Project-Team EXPRESSION

Expressiveness in Human Centered Data/Media

Vannes, Lannion & Lorient

Activity Report

2017

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
Joëlle Thépault, 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 (emeritus), 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

← Saeid Soheily-Khah, Post-doctoral researcher, Université de Bretagne Sud (until October 2017)
← Pamela Carreno, Université de Bretagne Sud, PhD, ATER (until August 2017)
Rémi Kessler, Post-doctoral researcher, Université de Bretagne Sud (since December 2016)
← Marie Tahon, Post-doctoral researcher, Université de Rennes 1 (until August 2017)
Gaëlle Vidal, Engineer, Université de Rennes 1 (from September 2017)

PhD students

⇌ Sandy Aoun, Université de Rennes 1, ARED/CD22, 1st year (resigned on February 2017)
→ Yonatan Carranza Alarcón, Université de Bretagne Sud, 1st year
Lei Chen, Université de Bretagne Sud/Univ. McGill, ARED, 3rd year
← Marc Dupont, Université de Bretagne Sud, Thèse CIFRE Thales, final year (defended on March 2017)
Cédric Fayet, Université de Rennes 1, DGA/ARED, 3rd year
Lucie Naert, Université Bretagne Sud, CDE, 2nd year
→ Nicolas Bloyet, Université de Bretagne Sud, Thèse CIFRE Seed, 1st year
Stefania Pecóre, Université de Bretagne Sud, 2nd year

→ Antoine Perquin, Université de Rennes 1, CD INSA, 1st year
⇌ Raheel Qader, Université de Rennes 1, MESR, final year (defended on March 2017)
Clément Reverdy, Université Bretagne Sud, CDE+ANR InGredible, final year
→ Meysam Shamsi, Université de Rennes 1, ARED/CD22, 1st year
→ Aghilas Sini, Université de Rennes 1, LABEX EFL/ANR SynPaFlex, 1st year
⇌ Sabiha Tahrat, Université de Rennes 1, ANR TREMoLo, 1st year (resigned on November 2017)

Master students
⇌ Antoine Perquin, Université de Rennes 1, ENSSAT

1.2 Evolution of the staff

The permanent staff has been stable. The team has developed its PhD supervision capacity at Lannion with the defence of Damien Lolive's HDR in November. The number of PhD students is also stable with 2 PhD defenses and three new hired PhD students: Antoine Perquin, Aghilas Sini and Meysam Shamsi in Lannion. The year has also been marked by the resignations of 2 PhD students (Sandy Aoun and Sabiha Tahrat) during the early stages of their contract.

The PhD defense of Marc Dupont was held on 28th March 2017. Marc Dupont holds since March 2017 a research & development engineer position at CISCO company, in Paris. The PhD defense of Raheel Qader was hold on 31st March 2017. Raheel now works as a postdoctoral researcher at *Laboratoire d'Informatique de Grenoble* (LIG). Finally, Marie Tahon left the team as she was hired as an associate professor in computer science at Le Mans Université as part of the LIUM lab.

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.

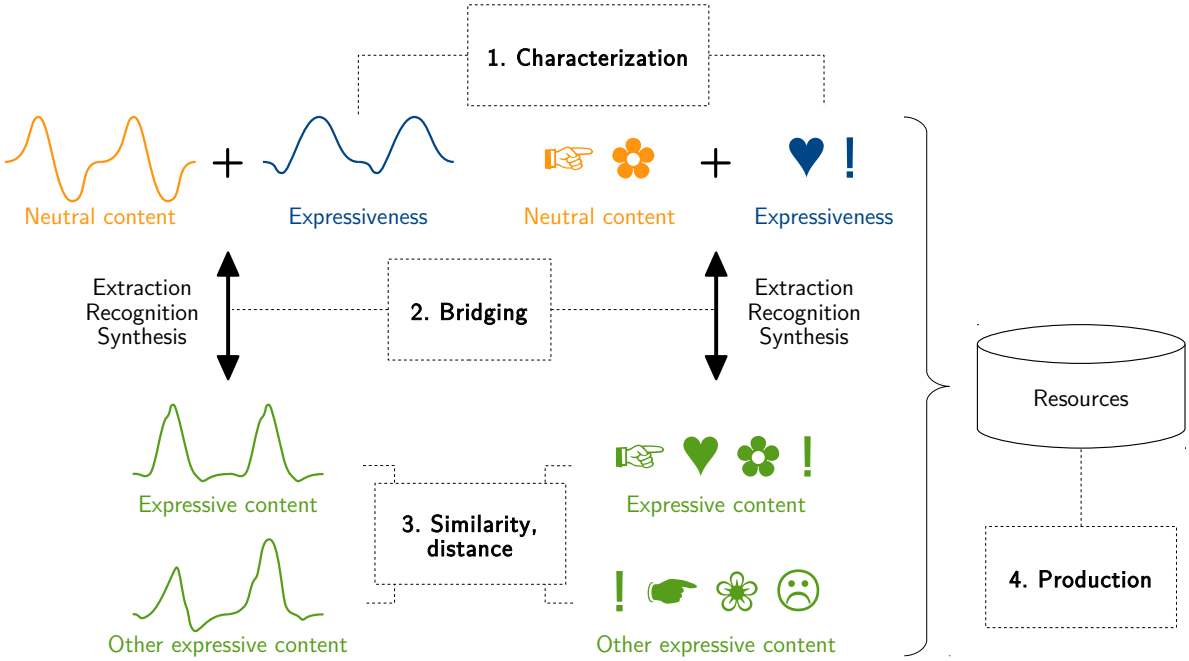

Figure 1: Overview of the main challenges considering both on continuous numerical (left) and discrete symbolic (right) data.

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.*, [6]). 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.

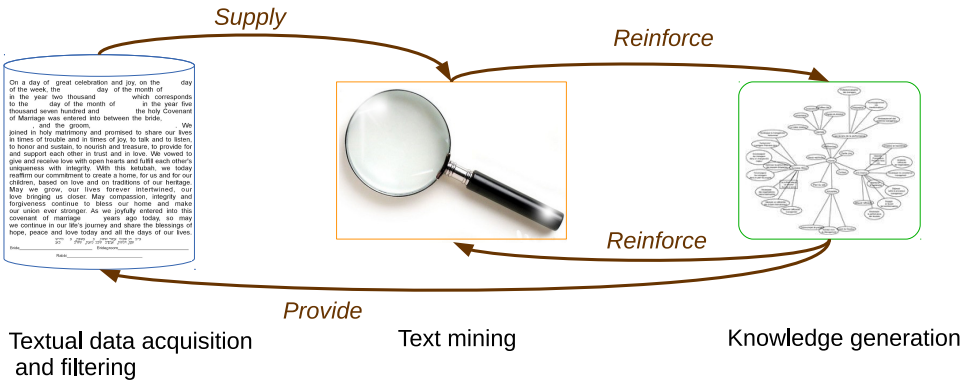

Figure 2: Possible interactions within the text axis.

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 [5], 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 Main events

5.2 New Results by Key Issues

In accordance with the Team Project, the main outcomes for 2017 are listed into the following key issues items defined above for the team:

Data acquisition, generation

Facial Animation from mocap Data: A new mocap data set with basic emotions (joy, sadness, disgust, fear, surprise, anger) has been created. It contains isolated expressions, as well as coarticulated expressions in synthetic transitions or sign language utterances. Using mocap data, facial expressions are represented through blendshape coefficients which are used both to analyze data and to animate $3D$ avatars. A complete synthesis system has been developed from mocap data. It includes a morphological adaptation process (retargetting), and a combination of two optimization processes. The final rendering process has also been improved.

Knowledge extraction

Automatic Annotation of Sign Language Motion Capture Data [18, 17]: A segmentation of the data is carried out on derivative features computed from the blenshape coefficients computed from facial mocap data. Based on this segmentation, an automatic machine learning process recognizes the basic emotions (joy, sadness, disgust, fear, surprise, anger) and annotates the data. A similar approach is a achieved on hand data. The main hand distances and their derivatives are first computed to segment the data. A machine learning process is then applied to classify and annotate the hand data. This approach is applied successfully on facial expressions and hand configurations in French sign language.

Knowledge extraction

[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.

Anomaly detection using personality traits vs. prosodic features [14, 13, 15]: This work presents the design of an anomaly detector based on three different sets of features, one corresponding to some prosodic descriptors and two extracted from Big Five traits. Big Five traits correspond to a simple but efficient representation of a human personality. They are extracted from a manual annotation while prosodic features are extracted directly from the speech signal. We evaluate two different anomaly detection methods: One-Class SVM (OC-SVM) and iForest, each one combined with a threshold classification to decide the "normality" of a sample. The different combinations of models and feature sets are evaluated on the SSPNET-Personality corpus which has already been used in several experiments, including a previous work on separating two types of personality profiles in a supervised way. In this work, we propose the above mentioned unsupervised methods, and discuss their performance, to detect particular audio-clips produced by a speaker with an abnormal personality. Results show that using automatically extracted prosodic features competes with the Big Five traits. In our case, OCSVM seems to get better results than iForest.

Data acquisition, Knowledge extraction

Study on the perception of expressivity in TTS [21]: Actually a lot of work on expressive speech focus on acoustic models and prosody variations. However, in expressive Text-to-Speech (TTS) systems, prosody generation strongly relies on the sequence of phonemes to be expressed and also to the words below these phonemes. Consequently, linguistic and phonetic cues play a significant role in the perception of expressivity. In previous works, we proposed a statistical corpus-specific framework which adapts phonemes derived from an automatic phonetizer to the phonemes as labelled in the TTS speech corpus. This framework allows to synthesize good quality but neutral speech samples. The present study goes further in the generation of expressive speech by predicting not only corpus-specific but also expressive pronunciation. It also investigates the shared impacts of linguistics, phonetics and prosody, these impacts being evaluated through different French neutral and expressive speech collected with different speaking styles and linguistic content and expressed under diverse emotional states. Perception tests show that expressivity is more easily perceived when linguistics, phonetics and prosody are consistent. Linguistics seems to be the strongest cue in the perception of expressivity, but phonetics greatly improves expressiveness when combined with and adequate prosody.

Generation

Probabilistic pronunciation modeling for spontaneous speech generation [20]: To bring more expressiveness into text-to-speech systems, this work presents a new pronunciation variant generation method which works by adapting standard, i.e., dictionary-based, pronunciations to a spontaneous style. Its strength and originality lie in exploiting a wide range of linguistic, articulatory and prosodic features, and in using a probabilistic machine learning framework, namely conditional random fields and phoneme-based n-gram models. Extensive experiments on the Buckeye corpus of English conversational speech demonstrate the effectiveness of the approach through objective and perceptual evaluations.

Generation

Automatic disfluency insertion towards spontaneous TTS - formalization and proof of concept [19]: This work is an exploratory work on the automatic insertion of disfluencies in text-to-speech systems. By inserting pauses, repetitions and revisions, the objective is to make synthetic speech more spontaneous and expressive. To achieve this task, we formalize the problem as a theoretical process, where transformation functions are iteratively composed. This is a novel contribution since most of the previous work either focus on the detection or cleaning of disfluencies in speech transcripts, or solely concentrate on pause insertion in text-to-speech. We present a first implementation of the proposed process using conditional random fields and language models, before conducting objective

and perceptual evaluations. These experiments lead to the conclusion that our proposition is effective to generate disfluencies, and highlights perpectives for future improvements. This work received a best paper award at TALN'2017.

Knowledge extraction, multi-level representation

Characterization of linguistic register: As part of the ANR TREMoLo project, preliminary work has been conducted to investigate the notion of linguistic register. This work has mainly been carried out during the internship of Jade Mekki, with a shared co-supervisision of IRISA/Expression and the MoDyCo lab (UMR 7114) in Nanterre. This work was focused on the manual analysis of register-specific texts, the proposition of linguistic characteristics (either from the state-of-the-art or new ones), and their validation through statistical results. As a results, a large set of descriptors has been proposed in [22]. This work is intended to be published in 2018.

Data acquisition, knowledge extraction

Corpus construction for linguistic registers: Linguistic registers have not really been studied in natural language processing. Hence, no dataset is available, exceptially for French. Hence, a semi-supervised process has been developed in order to build a large corpus of texts annotated in 3 linguistic registers: informal, neutral, formal. This work has been achieved through 2 interships (Benoît Fournier and Hugo Ayats) in the frame of the ANR TREMoLo project. This work exploits descriptors proposed by [22], and jointly enabled the training of a baseline model for register classification. This work is currently under extension in order to be published in 2018.

Use cases and evaluation

International Blizzard challenge: We participated for the third 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 [16] . The system is a concatenative system and uses a selection cost which integrates notably a DNN-based prosodic prediction and also a specific score to deal with narrative/direct speech parts. Unit selection is based on a Viterbi-based 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 [7] . Moreover, a fuzzy function is used to relax this penalty based on the concatenation quality with respect to the cost distribution. Integrating a lot of constraints, this system achieves average results compared to others.

Summary of the contributions

How many in each research focus?

5.3 Defended PhDs and HDRs

- Damien Lolive has defended his *Habilitation à Diriger des Recherches* (HDR) on the 29th of November 2017.
- Raheel Qader has defended his PhD on the 31st of March 2017.
- Marc Dupont has defended his PhD on the 28th of March 2017.

5.4 On going PhDs

1. Nicolas Bloyet has completed his 1st year of research (in the context of a CIFRE grant with the SEED company) addressing the research field of QSAR (quantitative activity - structure relation)

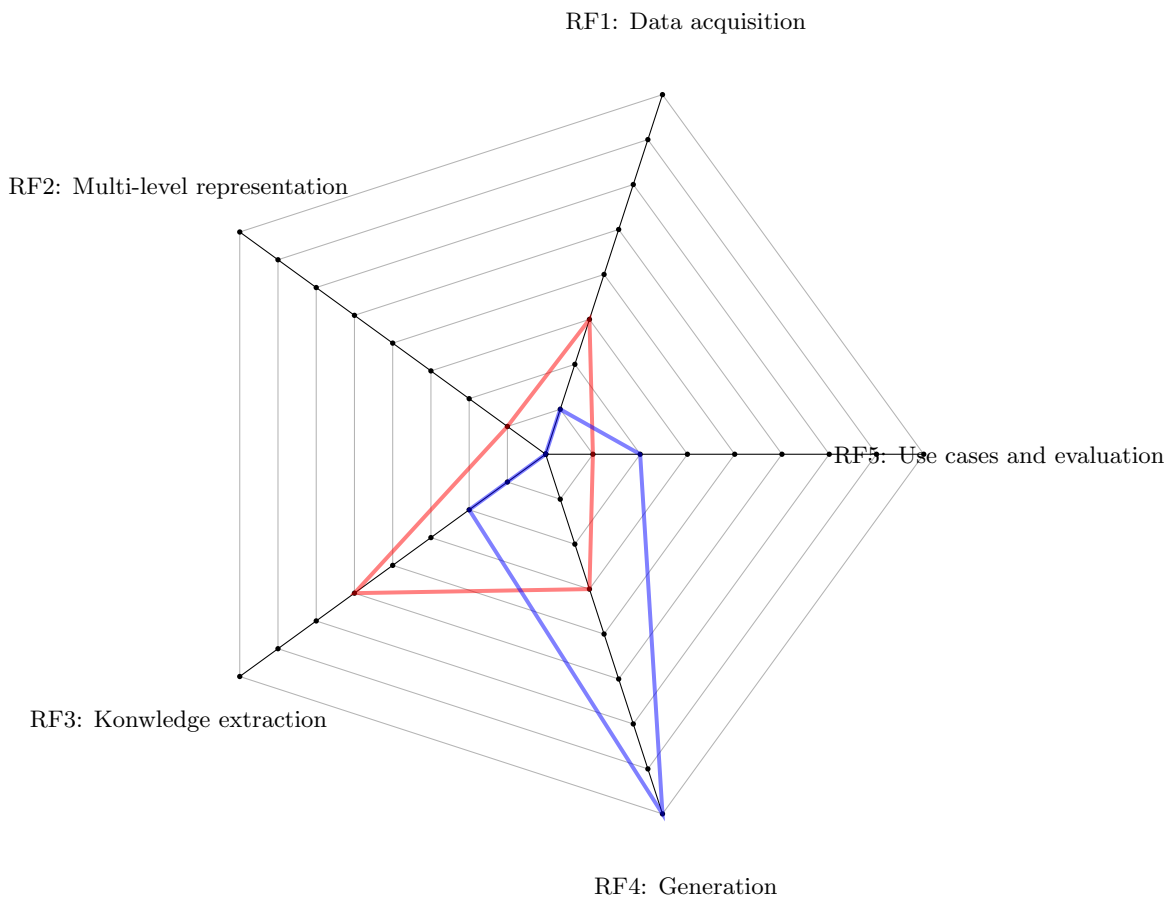

Figure 3: Contributions to each research focus of the team in 2017 (red) compared to 2016 (blue)

that seeks to establish links between observations of the structural nature of a molecule and an "activity", i.e. a physico-chimic property. During this first year, Nicolas Bloyet has carried out some experimentation around graph fragmentation and graph embeddings to improve regression models for QSAR property prediction.

2. Lei Chen has completed her five year of research addressing the definition and analysis of expressive conducting gestures. She has designed and constructed an original multimodal data set composed of gestures and musical excerpts, called CONDUCT. To this end, she defined a set of interacting gestures most suited to expressively control the modulation of sounds. Four main functional categories have been identified which correspond to well known musical effects. Within each category, several expressive variations have been characterized. She has submitted this work to LREC2018, which has been accepted. The data set is used for real-time recognition of gestures and their variations in order to control sound in live performances.

3. Clément Reverdy has completed his third year of PhD. His research addresses the problem of facial expression analysis and synthesis in the context of the animation of signing avatars. He has developed an original mocap data set composed of facial emotional expressions that make sense in the production of sign language utterances. His data set contains different levels of variations among the main affective expressions. He developed a whole blendshape-based motion synthesis pipeline. This blendshape representation is also used to segment motion capture data and automatically annotate this data according to the variations of emotions.

4. Lucie Naert has completed her first year of PhD. The PhD thesis aims at designing and animating signing avatars, i.e. virtual 3D characters signing in signed languages. This is part of 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. This year, she has first studied coarticulation and motion transitions between two consecutive signs and proposed an automatic refinement of manual segmentation based on velocity descriptors of movements [18]. She has then worked on the automatic annotation of one of the channels of sign language data: the hand configuration [17].

5. Stefania Pecore has completed 2nd year 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.

6. 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 have worked on the subjective evaluation of pronunciation modifications and also on disfluencies in spontaneous speech. He defended his PhD on the 31st of March 2017.

[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`.

7. Cédric Fayet has completed his second year. The topic of his research is the detection of anomaly from facial movements and speech signals of a human being. By "anomaly" 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. This year, as a preliminary approach to multimodal anomaly detection, he has focused on the detection of non professional speakers on radio using the acoustic signal (SSPNET-Personality corpus) with several classification methods (GMM, One class SVM and iForest) working on several types of traits (prosodic or Big Five). This work has lead to 3 publications [14, 13, 15]. The availability of a corpus for the detection of anomalies remaining a problem, Cédric has begun the construction of a specific one containing audio and video data. He has submitted this work to LREC2018, which has been accepted. With this corpus, he can begin to deal with multimodal anomaly detection.

8. Antoine Perquin has started his PhD in October. His research addresses new paradigms of speech synthesis opened by recent advances in neural networks and deep learning. The goal of the PhD is to enable generating flexible speech samples based on heterogeneous and massive data. A key aspect within this work lies in properly describing and representing speech variability without relying on expert knowledge. This differs from related work where models are usually trained to produce speech signals, not descriptions, and they are always trained on carefully limited data. During his first year, Antoine compiles pros and cons of many models from the state-of-art, and studies first solutions to adapt them towards the PhD objectives.

9. Meysam Shamsi started his PhD in June 2017 and replaced Sandy Aoun after her resignation. His 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.

10. Aghilas Sini has started her PhD in December 2016. He has completed his first year. 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). During 2017, he worked on the construction of a large speech corpus containing approximately 80 hours of speech. On this topic, we submitted a paper at LREC 2018 which has been accepted.

6 Software

6.1 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.

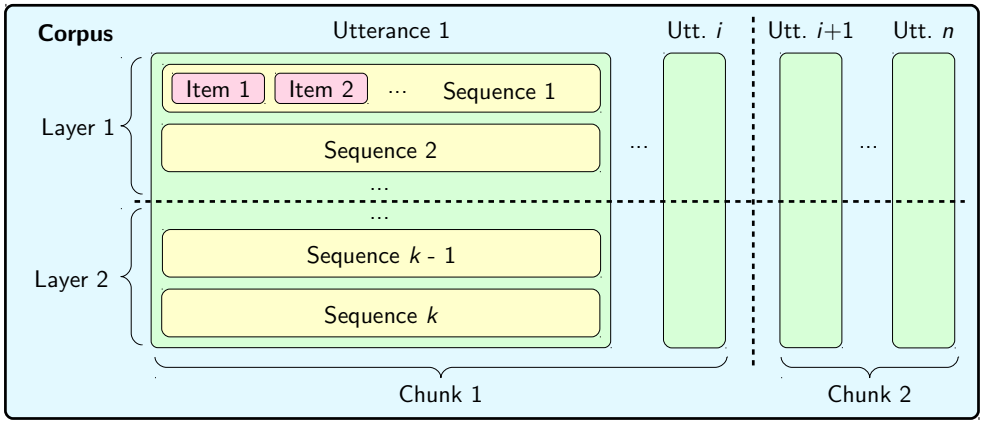

Figure 4: Hierarchical organization of data in ROOTS.

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 4, 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].
- Sequence: add/remove/get/update items; dump,
- Relation: get items related to another; link or unlink items; dump,

[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.

- Utterance: add/remove/get/update sequences; add/remove/get/update direct or composed relations; dump,
- Corpus: add/remove/get/update an utterance; add/remove chunks/layers; load/save; dump.

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

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

[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.

[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

6.2 Web-based listening test system

Participants: Cédric Fayet, Damien Lolive, Claude Simon [from IUT Lannion].

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.).

After a first platform implemented in PHP, we developed a second platform using Python and the Bottle framework to have a much more simple, adaptable and lightweight tool. This new platform is available through the following repository : `https://gitlab.inria.fr/dlolive/PercepEval`. The main idea is to generate a website for each test using simple templates. The platform is able to manage text, sound and video during a single test, ask several questions to the tester at each step of the test, and choose dynamically the right samples to evaluate. Recently, the platform has been extended to produce an tool usable to annotate an audio-video corpus.

6.3 Corpus-based Text-to-Speech System

Participants: Nelly Barbot, Jonathan Chevelu, Arnaud Delhay, 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?

---

sequential annotated data collections", Reykjavik, Iceland, *in : Language Resources and Evaluation Conference (LREC)*, May 2014, `https://hal.inria.fr/hal-00974628`.

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 have been 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. Concerning this point, in 2017, we introduced the possibility to use external complex models to provide target features during synthesis.

6.4 Recording Studio

Participants: Nelly Barbot, Vincent Barreaud, Damien Lolive, Cédric Fayet,.

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.

6.4.1 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).

6.4.2 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 screen or a tablet. Originally, we developed a software, 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.

In 2017, a second software has been developped to record multimodal corpora using both a webcam and a microphone. This software consists of a web-based software relying on HTML5 and providing the same functions as the previous one, but in a very portable way. This work has been done in the context of Cédric Fayet PhD thesis to record both the video of the speaker's face and a microphone. This software also enables to present a dynamic content to the speaker during the recording.

7 Contracts and Grants with Industry

7.1 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`.

7.2 TREMoLo

Participants: Gwénolé Lecorvé, Nicolas Béchet, Jonathan Chevelu, Sabiha Tahrat.

Expression is leader of the ANR project TREMoLo, which has been accepted in December 2016. 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. Official scientific kick-off has been given on the 1st of October 2017 but preliminary activities have been conducted before, between May and July, 2017.

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

8 Other Grants and Activities

8.1 International Collaborations

- In 2017, we have developed a collaboration with Ingmar Steiner and Sébastien Le Maguer from Saarland University, Saarbruck, Germany. Notably, we recruited an internship to work together on the construction of a common interface for Speech synthesis systems enabling to visualize and interact with several systems, like Expression TTS systems and also MaryTTS. In this context, we applied for a funding to continue this work.

- We also carried on our collaboration with John D. Kelleher from DIT Dublin / ADAPT research centre. Damien Lolive has been welcomed in May 2017 at DIT discuss possible collaborations on Language Modelling tools using Deep Neural Networks. We also welcomed John D. Kelleher who visited our team in Lannion, in October, and gave us a seminar on attention-based language modelling.

8.2 National Collaborations

- **VOCAGEN PME project**

  Participant: Nicolas Béchet, Giuseppe Berio, Rémy Kessler

  This project (funded by Pole Images et Réseaux and Région Bretagne) is focused on the building of a software in the field of the construction, allowing users to automatically fill forms starting for the output of a speech recognition system. To this end a concept and term taxonomy is required covering the construction domain. Expression Team is focused on the development of techniques for automatically extracting a relevant terminology and a list of hypernym/hyponym relationships between terms.

  This project is coordinated by ScriptGo compan based in Rennes and involves TyKomz company based in Lannion.

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 different international journals (IEEE Transactions on Affective Computing, IEEE Transactions on Neural Networks and Learning Systems). She has served as a reviewer for international conferences, including Motion Computing (MOCO2017), Sign Language Translation and Avatar Technology (SLTAT2017). She has been a reviewer for the PhD of Raphael Weber on the non supervised construction of an expressive facial model (Supelec).

- 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). He served as an expert for the French research agency (ANR).

- 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 (MOCO2017), 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.
- 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, ICASSP, 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. He is regularly invented hiring committees for the position of Associate Professor. He is part of the organization committee for the conference CORIA-TALN-RJC 2018. He is part of the program committee for JEP 2018.
- 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 had the benefit of a half-time CNRS delegation for 2016-2017. Damien Lolive defended his HDR degree (accreditation to supervise research), delivered by the university of Rennes 1, on the 29th of november 2017.
- Nelly Barbot has served as a reviewer for the International conference of the International Speech Communication Association (Interspeech).
- Nicolas Béchet is a member of program committee of the International Conference on Natural Language Information Systems, and the International Conference on Information Management and Big Data. He has served as a reviewer for different Journals/Conferences like the ACM Journal on Computing and Cultural Heritage and the Springer Scientometrics SCIM.

9.2 Teaching

- Nicolas Béchet teaches various computer sciences courses at the Statistique et Informatique Décisionnelle 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: algorithmic at Bachelor level (Python), bases of signal processing and machine learning (1st year master level), and mocap-based computer animation (2nd year master level).

- 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.

- Jonathan Chevelu teaches the following computer science courses at École Nationale Supérieure des Sciences Appliquées et de Technologie (ENSSAT): cybersecurity in Licence and Master level, operative systems in Licence level and natural language processing in Master level.

- 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 character animation as part of a new introductory lecture of computer science for freshmen and logic for undergraduates of the UFR SSI; C++ programming for ENSIBS graduate students; introduction to computer graphics and computer animation, rendering and interactive techniques at the master level (Master of Computer Science and Master WMR).

- Gwénolé Lecorvé teaches the following computer science courses at École Nationale Supérieure des Sciences Appliquées et de Technologie (ENSSAT): distributed algorithmics; artificial intelligence; and machine learning in Master level. He also teaches automatic speech recognition and speech synthesis in Research Master program of University of Rennes 1, in Rennes.

- 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, speech and language processing in Master level, and pattern recognition 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 de Bretagne Sud.

9.3 Conferences, workshops and meetings, invitations

- Pierre-François Marteau, Nicolas Béchet, and Gwénolé Lecorvé are part of the organization committee of CORIA-TALN-RJC 2018, in Rennes.

9.4 Graduate Student and Student internship

- Antoine Perquin has done his Master level internship in Lannion. He has worked on the construction of phoneme embeddings targeted at expressive speech synthesis. He 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] 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.

[3] 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`.

[4] 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`.

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

[6] 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.

[7] 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`.

[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`.

Publications in Conferences and Workshops

[13] C. Fayet, A. Delhay, D. Lolive, P.-F. Marteau, "Big Five vs. Prosodic Features as Cues to Detect Abnormality in SSPNET-Personality Corpus", *in : Interspeech*, Stockholm, Sweden, August 2017, `https://hal.inria.fr/hal-01583510`.

[14] C. Fayet, A. Delhay, D. Lolive, P.-F. Marteau, "First Experiments to Detect Anomaly Using Personality Traits vs. Prosodic Features", *in : 19th International Conference on Speech and Computer (SPECOM)*, Hatfield, Hertfordshire, United Kingdom, September 2017, `https://hal.inria.fr/hal-01583539`.

[15] C. Fayet, A. Delhay, D. Lolive, P.-F. Marteau, "Unsupervised Classification of Speaker Profiles as a Point Anomaly Detection Task", *Proceedings of Machine Learning Research*, *74*, p. 152–163, September 2017, https://hal.inria.fr/hal-01631385.

[16] D. Lolive, P. Alain, N. Barbot, J. Chevelu, G. Lecorvé, C. J. Simon, M. Tahon, "The IRISA Text-To-Speech System for the Blizzard Challenge 2017", *in : Blizzard Challenge*, Stockholm, Sweden, August 2017, https://hal.inria.fr/hal-01662361.

[17] L. Naert, C. Larboulette, S. Gibet, "Annotation automatique des configurations manuelles de la Langue des Signes Française à partir de données capturées", *in : Journees Françaises d'Informatique Graphique*, Rennes, France, October 2017, https://hal.archives-ouvertes.fr/hal-01649769.

[18] L. Naert, C. Larboulette, S. Gibet, "Coarticulation Analysis for Sign Language Synthesis", *in : International Conference on Universal Access in Human-Computer Interaction*, Vancouver, Canada, July 2017, https://hal.archives-ouvertes.fr/hal-01649815.

[19] R. Qader, G. Lecorvé, D. Lolive, P. Sébillot, "Automatic disfluency insertion towards spontaneous TTS : formalization and proof of concept", *in : Traitement automatique du langage naturel (TALN)*, Orléans, France, June 2017, https://hal.inria.fr/hal-01532031.

[20] R. Qader, G. Lecorvé, D. Lolive, M. Tahon, P. Sébillot, "Statistical Pronunciation Adaptation for Spontaneous Speech Synthesis", *in : Text, Speech and Dialogue (TSD)*, Prague, Czech Republic, August 2017, https://hal.inria.fr/hal-01532035.

[21] M. Tahon, G. Lecorvé, D. Lolive, R. Qader, "Perception of expressivity in TTS: linguistics, phonetics or prosody?", *in : Statistical Language and Speech Processing*, *10583*, p. 262–274, Le Mans, France, October 2017, https://hal-univ-lemans.archives-ouvertes.fr/hal-01623916.

Internal Reports

[22] J. Mekki, D. Battistelli, N. Béchet, G. Lecorvé, "" Nous nous arrachâmes promptement avec ma caisse " : quels descripteurs linguistiques caractérisent les registres de langue ?", *Technical report*, IRISA, équipe EXPRESSION ; MoDyCo, September 2017, https://hal.inria.fr/hal-01649948.