration with Elisabeth Delais-Roussarie and Hyion Yoo, researchers at LLF.

Use cases and evaluation

1. Speech synthesis system evaluation: Subjective evaluation is a crucial problem in the speech processing community and especially for the speech synthesis field, no matter what system is used. Indeed, when trying to assess the effectiveness of a proposed method, researchers usually conduct subjective evaluations by randomly choosing a small set of samples, from the same domain, taken from a baseline system and the proposed one. When selecting them randomly, statistically, samples with almost no differences are evaluated and the global measure is smoothed which may lead to judge the improvement not significant. To solve this methodological flaw, in [5] and this year in [20], we propose to compare speech synthesis systems on thousands of generated samples from various domains and to focus subjective evaluations on the most relevant ones by computing a normalized alignment cost between sample pairs. This process has been successfully applied both in the HTS statistical framework and in the corpus- based approach. We have conducted two perceptive experiments by generating more than 27000 samples for each system under comparison. A comparison between tests involving most different samples and randomly chosen samples shows clearly that the proposed approach reveals significant differences between the systems.

2. International Blizzard challenge: We participated for the second time to the challenge this year. The process followed to build the voices from given data and the architecture of our system is described in [17] . The search is based on a A* algorithm with preselection filters used to reduce the search space. A penalty is introduced in the concatenation cost to block some concatenations based on their phonological class following the work done in [24] . Moreover, a fuzzy function is used to relax this penalty based on the concatenation quality with respect to the cost distribution.

Despite some problems with the pause prediction module we tried to introduce this year, the results obtained by our system are encouraging, compared to the others results.

5.2 Defended PhDs and HDRs

- Vu Hai Hieu has defended his PhD on the 29th of Januray 2016.
- Jeanne Villaneau has defended her *Habilitation à Diriger des Recherches* (HDR) on the 9th of March 2016.
- David Guennec has defended his PhD on the 22nd of September 2016.
- Pamela Carreno has defended her PhD on the 25th November 2016.

5.3 On going PhDs

1. Marc Dupont has completed his 3rd year of research addressing gesture recognition using a data glove in the scope of controlling a moving robot in an adverse environment. His main achievements are:
   - the design of two "benchmarks" software for the comparison of hand gesture recognition algorithms: the first one is used to test isolated gestures, while the second one is used for gesture spotting and recognition in stream.
   - an improved data base containing isolated gestures and gestures in sequence (data stream).
   - a "sparse" approximation of DTW (dynamic time warping) algorithm, called Coarse-DTW, allowing alignment speed-up via a technique of adaptive sub-sampling of the time series.
   - an original algorithm to rank and prune 'bad' class examples (instances) in an on-line gesture spotting and recognition context.
2. David Guennec has completed his fourth year of research addressing the improvement of cost functions for unit selection speech synthesis. His main achievements concern the use of an A* algorithm and its comparison to the classical Viterbi algorithm and the introduction of a new target cost to constrain the duration of phonemes. For this last point, phoneme durations are predicted using a neural network and used to constrain the selection process. This year, he has worked on the use a phonologic penalty to improve the synthesis and also on prosodic constraints. He has defended his PhD on the 22nd of September 2016 with a jury composed of Nick Campbell (Professor, Trinity College Dublin), Ingmar Steiner (Research, University of Saarland), Phil N. Garner (Senior Researcher, IDIAP), Elisabeth Delais-Roussarie (Director of research, CNRS/LLF) and Yves Laprie (Director of research, CNRS/LORIA).
3. Raheel Qader has completed his third year of research addressing phonology modeling for expressive speech synthesis. During his first year, his main achievements were a review of the state of the art, the analysis of a spontaneous speech English corpus and first experiments towards a pronunciation variant predictive model for speaker and style adaptation. Last year, he has published an article to the SLSP conference on spontaneous pronunciation prediction. This year, we has worked on the subjective evaluation of pronunciation modifications and also on disfluencies in spontaneous speech. His defence is planned to occur in March 2017.
4. Pamela Carreno has defended her PhD thesis [13], after three years of research on the analysis and synthesis of expressive theatrical movements. She proposed a low-dimensional motion representation, consisting of the spatio-temporal trajectories of end-effectors (i.e., head, hands and

feet), and pelvis. Throughout her work, she showed that this representation is both suitable and sufficient for characterizing the underlying expressive content in human motion, and for controlling the generation of expressive whole-body movements. Several results highlight the validity of these hypothesis: (i) A new motion capture database inspired by physical theory, which contains three categories of motion (locomotion, theatrical and improvised movements), has been built for several actors; (ii) An automatic classification framework has been designed to qualitatively and quantitatively assess the amount of emotion contained in the data. It has been shown that the proposed low-dimensional representation preserves most of the motion cues salient to the expression of affect and emotions; (iii) A motion generation system has been implemented, both for reconstructing whole-body movements from the low-dimensional representation, and for producing novel end-effector expressive trajectories. A quantitative and qualitative evaluation of the generated whole body motions shows that these motions are as expressive as the movements recorded from human actors.

5. Lei Chen has completed her fourth year of research addressing the analysis of musical gestures for sound control. After studying timpani gestures through an automatic classification approach [19], she has designed a new gesture set inspired from conducting gestures performed by an orchestral conductor. She specifically focuses on expressive gestures, generally performed with the non dominant hand, which show some aspects of the music, such as sound texture and quality, atmosphere and expression, or variations in dynamics and intensity. She also assumes that these conducting gestures present highly-coded structural patterns also found in sign languages of the Deaf. The corpus has been partially recorded by two subjects, and a clustering approach is developed to extract significant behaviors for different categories of actions. Her defense is planned to occur in 2017.

6. Clément Reverdy has completed his second year of PhD. His research addresses the problem of facial expression analysis and synthesis in the context of sign language. He has developed a set of blendshape-based methods using motion captured data. A new synthesis system has been designed and implemented. This system can be described in two steps. First, it achieves the retargeting of the data in order to adapt the morphological differences between the positions of the mocap markers (located on the actor) and the corresponding positions on the avatar's mesh. Second, the blendshape coefficents are computed though an optimization method which uses two costs characterizing some energy of regularization. These costs act differently on the produced animation. The first cost – a geometrical cost – limits the space in which the blendshape vector can evolve. The second one – a physical cost – minimizes the mesh deformation. The blendshape representation is also used in a hidden Markov model to automatically annotate the facial data according to the expressed emotion [33].

7. Stefania Pecore has completed her first year of PhD. The topic of her research is sentiment analysis and, more precisely, detection of opinion from review extracted from French websites. Some experiments using classical statistical tools (SVM and Logistic Regression) have suggested directions to follow in order to address the shortcomings of the bag-of-words approach [PVS16]. Currently, the focus of the research are the study of the contribution of negation in opinion mining and the extraction of words and patterns from manually annotated data to enrich a french opinion lexicon.

8. Lucie Naert has started her PhD in October 2016. The PhD thesis aims at designing and animating signing avatars, i.e. virtual 3D characters signing in signed languages. This is part of

[PVS16] S. Pecore, J. Villaneau, F. Saïd, "Combiner lexique et régression logistique dans la classification d'avis laissés sur le Net : une étude de cas", *in : TALN*, 2016, `https://hal.archives-ouvertes.fr/hal-01447571/file/coltal2016.pdf`.

a larger project dedicated to the editing and generation of digital contents useful for Deaf and hearing people signing in French Sign Language (LSF). Her thesis follows an editing-generation / perception scheme. From the editing of simple sentences constructed on a limited LSF corpus, the synthesis system generates LSF movements, and simultaneously controls the animation of bodily movements, hand gestures, facial expressions, and gaze direction. With such an approach, novel sentences are built from the re-composition of movements extracted in annotated databases or synthesized according to the linguistic context. The objective of the thesis is to find appropriate representations allowing on the one hand to construct intelligible sentences, following grammatical and semantic rules in LSF, and on the other hand to access, retrieve, or synthesize in real time the movements chunks to finally control the animation of the virtual character, while preserving the co-articulation of the produced movements. The work also addresses some issues related to perceptual evaluation. Different types of experiments will be conducted to evaluate the quality of the signing utterances, considering both the comprehension and the acceptance of the 3-dimensional avatar by the Deaf.

9. Cédric Fayet has completed his first year. The topic of his research is the detection of abnormality from facial movements and speech signals of a human being. By "abnormality" we mean the existence of foreign elements to a normal situation in a given context. The study focuses in particular on the joint use of facial and vocal expression parameters to detect abnormal variations of expressivity in speech, not only related to emotion, but also to social interactions and psychological signals. For instance, these abnormal signals can appear in extreme stress situations for pilots or vehicle drivers. This study could also find applications in the medical field, e.g., detection of abnormal behaviors due to mental disabilities such as autism. We aimed at developing a system capable of detecting abnormal behaviors by the analysis of records of concrete situations. This year, he has focused on the study of literature and he searched the available corpora. He also begun to work on the detection of non professional speakers on radio using the acoustic signal. First results are promising and should be submitted for publication in 2017.

10. Sandy Aoun has started her PhD in December. Her research addresses the optimisation of recording scripts for the expressive reading of audiobooks. The originality of this work is that the problem is addressed by trying to find the best subset of the books we want to synthesize, that will be used to build a voice, then used to generate the remaining part of the books. This way, the goal is to find the best compromise between the size of what we need to record and the quality of the audiobooks we generate.

11. Aghilas Sini has started her PhD in December. His research addresses the characterisation and generation of expressivity in function of speaking styles for audiobook synthesis. This Phd takes places in the context of the ANR project SynPaFlex dealing with prosody modelling and the use of prosodic models in speech synthesis. This thesis is also co-funded by the Labex Empirical Foundations of Linguistics (EFL) and co-directed by Elisabeth Delais-Roussarie (DR, CNRS/LLF).

6 Software

6.1 SGN

Participants: Caroline Larboulette, Sylvie Gibet.

2016: The SGN library has been enriched with a library to compute motion descriptors (low-level and high-level motion descriptors).

Several QT-based small tools have been developed to study motion transitions in the context of sign language.

2015: The SGN library has been enriched with $3D$ rendering of the avatar: a skinning algorithm and a blendshape algorithm.

In addition, a dynamics module has been added for mass-spring and particle simulations to use SGN in a more general context (physically based simulation from captured data, not only avatars).

A QT interface has been built for a signing avatar demonstration based on SGN and interactive contruction of sentences from a sequence of gloses chosen by the user. The demo has first been presented at the public open house of IRISA and later refined and used for the HCERES evaluation of the laboratory.

6.2 ROOTS

Participants: Nelly Barbot, Vincent Barreaud, Jonathan Chevelu, Arnaud Delhay, Sébastien Le Maguer [University of Saarland], 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 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 3, 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[2].

[2]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", Florence, Italie, *in : Conference of the International Speech Communication Association (Interspeech)*, 2011.

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

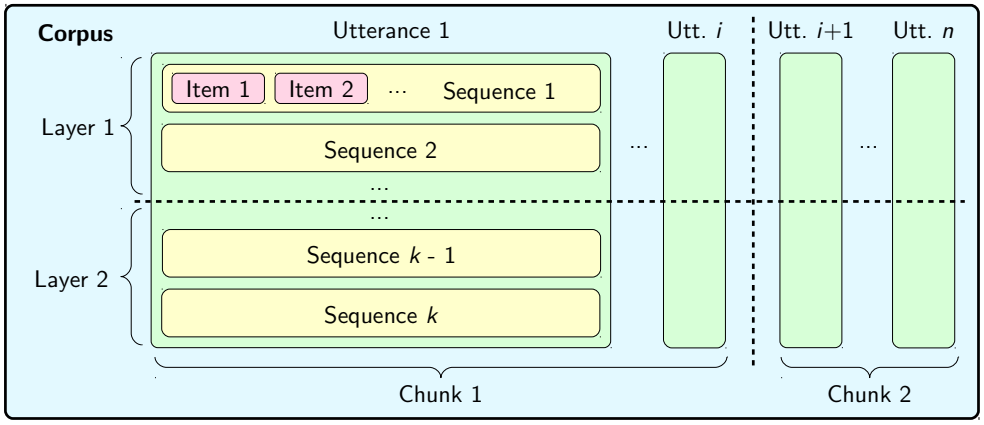

Figure 3: Hierarchical organization of data in ROOTS.

- Sequence: add/remove/get/update items; dump,
- 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

[CB02] D. CUNNINGHAM, H.AND MAYNARD, V. BONTCHEVA, K.AND TABLAN, "GATE: an architecture for development of robust HLT applications", *in: Proc. of the Annual Meeting of the ACL*, 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", *research report*, 2006.

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 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 [CLL14].

6.3 Web-based listening test system

Participants: Vincent Barreaud, Arnaud Delhay, Clause Simon [from IUT Lannion], 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.

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

[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.

[CLL14] J. CHEVELU, G. LECORVÉ, D. LOLIVE, "ROOTS: a toolkit for easy, fast and consistent processing of large sequential annotated data collections", Reykjavik, Iceland, *in : Language Resources and Evaluation Conference (LREC)*, May 2014, `https://hal.inria.fr/hal-00974628`.

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

6.4 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.

6.5 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 artifacts 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.

6.6 Recording Studio

Participants: 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.

6.6.1 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).

6.6.2 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.

7 Contracts and Grants with Industry

7.1 INGREDIBLE

Participants: Pamela Carreño, Sylvie Gibet, 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 outcomes of the project are:

- The creation of different motion corpora (fitness, interactive video games motions, theatrical gestures);
- The development of algorithms for affect recognition from expressive features;
- The development of a behavioral system linked to the recognition and synthesis modules;
- The development of synthesis algorithms for animating a virtual character in interactive situations.

7.2 SynPaFlex

Participants: Damien Lolive, Gwénolé Lecorvé, Marie Tahon, Gaëlle Vidal, Aghilas Sini.

Expression is leader of a ANR project named SynPaFlex and accepted in July 2015 and started the 1st of December 2015. This project is targeted at the improvement of Text-To-Speech synthesis engines through two main research axes:

- Pronunciation variants modelling and generation
- Context-adapted prosody modelling and generation

The main targeted applications are in the domains of entertainment (audiobook reading, video games), serious games (virtual environments), language learning (dictation, elocution style) or even for vocal aids designed for handicapped people. This project is mainly supported by IRISA, coordinated by Damien Lolive and involves members from LLF (Laboratoire de Linguistique Formelle) and from ATILF.

Up-to-date information are available at `https://synpaflex.irisa.fr`.

7.3 TREMoLo

Participants: Gwénolé Lecorvé, Nicolas Béchet, Jonathan Chevelu.

Expression is leader of a ANR project named TREMoLo and accepted in December 2016. Official scientific start is scheduled on the 1st of October 2017, and preliminary activities will begin in the meanwhile. The project studies the use of language registers and seeks to develop automatic methods towards the transformation of texts from a register to another. To do so, the project proposes to extract linguistic patterns which discriminate a register from another, and to integrate them into a probabilistic automatic paraphrase generation process. The language under study is French.

This project is mainly supported by IRISA, coordinated by Gwénolé Lecorvé and involves a member of MoDyCo (UMR 7114 Modèles, Dynamiques, Corpus).

7.4 VOCAGEN

Participants: Nicolas Béchet, Giuseppe Berio.

Expression is involved in a Brittany Region project named VOCAGEN, supported by the Image et Réseaux competitiveness cluster. It was accepted in September 2016 and it officially started on the 1st of December 2016 with the recruitment of a post-doctoral researcher Rémy Kessler. The project studies the automatic form input via the implementation of a voice recognition system. The implication of the Expression team is about the semi-automatic extraction of a knowledge model using data mining techniques and machine learning in order to improve the voice recognition system quality.

This project is mainly supported by the ScriptAndGo company, coordinated by Sébastien Macé and involves members of the TyKomz company.

8 Other Grants and Activities

8.1 International Collaborations

8.2 National Collaborations

- **Hybride ANR Project** Participant: Nicolas Béchet. The Hybride Research Project aims at developing new methods and tools for supporting knowledge discovery from textual data by combining methods from Natural Language Processing (NLP) and Knowledge Discovery in Databases (KDD). The consortium is made of INRIA/LORIA, GREYC, MoDyCo, Inserm (IRISA is associated to the GREYC in this project). The coordinator is Yannick Toussaint from INRIA/LORIA.

- **ANIMITEX CNRS MASTODONS Project** Participant: Nicolas Béchet. The ANIMITEX Project aims is to exploit the massive and heterogeneous textual data to provide crucial information in order to complete the analysis of satellite images. The consortium is made of LIRMM, TETIS, ICUBE, GREYC, LIUPPA, IRISA. The coordinator is Mathieu Roche, Cirad/TETIS.

9 Dissemination

9.1 Involvement in the Scientific Community

- Pierre-François Marteau served as a reviewer in international journals (IEEE TPAMI, IEEE TNNLS, IEEE TKDE, PRL). 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, the IEEE – TNNLS (Transactions on Neural Networks and Learning Systems). She has served as a reviewer for the international conference on Motion Computing (MOCO 2016). She has been a reviewer for the HDR thesis of Jean-Loïc Le Carrou (UPMC, October 2016). She is also an elected member of the administrative council of Université Bretagne Sud.

- Nelly Barbot has served as a reviewer for the International conference of the International Speech Communication Association (Interspeech).

- Jonathan Chevelu has served as a reviewer for the International conference of the International Speech Communication Association (Interspeech), the International Conference on Audio, Speech and Signal Processing (ICASSP).

- Arnaud Delhay has been re-elected as a member of the 'Commission Recherche' (Research comittee) of the IUT of Lannion in November 2015. He has served as a reviewer for the International conference of the International Speech Communication Association (Interspeech), the International Conference on Audio, Speech and Signal Processing (ICASSP).

- Caroline Larboulette is a member of various program committees for international conferences (CAe/EXPRESSIVE 2016, CASA 2016, MIG 2016, MOCO 2016 and SIGGRAPH Unified Jury 2016), a member of the editorial review board of the international journal of computer graphics and creative interfaces (IJCICG) and serves as a reviewer for various journals (Computer & Graphics, TVCG, CAVW). She is a member of the ACM SIGGRAPH Specialized Conferences Committee that attributes the ACM SIGGRAPH labels to conferences and supervises the budget of conferences sponsored by ACM SIGGRAPH. She is also vice-president of the ACM SIGGRAPH Madrid Professional Chapter that she created in 2007.

- Gwénolé Lecorvé is an elected member of the laboratory council of IRISA, and of the board of directors of the French speech communication association (AFCP). He also serves as a reviewer conferences and journals (Interspeech, ACM Multimedia *Traitement Automatique des Langues* journal, *Journées d'Études sur la Parole* conference). He served as an expert for the French research agency (ANR). He chaired the speech synthesis session of *Journées d'Études sur la Parole*. He was examiner (*examinateur*) in the PhD thesis of Hai Hieu Vu.

- Damien Lolive is an elected member of the 'Conseil Scientifique' (Scientific council) of ENSSAT, Lannion, and of board of directors of the French speech communication association (AFCP). He serves as a reviewer for the IEEE Transactions on Speech and Language processing, for the *Traitement Automatique des Langues* journal, for the International conference of the International Speech Communication Association (Interspeech), the International Conference on Audio, Speech and Signal Processing (ICASSP) and for the *Journées d'Études sur la Parole* conference. He has also served has an expert for the french research agency, ANR. He has the benefit of a half-time CNRS delegation for 2016-2017.

9.2 Teaching

- Nelly Barbot teaches the following mathematics courses at École Nationale Supérieure des Sciences Appliquées et de Technologie (ENSSAT): algebra and analysis basis, mathematical logic in Licence level, probability and statistics in Master level. In 2015 and 2016, she was responsible of the team of teachers of Mathematics and Human Sciences at ENSSAT and responsible of the teaching modules of economics and management.

- Nicolas Béchet teaches various computer sciences courses at the Statistique et Informatique D'ecisionnelle department of IUT Vannes.

- Arnaud Delhay teaches databases and web programming (server- and client-side) in Licence levels at IUT of Lannion, calculability and computational complexity of problems in Master level and web server-side programming in Licence level at École Nationale Supérieure des Sciences Appliquées et de Technologie (ENSSAT).

- Sylvie Gibet teaches the following computer science courses at the faculty of sciences, Université Bretagne Sud: functional programming and algorithmic in License level, and signal processing 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.

- Caroline Larboulette teaches logic and functional programming for undergraduates of the UFR SSI; C++ programming for ENSIBS graduate students; and computer animation, rendering and interactive techniques at the master level (Master WMR).

- Gwénolé Lecorvé teaches the following computer science courses at École Nationale Supérieure des Sciences Appliquées et de Technologie (ENSSAT): Java; distributed algorithmics; artificial intelligence; and machine learning in Master level.

- Damien Lolive teaches the following computer science courses at École Nationale Supérieure des Sciences Appliquées et de Technologie (ENSSAT): object-oriented programming in Licence level, compilers architecture and formal languages theory in Master level, and speech and language processing in Master level.

- Pierre-François Marteau teaches programming languages, logics, introduction to cryptography and information retrieval courses in computer sciences License and Master levels, mostly at École Nationale Supérieure de Bretagne Sud.

- Gildas Ménier teaches various computer sciences courses at the faculty of sciences, Université Bretagne Sud.

9.3 Conferences, workshops and meetings, invitations

- Pierre-François Marteau has co-organized with Ahlame Douzal (LIG, Grenoble Alpes University, France), and José A. Vilar (University A Coruna, Spain) the first ECML/PKDD 2015 Workshop on: Advanced Analytics and Learning on Temporal Data (AALTD'2015) Friday, September 11, 2015, Porto, Portugal. 20 papers have been presented orally or as posters, and a post-act volume in Springer LNAI series is in preparation.

9.4 Graduate Student and Student intern

- Sandy Aoun has done is Master level internship in Lannion and Beirut (University of Lebanon). She has worked on the optimization of corpus contents for text-to-speech synthesis. She is now pursuing this work as a PhD student.

10 Bibliography

Major publications by the team in recent years

[1] 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 : Proceedings of Interspeech*, p. 1501–1504, 2011.

[2] N. Barbot, O. Boëffard, J. Chevelu, A. Delhay, "Large Linguistic Corpus Reduction with SCP algorithms", *Computational Linguistics 41*, 3, 2015, p. 30.

[3] O. Boëffard, L. Charonnat, S. Le Maguer, D. Lolive, "Towards Fully Automatic Annotation of Audio Books for TTS", *in : Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, 2012.

[4] 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`.

[5] J. Chevelu, D. Lolive, S. Le Maguer, D. Guennec, "How to Compare TTS Systems: A New Subjective Evaluation Methodology Focused on Differences", *in : Interspeech*, Dresden, Germany, September 2015, `https://hal.inria.fr/hal-01199082`.

[6] S. Gibet, N. Courty, K. Duarte, T. Le Naour, "The SignCom System for Data-driven Animation of Interactive Virtual Signers : Methodology and Evaluation".

[7] 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.

[8] G. Ke, P.-F. Marteau, G. Ménier, "Improving the clustering or categorization of bi-lingual data by means of comparability mapping", October 2013.

[9] G. Lecorvé, D. Lolive, "Adaptive Statistical Utterance Phonetization for French", *in : Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5 p., 2 columns, Brisbane, Australia, April 2015, `https://hal.inria.fr/hal-01109757`.

[10] D. Lolive, N. Barbot, O. Boëffard, "B-spline model order selection with optimal MDL criterion applied to speech fundamental frequency stylisation", *IEEE Journal of Selected Topics in Signal Processing 4*, 3, 2010, p. 571–581.

[11] P.-F. Marteau, S. Gibet, "On Recursive Edit Distance Kernels with Application to Time Series Classification", February 2013.

[12] R. Qader, G. Lecorvé, D. Lolive, P. Sébillot, "Probabilistic Speaker Pronunciation Adaptation for Spontaneous Speech Synthesis Using Linguistic Features", *in : International Conference on Statistical Language and Speech Processing (SLSP)*, p. 12 p., 1 column, Budapest, Hungary, November 2015, `https://hal.inria.fr/hal-01181192`.

Doctoral dissertations and "Habilitation" theses

[13] P. Carreno-Medrano, *Analysis and Synthesis of Expressive Theatrical Movements*, Theses, Université Bretagne Sud, November 2016, https://hal.archives-ouvertes.fr/tel-01490785.

[14] D. Guennec, *Study of unit selection text-to-speech synthesis algorithms*, Theses, Université Rennes 1, September 2016, https://tel.archives-ouvertes.fr/tel-01439413.

[15] J. Villaneau, *Contributions à la modélisation du sens par approches formelles, linguistiques et statistiques*, Habilitation à diriger des recherches, Université Bretagne Loire, March 2016, https://hal.archives-ouvertes.fr/tel-01448487.

Articles in referred journals and book chapters

[16] P.-F. Marteau, S. Gibet, C. Reverdy, "Adaptive Down-Sampling and Dimension Reduction in Time Elastic Kernel Machines for Efficient Recognition of Isolated Gestures", *in : Advances in Knowledge Discovery and Management: volume 6*, F. Guillet, B. Pinaud, and G. Venturini (editors), *Studies in Computational Intelligence*, *Volume*, 665, Springer International Publishing, 2016, p. 39 – 59, https://hal.archives-ouvertes.fr/hal-01401453.

Publications in Conferences and Workshops

[17] P. Alain, J. Chevelu, D. Guennec, G. Lecorvé, D. Lolive, "The IRISA Text-To-Speech System for the Blizzard Challenge 2016", *in : Blizzard Challenge 2016 workshop*, Cupertino, United States, September 2016, https://hal.inria.fr/hal-01375897.

[18] P. Carreno-Medrano, S. Gibet, P.-F. Marteau, "From Expressive End-Effector Trajectories to Expressive Bodily Motions *", *in : 29th International Conference on Computer Animation and Social Agents*, Genève, Switzerland, May 2016, https://hal.archives-ouvertes.fr/hal-01367822.

[19] L. Chen, S. Gibet, P.-F. Marteau, F. Marandola, M. M. Wanderley, "Quantitative Evaluation of Percussive Gestures by Ranking Trainees versus Teacher", *in : MOCO'16: 3rd International Symposium On Movement & Computing*, Thessaloniki, Greece, July 2016, https://hal.archives-ouvertes.fr/hal-01367819.

[20] J. Chevelu, D. Lolive, S. Le Maguer, D. Guennec, "Se concentrer sur les différences : une méthode d'évaluation subjective efficace pour la comparaison de systèmes de synthèse", *in : Journées d'Études sur la Parole*, Paris, France, July 2016, https://hal.inria.fr/hal-01338918.

[21] E. Delais-Roussarie, D. Lolive, H. Yoo, D. Guennec, "Patrons Rythmiques et Genres Littéraires en Synthèse de la Parole", *in : Journées d'Études sur la Parole*, Paris, France, July 2016, https://hal.inria.fr/hal-01338959.

[22] E. Delais-Roussarie, D. Lolive, H. Yoo, D. Guennec, "Rhythmic Patterns and Literary Genres in Synthesized Speech", *in : Speech Prosody*, Boston, United States, 2016, https://hal.inria.fr/hal-01338873.

[23] M. Dupont, P.-F. Marteau, N. Ghouaiel, "Detecting Low-Quality Reference Time Series in Stream Recognition", *in : International Conference on Pattern Recognition (ICPR)*, IAPR (editor), IAPR, IEEE, Cancun, Mexico, December 2016, https://hal.archives-ouvertes.fr/hal-01435197.

[24] D. Guennec, D. Lolive, "On the suitability of vocalic sandwiches in a corpus-based TTS engine", *in : Interspeech*, San Francisco, United States, September 2016, https://hal.inria.fr/hal-01338839.

[25] D. Guennec, D. Lolive, "Une pénalité floue fondée phonologiquement pour améliorer la Sélection d'Unité", *in : Journées d'Études sur la Parole*, Paris, France, July 2016, https://hal.inria.fr/hal-01338948.

[26] C. Larboulette, S. Gibet, "A Review of Computable Expressive Descriptors of Human Motion", *in : MOCO 2015 : 2nd International Workshop on Movement and Computing, Proceedings of the International Workshop on Movement and Computing*, p. 21–28, Vancouver, Canada, August 2015, https://hal.archives-ouvertes.fr/hal-01196267.

[27] C. Larboulette, S. Gibet, "I Am a Tree: Embodiment Using Physically Based Animation Driven by Expressive Descriptors of Motion", *in : 3rd International Symposium on Movement and Computing, Proceedings of the International Symposium on Movement and Computing*, ACM, Thessaloniki, Greece, July 2016, https://hal.archives-ouvertes.fr/hal-01344746.

[28] G. Lecorvé, D. Lolive, "Phonétisation statistique adaptable d'énoncés pour le français", *in : Journées d'Études sur la Parole*, Paris, France, July 2016, https://hal.inria.fr/hal-01321358.

[29] P.-F. Marteau, "Assessing pattern recognition or labeling in streams of temporal data", *in : 2nd ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data*, Riva del Garda, Italy, September 2016, https://hal.archives-ouvertes.fr/hal-01403948.

[30] B. Möbius, S. Le Maguer, I. Steiner, D. Lolive, "De l'utilisation de descripteurs issus de la linguistique computationnelle dans le cadre de la synthèse par HMM", *in : Journées d'Études sur la Parole*, Paris, France, July 2016, https://hal.inria.fr/hal-01338953.

[31] S. Pecore, J. Villaneau, F. Saïd, "Combiner lexique et régression logistique dans la classification d'avis laissés sur le Net : une étude de cas", *in : TALN 2016*, Paris, France, July 2016, https://hal.archives-ouvertes.fr/hal-01447571.

[32] R. Qader, G. Lecorvé, D. Lolive, P. Sébillot, "Adaptation de la prononciation pour la synthèse de la parole spontanée en utilisant des informations linguistiques", *in : Journées d'Études sur la Parole*, Paris, France, July 2016, https://hal.inria.fr/hal-01321361.

[33] C. Reverdy, S. Gibet, C. Larboulette, P.-F. Marteau, "Un système de synthèse et d'annotation automatique à partir de données capturées pour l'animation faciale expressive en LSF", *in : Journées Françaises d'Informatique Graphique (AFIG 2016)*, Grenoble, France, November 2016, https://hal.archives-ouvertes.fr/hal-01490780.

[34] M. Tahon, R. Qader, G. Lecorvé, D. Lolive, "Improving TTS with corpus-specific pronunciation adaptation", *in : Interspeech*, San Francisco, United States, September 2016, https://hal.inria.fr/hal-01338111.

[35] M. Tahon, R. Qader, G. Lecorvé, D. Lolive, "Optimal feature set and minimal training size for pronunciation adaptation in TTS", *in : International Conference on Statistical Language and Speech Processing (SLSP)*, Pilsen, Czech Republic, October 2016, https://hal.inria.fr/hal-01338853.

Miscellaneous

[36] P.-F. Marteau, "Times series averaging and denoising from a probabilistic perspective on time-elastic kernels", working paper or preprint, November 2016, https://hal.archives-ouvertes.fr/hal-01401072.