PréGDR TAL
Axe de réflexion : Intermodalité et multimodalité

Benoît Favre (LIS), Pascale Sébillot (IRISA)

2 juillet 2018

1 Participants

Nous tenons à remercier les personnes ci-dessous qui ont répondu favorablement à notre sollicitation et nous ont fourni du matériau pour la réalisation de cette synthèse.

- Bailly, Gérard, GIPSA-lab, CNRS, Gerard.Bailly@gipsa-lab.grenoble-inp.fr
- Barras, Claude, LIMSI, Université Paris Sud, claude.barras@limsi.fr
- Beautemps, Denis, GIPSA-lab, CNRS, denis.beautemps@gipsa-lab.grenoble-inp.fr
- Béchet, Frédéric, LIS, Aix Marseille Université, frederic.bechet@lis-lab.fr
- Blondel, Marion, SFL, CNRS, marion.blondel@cnrs.fr
- Braffort, Annelies, LIMSI, CNRS, annelies.braffort@limsi.fr
- Bredin, Hervé, LIMSI, CNRS, herve.bredin@limsi.fr
- Carrive, Jean, INA, jean.carrive@ina.fr
- Clavel, Chloé, LTCI, Telecom-ParisTech, chloe.clavel@telecom-paristech.fr
- Cerisara, Christophe, LORIA, CNRS, christophe.cerisara@inria.fr
- Damnati, Géraldine, Orange Labs, geraldine.damnati@orange.com
- Estève, Yannick, LIUM, Le Mans Université, Yannick.Esteve@univ-lemans.fr
- Fohr, Dominique, LORIA, CNRS, dominique.fohr@loria.fr
- Fredouille, Corinne, LIA, Université d'Avignon et des Pays de l'Adour, corinne.fredouille@univ-avignon.fr
- Grau, Brigitte, LIMSI, ENSIIE, brigitte.grau@limsi.fr
- Gravier, Guillaume, IRISA, CNRS, guillaume.gravier@irisa.fr
- Le Borgne, Hervé, CEA-LIST, herve.le-borgne@cea.fr
- Lefèvre, Fabrice, LIA, Université d'Avignon et des Pays de l'Adour, Fabrice.Lefevre@univ-avignon.fr
- Linarès, Georges, LIA, Université d'Avignon et des Pays de l'Adour, georges.linares@univ-avignon.fr
- Ringeval, Fabien, LIG, Université Grenoble-Alpes, fabien.ringeval@imag.fr
- Viard-Gaudin, Christian, LS2N, Université de Nantes, Christian.Viard-Gaudin@ls2n.fr

2 Thèmes et périmètre d'étude

L'axe de réflexion *Intermodalité et multimodalité* s'intéresse au traitement du langage dans un contexte où d'autres modalités sont présentes : l'audio (contenant la parole du locuteur, ainsi que des informations sur son identité et son état, ses émotions, ses pathologies, l'environnement acoustique permettant de contextualiser la scène...), l'image (vidéos ou images fixes montrant des mouvements articulatoires, des expressions faciales, la posture des participants à une conversation, la direction du regard, des gestes extra-verbaux, des personnes communiquant en langue des signes, du contenu visuel, des textes, de l'écriture manuscrite...), mais aussi des informations non textuelles issues de capteurs (mesures physiologiques, distances interpersonnelles...) ou de données diverses (localisation physique d'un robot, météo, agenda, cours de la bourse...).

La multimodalité en TAL peut être décrite à travers plusieurs aspects. Le premier est le **langage multimodal** : la communication ne se limite pas à l'échange d'un message textuel, car un message verbal peut être porté par la parole (audio), des gestes (langue des signes, langue parlée complétée) ou des attitudes sociales. Les interactions non verbales étendent le champ de la communication, qu'elles soient volontaires (mime, rires, regards, mouvements de tête, gestes bracchio-manuels) ou involontaires (état et identité du locuteur,

de l'auteur, émotions et opinions, degré de spontanéité, co-adaptation des participants à une conversation, pathologies...), certains aspects pouvant être simulés, créant de ce fait une frontière floue entre volontaire et involontaire. Le deuxième aspect est l'utilisation du langage pour faire référence à des **concepts inter- ou multimodaux** : le langage permet de décrire des concepts et phénomènes du monde réel qui peuvent être saisis par des capteurs, et sont donc représentés à la fois par leur usage dans la langue (définition du dictionnaire, propriétés distributionnelles) et les données physiques issues de ces capteurs (son, image, mesures physiques). On peut souhaiter construire ou manipuler conjointement ces représentations, les comparer, les confronter, exploiter leur co-distribution. Le troisième aspect de la multimodalité en TAL est celui des **applications multimodales** : indexation multimédia, recherche d'information cross-modale, *data mining* dans des corpus oraux ou des documents manuscrits, dialogue homme-machine, robotique, traduction ou reconnaissance de la parole, transcription et traduction en langue des signes, applications à la santé...

L'axe *Intermodalité et multimodalité* est fortement relié aux axes *Modélisation* (étude linguistique et statistique des phénomènes multimodaux), *Apprentissage automatique* (méthodes d'apprentissage spécifiques aux modalités non textuelles, et au contenu intermodal), et *Données* (constitution et annotation de corpus multimodaux). Toutefois, nous prenons soin dans ce document de ne traiter que les spécificités de ces axes dues à la multimodalité. Les éléments d'état des lieux et les enjeux qui sont aussi valables pour le TAL de manière générale ne sont donc pas développés ici.

Actuellement, il n'existe pas en France d'équipe de recherche dont la thématique principale porte sur l'ensemble des aspects recouverts par les notions d'intermodalité et de multimodalité au sein du TAL, et qui s'affiche donc sous cette appellation. La présentation de l'axe au sein de ce document est pour cette raison organisée selon les sous-thématiques suivantes – bien identifiées dans la communauté française – qui se confrontent au TAL dans certains de ses aspects inter- et cross-modaux :

- **le dialogue humain-machine multimodal**, qui concerne d'une part les interactions orales dans le cadre des serveurs vocaux interactifs et les assistants vocaux personnalisés (par ex. Google Home, Siri...), et, d'autre part, les interactions en face à face avec des robots et avec des personnages virtuels ou des objets communicants pouvant prendre des formes pseudo-anthropomorphiques ;
- **les langues des signes**, langues naturelles visuo-gestuelles pratiquées par des communautés de sourds en utilisant des composants corporels et dont les interlocuteurs perçoivent le message par le canal visuel, ou **le langage parlé complété** dans lequel des gestes de la main à proximité de la bouche viennent accompagner le mouvement des lèvres pendant la production de la parole et désambiguïser en particulier la lecture labiale ;
- **les représentations multimodales et modèles joints**, qui s'intéressent à la représentation de plusieurs modalités dans un espace commun ;
- **le traitement du langage oral**, c'est-à-dire les travaux qui exploitent le signal acoustique pour le TAL (sans la reconnaissance automatique de la parole qui est décrite par ailleurs ; mais en incluant la parole pathologique dédiée aux troubles de la parole et de la voix) ;
- **la reconnaissance automatique de la parole**, qui permet de transformer la parole d'un locuteur en un texte, et **celle des caractères**, qui concerne principalement la reconnaissance d'écriture manuscrite ou la lecture de documents imprimés ;
- **les applications multimodales et cross-modales** prenant en compte plusieurs modalités ou le passage d'une modalité vers une autre.

Les applications phares de cet axe *Intermodalité et multimodalité du TAL* concernent la robotique (interactions face à face avec l'utilisateur), les assistants intelligents (domotique, services à la personne), la gestion, l'analyse et la génération de contenu, la recherche d'information, la transcription et traduction de la parole. Le TAL multimodal touche aussi d'autres disciplines scientifiques en tant qu'outil méthodologique, par exemple en humanités numériques (linguistique, histoire, psychologie...), en justice, en journalisme, en éducation, ou dans le domaine médical.

Cet axe a la particularité d'être pluridisciplinaire par nature, d'une part à cause de son interaction avec les disciplines propres à chaque modalité (traitement du signal, vision artificielle, robotique, psychologie...), mais aussi car intégrer le langage dans son contexte de production/captation est une étape souvent nécessaire à la réalisation d'applications pertinentes.

3 État des lieux

L'avancée des travaux dans les domaines du TAL inter- et multimodal dépend de la taille des communautés s'intéressant aux problèmes reliés, de la disponibilité de données et de cadres expérimentaux pertinents, et de l'intérêt des applications dans le monde industriel. Dans certains sous-thèmes, la description des phénomènes n'est pas stabilisée, alors que dans d'autres, des systèmes industrialisés sont déployés. Les systèmes créés pour traiter la multimodalité reposent souvent sur l'apprentissage artificiel et en suivent les avancées, en particulier lorsque beaucoup de données sont disponibles. Contrairement au TAL non multimodal qui repose sur des connaissances linguistiques pour la compréhension et la modélisation de phénomènes, la multimodalité rend le travail descriptif plus difficile, ce qui a pour conséquence de favoriser les approches boîte noire / de bout-en-bout, en particulier issues de l'apprentissage profond. On peut décrire l'état actuel de chacune des sous-thématiques comme suit.

En dialogue multimodal concernant les interactions orales avec des serveurs vocaux interactifs et des assistants personnalisés, les recherches se focalisent sur la perception des comportements verbaux et non verbaux de l'utilisateur. La qualité atteinte en reconnaissance de la parole permet la résolution des références en dialogue, la reconnaissance et classification des actes de langage et l'analyse multimodale des sentiments. Les méthodes de prédiction multimodale de comportements les plus récentes reposent sur de l'apprentissage profond, mais sont encore à développer. Les travaux dédiés aux interactions avec des robots ou personnages virtuels portent sur la sélection et la génération d'énoncés de l'agent en fonction du comportement de l'utilisateur grâce à des méthodes reposant sur l'apprentissage pour la gestion du dialogue. À ceci s'ajoutent des travaux prenant en compte l'aspect situé, c'est-à-dire fusionnant des informations apportées par des modalités de perception (vision, haptique).

Les langues des signes (LS) sont encore peu décrites, et une part des travaux actuels de ce domaine concerne donc leur étude et modélisation linguistique, de la conception d'une forme graphique pour leur transcription à la formalisation de leur grammaire en s'inspirant ou pas de modélisations de langues dites vocales. Dans le premier cas, les énoncés en LS sont modélisés (à base de règles ou par apprentissage) par une séquence d'unités ou classes mais les aspects de multilinéarité, d'iconicité et de spatialisation peinent à être décrits. Dans le second – et c'est le cas pour la LS Française – des approches descriptives propres aux LS permettent le développement de travaux de traduction et de génération. Un pan de recherches est également dédié à l'élaboration de ressources (lexiques et corpus) et d'outils pour les manipuler (par ex. outils d'aide à l'annotation de corpus vidéos par traitement d'images). Notons toutefois que les corpus annotés sont, à de rares exceptions près, encore de petite taille, limitant les possibilités d'applications de méthodes d'apprentissage.

Les travaux en traitement automatique du langage parlé complété (LPC) portent sur la synthèse et la reconnaissance. Des systèmes vidéos exploitant la reconnaissance de la parole d'un locuteur pour superposer à l'image vidéo de son visage une main de synthèse codant les clés du *Cued Speech* (CS, ensemble codifié des gestes codeurs) ont par exemple été développés, ainsi que des systèmes de synthèse visuelle de la parole avec modalité LPC à partir de texte intégrant des règles d'avance de la main pour faciliter la perception par les utilisateurs. Des travaux récents s'intéressent à prédire la position LPC de la main à partir d'une modélisation fondée sur des mélanges de gaussiennes des paramètres spectraux du signal acoustique. En reconnaissance de LPC à partir d'images vidéos, dans lesquelles des artifices sont appliqués aux codeurs LPC (bleu sur les lèvres, pastilles sur le dos de la main...) pour marquer l'information pertinente, des recherches ont permis d'aboutir à des résultats intéressants en reconnaissance des voyelles et de phonèmes extraits de parole continue.

Les représentations multimodales et modèles conjoints visent à exploiter la complémentarité et les redondances apportées par diverses modalités pour améliorer la compréhension (ou à défaut la modélisation) par les machines. Les représentations jointes des modalités (essentiellement) textuelles et visuelles des années 2000, fondées sur des modèles thématiques (*topic models*, par ex. LSA et LDA) ou sur des approches reposant sur de l'analyse de corrélations canoniques (KCCA), sont peu à peu supplantées depuis les années 2010 par de l'apprentissage profond, avec par ex. Devise en 2013, un des premiers *embeddings* texte/image traitant aussi le problème de reconnaissance sans exemple d'apprentissage. L'état de l'art consiste en un traitement parallèle des deux modalités textuelles et visuelles au moyen de réseaux convolutifs ou récurrents, puis en leur projection dans un espace commun avec optimisation par une fonction *triplet loss*. Des méthodes de régularisation peuvent améliorer l'alignement sémantique, ou une mise en correspondance directe des modalités peut être proposée.

L'exploitation du signal acoustique pour le TAL a donné lieu à de nombreux travaux intégrant des indices acoustiques (de bas niveau, prosodiques, phonétiques...) ou, de façon plus indirecte, prenant en compte l'incertitude liée au processus de reconnaissance de la parole au sein des systèmes développés. Les mesures de confiance servent ainsi à limiter l'impact de certains mots dans des représentations de transcriptions et donc leur prise en compte (par ex. en reconnaissance ou segmentation thématique). Des index phonétiques ou lexico-phonétiques sont proposés pour la recherche de segments de parole. Les indices prosodiques sont quant à eux fortement sollicités, en résumé de documents oraux, en analyse de sentiments et détection d'émotions — pour proposer des réactions de systèmes appropriées —, combinés à des modèles de langue pour améliorer la transcription de conversations de réunions, en prédiction du stress d'un locuteur, reconnaissance des actes de dialogue ou encore apprentissage des langues par exemple. La généralisation récente de méthodes d'apprentissage profond intensifie cette intégration d'informations de bas niveau dans des tâches de TAL de haut niveau, et une tendance actuelle est de développer des méthodes *end-to-end* qui font correspondre directement un signal audio avec des sorties symboliques. Des recherches récentes proposent ainsi une architecture de bout-en-bout (encodeur-décodeur avec attention par convolution) de traduction automatique de parole en langue source (prenant en entrée des vecteurs de coefficients cepstraux) vers du texte en langue cible, bousculant la cascade standard d'un système de reconnaissance et d'un système de traduction des transcriptions obtenues. Si ce système peine encore à dépasser les combinaisons précédentes, l'utilisation de décodeurs fonctionnant au niveau des caractères semble ouvrir des perspectives. Pour ce qui concerne la parole pathologique, des traitements automatiques sont dédiés à l'évaluation objective des troubles de la parole et de la voix, mais également appliqués comme outils de détection, thérapeutiques ou de communication alternative pour aider les patients atteints d'un handicap lourd à interagir avec leur environnement. À titre d'exemple, la base de données PC-GITA initiée par l'université d'Antioquia (Colombie) a été utilisée dans le challenge ComParE à Interspeech 2015 pour prédire l'état neurologique de patients atteints de la maladie de Parkinson selon une échelle couramment utilisée par des cliniciens, le corpus CPSD initié par l'université Pierre et Marie Curie a été employé lors de ComParE à Interspeech 2013 pour détecter des troubles autistiques, ou encore le corpus BD collecté par l'université d'Istanbul a été retenu par l'*Audio/Visual Emotion Challenge* lors d'ACM Multimedia 2018 pour classifier des troubles bipolaires. D'autres travaux portent sur l'évaluation de l'intelligibilité. Les approches automatiques de ce domaine se heurtent toutefois au manque flagrant de données disponibles, et à la difficulté de rassembler des corpus volumineux et variés, en particulier dans ce cadre médical soulevant des problèmes éthiques quant à la collecte et à l'ajout de méta-données personnelles aux données.

La reconnaissance automatique de la parole a, elle aussi, subi une révolution depuis les années 2010 avec le développement d'approches neuronales en combinaison, voire en remplacement des modèles de Markov cachés (HMM). Ainsi les modèles de mélanges de gaussiennes (GMM) des modèles acoustiques sont remplacés par des réseaux neuronaux profonds, conduisant à des chutes du taux d'erreur en mots et des performances comparables à celles de l'humain sur certaines tâches. La tendance actuelle vise également à faire disparaître les HMM de la modélisation acoustique, au sein de modèles neuronaux de bout en bout estimés à l'aide d'une fonction de coût CTC (*Connectionist Temporal Classification*). L'apprentissage de ces modèles acoustiques ou des modèles de langage fondés sur des n-grammes ou réseaux neuronaux récurrents nécessite cependant de très grands corpus pour que les systèmes atteignent de bons résultats.

La reconnaissance d'écriture manuscrite non contrainte a connu des avancées significatives avec l'apport des réseaux récurrents, récemment combinés à des réseaux profonds convolutifs. Les systèmes de lecture optique appliqués aux documents imprimés, qui obtiennent de très bons résultats sur des documents contemporains numérisés dans de bonnes conditions, voient leurs performances diminuer lorsque ces dernières se dégradent. Ces systèmes peinent encore à comprendre la mise en page des documents qu'ils doivent analyser et à distinguer finement les objets variés que l'on peut y trouver. Les modèles à attention soulèvent actuellement l'intérêt de la communauté pour intégrer l'ensemble des traitements nécessaires au développement de ces systèmes.

Les avancées en TAL, compréhension et reconnaissance de la parole, vision et apprentissage ont conduit au développement de nombreuses applications multimodales ou cross-modales. Les agents conversationnels, robots-compagnons et assistants vocaux personnalisés, s'appuyant sur les travaux en dialogue homme-machine multimodal, ont ainsi connu un essor majeur. Si certains de ces systèmes sont développés en contexte sémantique bien défini, d'autres sont multi-domaines et souvent couplés à des bases de connaissances générales pour jouer le rôle de systèmes de question/réponse. Tirant parti de l'analyse de l'ensemble des modalités disponibles, des applications visant à donner accès à de vastes archives multimédias ont été créées, que ce

soit sous forme de moteurs de recherche multimédias ou cross-médias, de systèmes de recommandation, ou de systèmes de structuration et de navigation au sein des collections. Au-delà de la gestion de contenus, la production audiovisuelle automatisée s'est également développée avec le montage automatique de vidéos à partir de textes et de résumés automatiques (par ex. Wibbitz) ou l'utilisation des scripts pour le montage. L'accent a également été mis sur l'accessibilité des contenus audiovisuels avec le sous-titrage automatique utilisé de façon opérationnelle (par ex. YouTube), éventuellement via un système de relocution, ou avec l'audio-description d'images ou vidéos, par exemple dans le domaine du sport où des procédés automatiques réalisent de la synthèse vocale à partir des commentaires textuels (phases de jeu, actions, scores). Pour clore ce tour d'horizon applicatif, citons également les travaux exploitant l'analyse des mouvements de la bouche en sus de la modalité audio pour la transcription de la parole ou encore, les travaux qui cherchent à combiner synthèse de la parole et animation faciale.

Comme souligné précédemment, reposer sur l'apprentissage automatique nécessite de grandes quantités de données annotées pour toutes les communautés du TAL multimodal. Des efforts importants ont permis l'émergence de corpus multimodaux de grande taille (parole transcrite, vidéos, images et légendes...) ciblant quelques applications comme la recherche d'information, les questions-réponses visuelles, la traduction... Les annotations associées à ces corpus sont souvent produites à faible coût en tirant parti du *crowd sourcing*, mais leur succès repose sur la facilité d'annotation (descriptions en langage naturel au lieu de structures abstraites). *A contrario*, les tâches demandant une expertise en terme d'annotation sont moins bien dotées ou vraiment sous-dotées (affects, langue des signes, parole complétée, parole pathologique). Le dialogue homme-machine reste un domaine particulier pour lequel les plus grandes bases de données prennent la forme d'interactions humain-humain, différentes de la tâche visée, et n'incluent pas souvent des données liées aux actions et perceptions des systèmes (dialogue situé). Contrairement aux données annotées précédentes majoritairement fermées et payantes (accessibles par LDC ou ELDA), les grands corpus récents peuvent facilement être obtenus. Toutefois, les industriels ont accès à de beaucoup plus grands jeux de données en interne.

4 Grands enjeux

Encore plus que pour le traitement du texte seul, la question de la pertinence de l'apprentissage automatique comme approche majoritaire se pose dans le contexte de la multimodalité. La quantité de phénomènes multimodaux bien compris est relativement faible, et les méthodologies à base d'apprentissage (collecte et annotation de données, création de systèmes, développement par les applications) pourraient appauvrir notre connaissance du domaine. L'hypothèse de représentativité des données n'est pas forcément respectée par les corpus existants, menant à des problématiques d'adaptation et questionnant la valeur scientifique des résultats empiriques. Il semble donc important de construire des lignes de recherche qui permettent de voir au-delà des limites imposées par l'approche majoritaire. À la lumière de cette recommandation, nous abordons d'abord les enjeux principaux de la thématique, puis les enjeux de chacune des sous-thématiques.

Les spécificités du TAL multimodal apparaissent avant tout dans la nature des données traitées. Une piste de recherche dans ce domaine est la prise en compte des signaux faibles et de l'asynchronie entre les sources d'informations issues des différentes modalités. De manière générale, il existe un déséquilibre profond entre les modalités en terme de richesse et de difficulté d'annotation (la subjectivité intrinsèque des émotions face aux efforts de normalisation pour la transcription par exemple). Certaines modalités ont été principalement étudiées dans des conditions favorables et il serait intéressant de les explorer dans un environnement plus écologique et des conditions plus adverses (microphone lointain pour la parole, conversations très spontanées, mauvaise qualité d'image...). Un enjeu particulièrement important est de mieux exploiter les grandes quantités de données brutes audio, textuelles et vidéo disponibles afin de créer des représentations multimodales génériques, hiérarchiques et complexes qui soient indépendantes de la tâche et qui puissent être utilisées dans de nombreuses applications.

L'apprentissage automatique étant de plus en plus incontournable dans le domaine, ses enjeux sont aussi ceux du TAL multimodal. On peut identifier les points suivants qui sont particulièrement importants. Tout d'abord, il serait intéressant de développer la perméabilité de l'apprentissage aux connaissances explicites et, réciproquement, l'extraction de connaissances explicites à partir de systèmes d'apprentissage pour enrichir des bases de connaissances ou ontologies à partir d'expériences multimodales interactives. Savoir mesurer l'utilité de ces connaissances quand de grandes quantités de données sont disponibles est encore un challenge. L'apprentissage de représentations multimodales est une des particularités du domaine, et il est important de

le développer, notamment en raffinant les processus à supervision indirecte qui permettent de les apprendre pour créer des représentations plus interprétables, plus universelles et qui soient capables de s'étendre à de nouvelles modalités. Les modèles de bout en bout, s'ils sont développés dans des applications phares comme la traduction automatique ou la transcription de parole, sont encore difficiles à entraîner dans des domaines où il y a peu de données, et il reste des tâches pour lesquelles cette approche n'a pas encore été envisagée et pourrait être bénéfique. Il semble aussi important de réussir à faire coopérer des briques logicielles traitant des sous-tâches dans un contexte de bout en bout.

Dans le contexte des applications interactives, la multimodalité peut être associée à un certain nombre d'enjeux. Ainsi, les modèles engagés dans une boucle d'interaction doivent être capables d'opérer avec des horizons courts, quitte à corriger les estimations en ligne. Cette incrémentalité est intéressante dans les champs d'application où les données évoluent rapidement, et, en particulier, dans un contexte industriel. Alors que les phénomènes de co-adaptation (convergence...) sont bien connus en sciences du langage, ils sont encore peu exploités dans les systèmes automatiques. Les modèles interactifs doivent pouvoir s'adapter rapidement à l'usager et être capables de percevoir l'adaptation de ce dernier. Cette co-adaptation est d'autant plus complexe qu'elle opère à de multiples niveaux : phonétique, phonologique, lexical, syntaxique... L'évaluation des systèmes multimodaux et interactifs en particulier est balbutiante. Il reste à organiser des challenges spécifiques permettant aux équipes participantes de situer leurs travaux de manière équitable et reproductible.

Un autre enjeu de la communauté est celui de réussir l'interdisciplinarité au sens large, alors que les laboratoires sont surtout structurés par discipline. On peut citer les exemples de collaboration avec les sciences cognitives et les neurosciences pour mieux comprendre les fonctions linguistiques dans le cerveau, ou encore avec les sciences du mouvement pour mieux capturer les gestes non-verbaux en reconnaissance de LSF. Le TAL multimodal appelle à la collaboration de disciplines, certainement bien au-delà de ce qui est tenté actuellement au vu des découpages sous-thématiques des équipes et laboratoires de recherche.

Au-delà de ces enjeux globaux au domaine, chaque sous-thématique du TAL multimodal a des enjeux spécifiques qui sont décrits dans les paragraphes suivants.

L'arrivée sur le marché des robots-compagnons et le développement des assistants vocaux personnalisés et des *chatbots* vont engendrer des attentes fortes de la part des usagers de ces systèmes de dialogue multimodal. Les verrous scientifiques auxquels ce domaine doit se confronter portent d'une part sur l'intégration de la composante socio-émotionnelle dans l'interaction afin de doter les systèmes d'une intelligence sociale ; d'autre part sur la gestion de la parole spontanée – potentiellement dans des cadres multi-parties – pour laquelle les corpus écologiques de grande taille font globalement défaut ; enfin sur l'utilisation de méthodes d'apprentissage profond pour mieux prédire les comportements utilisateurs tout en s'adaptant à leurs particularités et d'apprentissage par renforcement pour la gestion du dialogue. Des partenariats plus resserrés avec des disciplines telles que la robotique (en particulier pour le dialogue situé), la psychologie et la sociologie sont à développer.

L'étude des langues des signes dans un cadre pluridisciplinaire (linguistique, informatique, science du mouvement...) est encore à encourager pour mieux maîtriser les systèmes linguistique et biomécanique de ces langues et leurs modélisations, étendre les capacités de génération d'énoncés en LS en incluant des structures linguistiques plus complexes, aboutir à des animations de signeurs virtuels moins robotiques qu'actuellement, et ouvrir la voie à des applications pour l'accessibilité et la traduction assistée par ordinateur. La constitution de corpus annotés de très grande taille, permettant l'application de techniques d'apprentissage en particulier pour l'analyse de vidéos de LS et le développement d'applications de saisie gestuelle, d'indexation de vidéos et de recherche de contenus dans celles-ci, est également un objectif à atteindre.

Au même titre que pour les LS, la synthèse de LPC et sa reconnaissance à partir d'images d'un locuteur posent encore un nombre important de défis, parmi lesquels on peut par exemple citer l'asynchronie des flux d'informations, ou la perception et l'extraction des informations utiles sans marquage *a priori*. Diverses méthodes d'apprentissage neuronales peuvent être envisagées pour extraire, à partir d'images brutes, des descripteurs visuels pertinents au regard de la discrimination phonétique, pour modéliser l'aspect séquentiel et la dé-synchronisation lèvres-mains, pour intégrer des connaissances linguistiques, ou pour repérer des régularités de régions d'intérêt détectées automatiquement par des techniques issues de vision par ordinateur.

Les enjeux du domaine des représentations multimodales portent d'une part sur les représentations jointes elles-mêmes, avec l'étude de l'ancrage des symboles dans différentes modalités. Les niveaux auxquels les informations issues des diverses modalités (fusion précoce, tardive ou intermédiaire) doivent être intégrés pour construire des représentations sont encore à préciser pour résoudre une tâche de prédiction. L'apprentissage

sans exemple (par ex. reconnaître un concept dans une image à partir de sa seule description mais sans exemple visuel) reste un défi, tout comme l'apprentissage joint consistant à exploiter les relations entre informations issues de différentes modalités pour apprendre des modèles. Le problème de l'alignement bien connu en TAL (entre énoncés en relation de paraphrase, d'implication...) se décline dans le cadre multimodal en s'intéressant à la façon d'identifier des relations entre éléments issus de l'analyse d'une modalité (entités, concepts, objets...) pour plusieurs modalités. La traduction, c.-à-d. l'exploitation de modalités en entrée pour en produire une autre (production de gestes visuels d'avatars à partir de transcription de la parole et du flux audio, génération de légendes pour des images...) est encore à étudier. Enfin, la production même de corpus multimodaux pose question : quelles images sont par exemple pertinentes pour illustrer des concepts et des instances ?

Pour ce qui est du TAL oral, un premier enjeu, particulièrement crucial pour le traitement de l'oral spontané, est le développement de modules de TAL aptes à exploiter un espace de transcriptions possibles avec des mesures de confiance associées à chaque symbole verbal mais aussi non-verbal émis (prise en compte des dimensions acoustiques du message (prosodie, bruits non verbaux), intégration des disfluences produites). Par ailleurs, l'oral étant souvent produit en contexte dialogique, traiter des conversations nécessite de considérer l'aspect dynamique et collaboratif résultant de l'échange entre les participants, décoder les signes verbaux ou non-verbaux indiquant la compréhension ou non-compréhension en situation d'analyse, et savoir générer de tels signaux afin de rendre la communication naturelle dans le cadre de l'interaction personne-machine. Concevoir des modèles qui apprennent le contenu informationnel présent dans des données orales est aussi un défi, par exemple apprendre la segmentation en phonèmes et des unités linguistiques de façon non supervisée. Enfin un des enjeux majeurs de ce domaine est le développement de méthodes de bout en bout performantes qui transforment directement le signal audio en séquences de mots (éventuellement annotés) sans passer par l'étape de transcription de parole. Au-delà des avancées technologiques, pour que le traitement automatique de la parole pathologique puisse envisager des avancées significatives, une perspective évidente est la collecte de grands volumes de données tenant compte d'un maximum de variabilité intra- et inter-populations. Le développement de plateformes collaboratives telles que CloudCast (univ. Sheffield) permettant de centraliser la collecte et la sécurisation des données pourrait être une piste à court terme.

En reconnaissance de la parole, les enjeux principaux portent sur le développement de systèmes neuronaux de bout en bout apprenant simultanément tous les composants et modélisant conjointement contenu linguistique, para-linguistique et conditions acoustiques ; sur l'étude de la robustesse au bruit, aux conditions acoustiques, aux accents, changements de langue au sein d'une phrase, en un mot à la variabilité ; sur l'intégration de la reconnaissance à d'autres tâches (traduction, extraction de concepts...) ; sur l'apprentissage faiblement supervisé et l'apprentissage par transfert pour les langues peu dotées ; ou encore, sur la compréhension de ce que dit le locuteur ou de son intention en prenant en compte le contexte de l'interaction et éventuellement la gestuelle.

Dans le domaine de la lecture automatique d'imprimé et de manuscrit, un des objectifs est le développement de systèmes de bout en bout intégrant connaissances optiques, linguistiques et sémantiques. La technologie OCR se focalise actuellement fortement sur la conversion image vers texte et un des enjeux futurs concerne l'accès à la sémantique des documents (champs informatifs à extraire, compréhension selon les domaines métiers). La spécialisation des systèmes à des écritures avec peu de données ou pour une langue inconnue reste aussi un challenge.

Les applications multimodales à destination du grand public ont encore de nombreux défis à relever. Les assistants vocaux doivent par exemple parfaire leur compréhension fine de la parole, tout en étant capables de détecter le domaine abordé et les connaissances générales à convoquer. La détection du focus dans une conversation multi-locuteurs reste un problème ouvert pour ces technologies (traité actuellement avec de la détection d'un mot-clé d'activation) qui nécessitera probablement une approche multimodale. L'accès en vocal à des bases de connaissances *via* des systèmes de question-réponse est également à développer. Les applications de dialogue, devenant un objet familier des foyers, doivent savoir traiter le langage des enfants. Les interactions doivent aussi gagner en naturel. Les utilisateurs des diverses applications sont également friands d'explications, c.-à-d. d'interprétabilité des algorithmes (pourquoi un système me recommande tel achat ?). Dépasser le paradigme du tout automatique pour intégrer le collaboratif et savoir propager des corrections faites dans un élément de base à toutes les phases d'un système est également un enjeu.

Pour conclure, un enjeu primordial est celui de ramener la multimodalité au centre du champ disciplinaire et de la considérer comme objet d'étude à part entière. Rassembler une communauté identifiée et

positionnée devrait permettre de fédérer les initiatives pluridisciplinaires et de rendre évidentes des questions fondamentales pour le domaine sur lesquelles il était difficile jusqu'à aujourd'hui de se concentrer.

5 Positionnement

5.1 Éthique

Les applications de traitement automatique de LSF issues des recherches ont pour objectif de répondre à un besoin sociétal, appuyé par une loi parue en 2005 sur l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées, qui a aussi reconnu la LSF comme langue à part entière. À cela, s'ajoutent les aspects juridiques et éthiques liés au droit à l'image pour la constitution de corpus de LSF.

Le déploiement des assistants vocaux à large échelle, avec des systèmes comme Alexa d'Amazon ou Google Home, pose le problème de collecte de données et vie privée.

Dans le cadre du traitement automatique des troubles de la parole et de la voix, la collecte des données est le plus souvent régie par les établissements hospitaliers sous une réglementation et des procédures propres à chaque pays et qui tendent à se renforcer avec la protection accrue des données personnelles. Cette collecte est très dépendante de la disponibilité des patients, de leur nombre, de leur consentement à participer au programme d'enregistrements, de l'évolution de leur maladie, de leur degré de fatigabilité... Répondre à des exigences quantitatives (large nombre de patients et de sujets contrôles appareillés en genre et en âge, équilibre des populations de patients) et qualitatives (pour répondre aux critères d'inclusion, à des niveaux de variabilité différents dans le stade de la maladie, dans les degrés de sévérité, dans l'âge des patients, etc.) est un objectif difficile à atteindre et qui soulève de nombreuses interrogations quant au respect de la vie privée. Par ailleurs, l'enrichissement des données audio par des informations relatives au patient, à sa pathologie, au stade d'évolution de sa maladie, aux traitements qu'il suit, aux évaluations perceptives réalisées par des cliniciens lors de consultations peut s'avérer fournir des éléments aussi importants que le signal lui-même et aider à l'analyse des résultats et la comparaison entre patients. Mais à nouveau, la préservation de l'anonymat d'une part et l'immiscion très poussée dans la vie privée d'autre part sont des aspects à considérer avec soin.

5.2 Interface

Principaux laboratoires Les principaux laboratoires académiques en France ayant des groupes de recherches travaillant sur les aspects multimodaux du TAL sont le CEA-LIST à Saclay, le GIPSA-Lab à Grenoble, le LATTICE à Paris, le LIP6 à Paris, l'IRISA à Rennes, l'IRIT à Toulouse, le LITIS à Rouen, le LIA à Avignon, le LIG à Grenoble, le LIMSI à Paris, le LIS à Marseille, le LIUM au Mans, le LORIA à Nancy, le LPP à Paris, le LS2N à Nantes, le L3I à La Rochelle et Telecom-ParisTech/LTCI, auxquels s'ajoutent des groupes de R&D tels que Orange Labs par exemple.

Il n'existe en fait pas vraiment d'équipes concrètement identifiées par un thème principal « TAL multimodal », mais des groupes de recherche ayant des travaux pouvant se positionner sous cette thématique mais affichant des axes ou sujets d'études plus traditionnels : traitement de la parole, traduction automatique, recherche d'information...

Sociétés savantes En France, le TAL est représenté par la société savante ATALA (Association pour le Traitement Automatique des Langues), regroupant des chercheurs provenant des domaines de la linguistique, de l'intelligence artificielle et de l'informatique ; le traitement automatique de la parole se retrouve dans l'AFCP (Association Francophone de la Communication Parlée), société savante issue à l'origine de la Société Française d'Acoustique et regroupant des chercheurs en traitement de signal, des phonéticiens et des informaticiens. Diverses autres associations sont de près ou d'un peu plus loin concernées par la thématique du TAL vu sous ses aspects *Intermodalité* et *Multimodalité*, parmi lesquelles on peut citer l'AFIA (Association Française pour l'Intelligence Artificielle) et notamment son groupe de travail « Affects, Compagnons Artificiels et Interactions » (ACAI) qui regroupe les activités en France autour de l'informatique affective et de l'interaction avec des compagnons artificiels et organise notamment le *workshop* biannuel WACAI, l'ARIA (Association francophone de Recherche d'Information et Applications), l'AFIHM (Association Francophone d'Interaction Homme-Machine), l'AFRIF (Association Française pour la Reconnaissance et l'Interprétation des Formes) et la SSFAM (Société Francophone d'Apprentissage Machine). Un des enjeux pour développer

cette thématique du TAL est de faire en sorte que les membres de ces associations se côtoient, prennent connaissance des avancées des uns et des autres, et travaillent ensemble.

Lien avec les GDR et autres instituts

- GDR ISIS : la communauté du signal, des images, et de la vision par ordinateur entretient déjà des rapports forts avec le TAL, notamment dans le domaine du traitement conjoint de documents multimédias (textes, sons, images...), ceci étant traité dans le thème transverse « Apprentissage pour l'analyse du signal et des images » ;
- GDR Robotique : le groupe GT5 « Interactions personnes/systèmes robotiques » rassemble des chercheurs travaillant sur les interactions homme-robot, robot-robot avec des ouvertures sur l'Internet des objets et la réalité virtuelle ;
- GDR IG-RV : GTAS, groupe de travail sur l'animation et la simulation ;
- ILCB (Institute for Language, Communication and the Brain) : projets entre TAL et sciences cognitives (modèles de la lecture, amorçage sémantique, simplification de textes), neurologie (neuro-imagerie, *mind reading*, modèles de compréhension de la parole) et linguistique (co-adaptation, dialogue oral).

Tâches partagées et compétitions Un certain nombre de tâches partagées et compétitions scientifiques permettent d'évaluer les avancées dans le domaine du TAL multimodal, parmi lesquelles ASPiRE (reconnaissance de la parole), ComParE (émotions et traits paralinguistiques), ImageCLEF (captioning), IWSLT (traduction de la parole), MediaEval (recherche d'information, accès et exploration de contenus multimédias), MGB-challenge (reconnaissance de la parole et du locuteur), SRE (reconnaissance du locuteur), TrecVid (recherche d'information vidéo, hyperliage...), VQA (question-réponse visuelle), WMT (traduction multimodale).

Points marquants Ci-dessous, une liste de points marquants montrant des forces présentes en France dans le domaine du TAL inter- et multimodal :

- langue des signes : prise en compte de la multilinéarité, de l'iconicité et de la spatialisation grâce à une modélisation linguistique originale développée au SFL ; existence d'un projet traitant ensemble les problématiques de la LPC et des langues des signes au GIPSA-lab
- traduction multimodale au LIUM, avec la co-organisation de WMT 2017 (`http://www.statmt.org/wmt17/multimodal-task.html`)
- traduction *speech-to-text* au LIG par des approches neuronales *end-to-end* (1^{er} système de ce type)
- représentation multimodale : IRISA vainqueur de la tâche *Hyperlinking* de TRECVid 2016 à l'aide d'une représentation cross-modale texte-image
- Défi REPÈRE 2015 gagné par un consortium LIS, LIA, Orange Labs, LIFL
- journée GDR-ISIS sur TAL et image/vidéo organisée en particulier par le CEA-LIST et le LIMSI (`http://www.gdr-isis.fr/index.php?page=reunion&idreunion=358`)

Logiciels Alors qu'il existait plus de logiciels directement dédiés à des tâches finales par le passé, les logiciels développés à l'heure actuelle sont des boîtes à outils destinés aux développeurs. Pour l'analyse de la parole, le logiciel le plus utilisé actuellement en recherche est Kaldi (Povey et al., 2011). En image, OpenCV est une référence. Pour développer des réseaux de neurones, Tensorflow (Google, `https://www.tensorflow.org/`), pytorch (Facebook, `https://pytorch.org/`) et CNTK (Microsoft, `https://github.com/Microsoft/CNTK`) permettent de définir et d'entraîner différentes architectures et incluent de nombreux exemples liés à la multimodalité, ainsi que des fonctionnalités pour traiter les données non textuelles (chargement et traitement des fichiers, augmentation des données, métriques d'évaluation). Les *toolkits* SRILM et IRSTLM permettent d'entraîner des modèles de langage pour la reconnaissance de la parole. En analyse d'émotions, on peut citer les logiciels SSI (Social Signal Interpretation), EMospeech (`https://emospeech.net/`), Audeering, IBM Watson Tone Analyzer (`https://tone-analyzer-demo.mybluemix.net/`). En dialogue homme-machine et systèmes multi-agents, on peut lister les outils : GRETA (`http://pages.isir.upmc.fr/~pelachaud/site/resources.html`), Virtual human Toolkit (`https://vhtoolkit.ict.usc.edu/`), Flipper Multi-Model Dialogue System (`https://hmi-utwente.github.io/FlipperMMDS/`), Disco - Collaborative Discourse Manager (`https://github.com/charlesrich/Disco`). Il est en train de devenir standard dans la communauté de distribuer les sources des logiciels ayant servi aux publications scientifiques à des fins de réplication, mais

sans qu'il y ait toujours un réel souci de pérennité, réutilisation et reproduction des résultats scientifiques (souvent sur la plateforme GitHub).

Médiation scientifique De nombreuses interventions de Laurence Devillers (LIMSI) sur le thème des interactions robots-humains : https ://www.franceculture.fr/personne-laurence-devillers livre : Des robots et des hommes chez Plon Les Rendez-vous du Futur : février 2018...

6 Programmatique

6.1 Communauté

Le langage naturel est multimodal par essence, alors que la communauté TAL s'est principalement concentrée jusqu'à présent sur le message verbal véhiculé et ce, très majoritairement, sous sa forme écrite. Dans le contexte actuel de mise en avant d'applications liées à l'intelligence artificielle et à la robotique, il semble stratégique de développer des forces autour du traitement du langage multimodal, c'est-à-dire du langage porté par les différentes modalités. Les avancées en apprentissage automatique, notamment avec le *deep learning*, rendent les approches multimodales beaucoup plus accessibles (logiciels et modèles pré-entraînés disponibles par exemple), et l'expertise présente en France sur un certain nombre de thématiques du TAL mettant en jeu diverses modalités montre que les forces en présence sont prêtes à aborder un tel challenge.

Comme indiqué précédemment, bien qu'il y ait de nombreux travaux portant sur des aspects du TAL inter- et multimodal, il n'existe pas, à l'heure actuelle, en France de communauté visible travaillant sur le TAL multimodal en temps qu'objet de recherche. Il serait donc souhaitable de créer une telle communauté s'intéressant à tous les aspects de la communication langagière verbale et non verbale au-delà de la seule modalité texte. Cette communauté aurait pour objectif d'étendre le champ d'application du TAL sur tous les aspects de la multimodalité, tant en analyse qu'en production : reconnaissance et synthèse de la parole, de langues des signes, de caractères ; prise en compte des phénomènes non verbaux tels que les regards, mouvements de tête, gestes bracchio-manuels, mimes, rires ; prise en compte de l'état et de l'identité du locuteur, de l'auteur, des émotions et des opinions, du degré de spontanéité du langage, des pathologies ; modélisation de la co-adaptation des participants à une conversation...

Un groupe de travail (GT) pourrait être créé pour faire émerger et animer cette communauté du TAL multimodal. Les objectifs de ce groupe de travail seraient de développer un modèle du langage multimodal, de l'exploiter dans des applications à fort impact sociétal, et de renforcer les liens avec d'autres disciplines. Ce GT pourra explorer un certain nombre de questions importantes pour le TAL qui sont particulièrement intéressantes dans un cadre multimodal, comme par exemple l'adaptation à des publics spécifiques (prise en compte du handicap, du niveau de langue, de l'âge...), la mutualisation des méthodologies et des modèles pour l'analyse et la génération, l'évaluation de la pertinence des systèmes de TAL multimodal (comment évaluer la production de contenu multimodal ? Comment établir le meilleur canal de communication d'un message ?)...

Dans un premier temps, nous proposons de mettre en avant trois thématiques liées à ces objectifs :

— **modèle du langage multimodal** : construire des corpus et cadres d'évaluation pour mieux étudier la multimodalité du signifiant en intégrant les dernières avancées en linguistique et disciplines associées conjointement avec l'établissement de modèles computationnels correspondants ;
— **dialogue et robotique** : créer des outils de TAL multimodal à destination de la communauté robotique mettant en avant la généricité et la couverture des applications possibles ;
— **TAL multimodal et cerveau** : étudier comment l'être humain manipule le langage multimodal en collaboration avec les sciences cognitives et neurosciences, dans le but de mieux comprendre et de bio-inspirer la création de représentations multimodales.

Les actions à mener dans le cadre de ce GT pourraient être :

— l'organisation de conférences jointes entre la communauté du TAL et celle d'une communauté sœur afin d'aider à la constitution de la communauté TAL multimodal
— l'organisation de campagnes d'évaluation spécifiquement dédiées aux aspects multimodaux du TAL, pérennes pendant la durée du GT pour permettre de mesurer les progrès
— la création de corpus multimodaux
— l'organisation de journées d'étude thématiques

— le soutien de sessions spéciales et *workshops* dans de grandes conférences

6.2 Jeunes chercheurs

Afin d'attirer les jeunes chercheurs dans cette communauté, il serait souhaitable de créer ou participer à l'organisation d'une école d'été avec des cours spécifiques sur la multimodalité, mettant en avant la multidisciplinarité. L'organisation de sessions de travail monolocalisées longues à finalité commune (de type hackathon, *JHU workshop*), rassemblant jeunes chercheurs et mentors issus de la communauté, permettrait d'avancer significativement sur des problèmes spécifiques tout en motivant les jeunes chercheurs à travailler sur ces thématiques par la suite.

Il serait aussi intéressant de mettre en place des sessions jointes dans les conférences à destination des jeunes chercheurs des sous-communautés du TAL multimodal pour que ceux-ci se rencontrent et établissent un champ de recherche commun.

Il faudra enfin entreprendre des actions de lobbying auprès des financeurs de thèses cross-disciplinaires pour mettre en avant des sujets sur la multimodalité, en particulier sur les applications à fort impact sociétal.

6.3 Médiation scientifique

Une possibilité de médiation scientifique serait de s'associer à un ou plusieurs grands événements de vulgarisation pour mettre en avant la multimodalité.