Proposition de recherche doctorale

Analyse multimodale de scènes musicales et interaction avec un agent émotionnel

Mots clés :

- **Directeur de thèse** : GAEL RICHARD
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire Traitement et Communication de l'Information
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

{{Contexte}} Les robots sont de plus en plus présents dans les espaces sociaux, tels que les écoles, les hôpitaux ou les musées, et l'homme doit alors interagir au quotidien avec eux, de la manière la plus naturelle possible. Pour améliorer la qualité de ces interactions, les robots sont maintenant dotés d'intelligence sociale, qui leur permet d'analyser plus en profondeur la situation, et de réagir en démontrant des changements d'humeur ou d'émotion. Certains robots sont alors considérés comme des compagnons, et créent de véritables liens sociaux avec les humains. Ils jouent par exemple un rôle clé dans des traitements contre l'autisme par leur capacité à divertir . Ainsi, il existe des robots acteurs, dessinateurs, acrobates [Pais2011], et une multitude de robots pouvant jouer d'un instrument de musique, comme WF-4RII, un joueur de flûte doté de poumons . L'expressivité de leur jeu musical reste néanmoins limitée, cela étant dû entre autres à la complexité de la relation entre geste musical et émotion. Nous allons ici nous pencher sur cette relation en prenant la situation où le robot détecte et interprète les émotions transmises par le musicien, puis signifie sa compréhension en reproduisant l'émotion perçue. Dans le cadre de la thèse, il est également envisagé, au moins dans un premier temps de représenter le robot par un avatar (ou agent) numérique ce qui permet de limiter les contraintes mécaniques et d'aborder plus directement les applications du monde numérique. Il s'agit d'abord de développer un système de classification émotionnelle, traitant des informations multimodales (vidéo, son, voire capture de mouvements), qui permettra la modélisation des émotions transmises. Il s'agit ensuite de modéliser l'impact de ces émotions sur le robot, et de proposer un modèle de réactivité émotionnelle. La génération de comportement reposera sur la plateforme GRETA [Niewiadomski2011]. Les deux situations d'étude privilégiées seront d'une part les situations en intéraction (ou « live ») où le robot est le spectateur d'un musicien et d'autre part les situations où le robot visualise un flux audio visuel (vidéo musicale en particulier). Dans le premier cas, les réactions du robot (ou agent numérique) pourraient servir comme retour (ou « feedback ») au musicien, qui lui permet de décider de l'évolution de sa performance alors que dans le second cas, les réactions constitueront de l'enrichissement de contenu. {{Sujet Détaillé}} Les émotions impliquent des phénomènes complexes et ne peuvent pas être décrites sans l'implication de multiples dimensions (physiologiques, cognitive, contexte,..). Il y a ainsi de très nombreux travaux sur l'analyse et la caractérisation des émotions et sur les modèles théoriques que l'on peut construire même s'il n'existe pas de consensus sur les modèles appropriés (voir par exemple pour un état de l'art du domaine dans [Pelachaud2011]). Les émotions ont été utilisées dans des systèmes informatiques depuis une quinzaine d'années faisant naître un nouveau domaine : l'informatique affective (ou Affec-tive computing) [Picard1997]. Il est classique de distinguer l'émotion exprimée (par exemple par le musicien) de l'émotion perçue (par exemple par un spectateur) [Gabrielsson2002], mais c'est surtout cette dernière qui est modélisée notamment dans les systèmes d'indexation automatique car elle est moins sensible à la situation et au contexte d'écoute (v. par exemple [Yang2011]). Un système de reconnaissance automatique de l'émotion peut suivre les principes généraux d'un sys-tème de classification (extraction de caractéristiques, classificateur). Historiquement, les premiers sys-tèmes s'attachaient à reconnaître des émotions prototypiques (joie, tristesse, peur,..). Les modèles plus récents décrivent l'émotion dans un espace continu bidimensionnel (ou de plus grande dimension) ap-pelé espace émotionnel en utilisant la « valence » (faisant référence au plaisir, l'aspect positif ou négatif d'un état émotionnel et « l'Arousal » (qui fait référence au niveau d'activation ou de stimulation) [Russel1980][Grimm2007]. Ainsi, le problème de reconnaissance d'un état émotionnel s'apparente à prédire un score émotionnel (en utilisant les échelles continues valence et arousal). De telles approches peuvent être utilisées pour analysées des extraits de films [Clavel2008] ou catégoriser des collections de documents audio [Yang2011]. De nombreux travaux visent à découvrir de nouveaux descripteurs de signaux qui sont corrélés avec les scores émotionnels et certains travaux peuvent utiliser en parallèle plus de 300 descripteurs [Schul-ler2012]. En termes de méthodes de classification, un large panel de méthodes ont d'ores et déjà été utilisées mais la récente prise en compte d'espaces émotionnels continus de dimension croissante ont renforcé l'intérêt des méthodes de régression. Le sujet de cette thèse {{Analyse multimodale de scènes musicales et interaction avec un agent émotionnel spectateur}} entre tout à fait dans le cadre de l'informatique affective mais vise des applications originales pour lesquelles de nouvelles méthodes devront être développées. En particulier, il sera nécessaire de développer de nouvelles méthodes dédiées pour l'analyse, la reconnaissance et la caractérisation des émotions dans ce contexte et de développer de nouveaux modèles d'interaction émotionnelle par rapport à l'état de l'art et aux travaux déjà menés à Télécom ParisTech [Clavel2008, Pelachaud2012].

Résumé du projet de recherche (Langue 2)

En particulier, ce travail s'attachera à traiter les aspects cruciaux suivants : {{Définir l'espace émotionnel approprié aux situations contextuelles envisagées.}} Il s'agira dans un premier temps de définir « l'émotion musicale », quels concepts elle intègre (émotion exprimée par le musicien, émotion ressentie,..) et quel espace émotionnel est pertinent. Sur ce dernier point, si une approche catégorielle reste possible en considérant que plusieurs émotions peuvent définir un état émotionnel, il semble maintenant incontournable de considérer l'émotion dans un espace continu. Dans le cadre de ce travail de caractérisation, il pourra être fait appel à des notions de psychologie cognitive [Juslin2011] notamment pour les parties relatives à la musique, aux spécificités et relations des modalités utilisées, et à la contagion émotionnelle où des critères dépendant du contexte et de l'expérience devront être modélisés. Un autre aspect particulièrement important, bien que souvent sous-estimé dans les études récentes, concerne la dynamique de l'émotion. Il s'agira en particulier de construire des modèles prenant en compte cette dynamique en intégrant d'éventuelles dépendances avec le genre ou style musical, le passé proche ou lointain du contexte d'écoute, … Prédire un score émotionnel (ou reconnaître un état émotionnel) à partir d'un flux de données multimodales hétérogènes : Champs aléatoires conditionnels et Régression à base de noyaux. De nombreux travaux se sont limités à utiliser des approches de classification ou de régression sans prendre réellement en compte la spécificité du phénomène émotionnel et notamment sa dynamique. Dans cette thèse, un enjeu particulièrement important sera de détecter et suivre les changements d'émotion au cours du temps. Plusieurs approches pourront être envisagées, mais dans un premier nous utiliserons les Champs Conditionnels Aléatoires (Conditional Random Fields ou CRF) qui permettent à la fois de modéliser de telles dynamiques temporelles tout en prenant en compte de manière adéquate l'hétérogénéité des données. Par ailleurs, de tels modèles ont déjà donnés d'excellents résultats pour des problèmes d'alignement audio sur partition midi [Joder2012] et des résultats très prometteurs en reconnaissance des émotions sur de petites bases de clips musicaux [Schmidt2011]. L'utilisation des CRF pour un problème de régression est ici rendu possible en discrétisant l'espace émotionnel, sans pour autant définir des catégories émotionnelles précises. On considérera également d'autres approches de régression qui semblent particulièrement pertinentes pour ce problème telles que les Support Vector Regressor (déjà utilisés dans des études préliminaires sur les émotions [Han2009][Yang2011]), les approches plus prometteuses tels que les Kernel Rigde Regressor (KRR) et leurs variations parcimonieuses, les Reduced Rank Kernel Ridge Regressor (RRKRR) [Basak2007] et de manière plus générale l'ensemble des approches de régression à noyau (Kernel Based regression ou KBR). En particulier pour les approches pré-citées, et à l'instar d'études menées dans d'autres domaines [Sun2008], on s'intéressera à les rendre suffisamment flexibles pour accepter diverses formes d'information a priori permettant de contraindre le problème et de trouver des solutions plus satisfaisantes au regard des connaissances et de la spécificité de la scène analysée. Une autre approche envisagée à court terme consistera à exploiter de multiples régresseurs en parallèle dans le but de les fusionner ou de les intégrer dans un méta-classifieur à l'image des approches de boosting. Ces régresseurs spécifiques pourront par exemple être appris pour des sous-ensembles de caractéristiques identiquement corrélées avec les scores émotionnels. {{Obtention de bases de données représentatives.}} Comme pour tout problème de classification, il est essentiel de pouvoir disposer de bases de données annotées représentatives du problème à étudier pour l'apprentissage des modèles et l'évaluation des systèmes développés. Plusieurs approches seront ici suivies. Dans un premier temps, le travail se concentrera sur l'utilisation de données existantes. Il existe par exemple un grand nombre de vidéos en ligne, informées (ou taggées) en émotion par les utilisateurs qui permettent de se constituer rapidement des bases intéressantes même si les annotations sont dans ce cas toujours assez bruitées [Bischoff 2009]. Une alternative est d'enrichir des bases de données existantes. Actuellement, plusieurs projets d'obtention de données par jeu collaboratif ont été réalisés. Moodswings qui est l'un de ces projets [Kim2008] a notamment combiné de tels jeu avec l'analyse de tags libres. Considérant que ces projets ne prennent en compte que la modalité audio, un des enjeux sera d'étendre de telles bases avec des données vidéos. Ces données pourront soit être retrouvées par recherche sur Internet soit être générées en studio. L'une des originalités de la seconde approche pourrait être d'enregistrer une performance mimée d'un musicien sur les musiques correspondantes un peu à l'image des prestations de « Air guitar » très populaires. Les modalités vidéos et audio pourraient ainsi être corrélées en termes d'émotion même si la modalité vidéo ne correspondrait pas précisément au rendu visuel du contenu audio. Enfin, il apparaît important de devoir constituer une base de données spécifique au problème de la thèse. Une telle base qui sera enregistrée dans le tout nouveau studio multimodal de Télécom ParisTech sera conçue au cours de la première année de thèse avec l'ambition d'une diffusion large à la communauté pour renforcer l'impact et la reproducibilité des travaux réalisés dans cette thèse {{Modélisation de l'Interaction émotionnelle avec robots ou agents virtuels}} C'est une partie essentielle de cette thèse qui sera probablement menée dans un second temps. Cependant, les problématiques et les enjeux de cette interaction seront pris en compte dans les approches d'analyse de l'émotion développées au cours de la thèse. Afin de limiter les différentes contraintes d'une interaction avec un robot physique (contraintes mécaniques et physiques, impacts sur la qualité des signaux d'acquisition et sur les éventuels problèmes d'implémentation matériel si les algorithmes sont déportés dans le ou les robot(s)), on se consacrera dans un premier temps à un spectateur sous la forme d'un agent virtuel (ou avatar). L'enjeu ici sera de modifier l'activité habituelle de cet avatar par des démonstrations d'émotion [Grunberg2012]. L'un des aspects novateurs sera de considérer l'intégration d'éléments subjectifs pour reconstituer une émotion induite personnalisée suivant l'avatar (par exemple en fonction de ses goûts ou de sa personnalité) et suivant le musicien (voir par exemple [Yang2007] pour des travaux préliminaires dans cette direction). Le calcul des mouvements du robot et de l'agent virtuel reposeront sur la plateforme GRETA [Niewiadomski2011]. Celle-ci permet de calculer le comportement multimodal expressif pour communiquer des intentions communicatives et émotionnelles. La plateforme GRETA peut contrôler aussi bien un agent virtuel qu'un agent physique comme le robot NAO. Cependant, il apparaît d'ores et déjà que des aspects de synchronisation dynamique seront essentiels notamment lorsque les gestes seront rythmés et en phase avec le contenu musical. On pourra ici considérer des approches d'optimisation des paramètres de synchronisation entre les émotions, le mouvement multimodal et le rythme. Le passage à une interaction avec des robots physiques sera réalisée dans un second temps tout d'abord en s'affranchissant de l'aspect temps réel (comme par exemple dans l'étude [Xia2012] où le robot écoute d'abord une première fois la musique pour générer une chorégraphie qu'il synchronisera à la musique lors de la deuxième écoute). L'aspect temps réel sera ensuite considéré sachant le temps d'analyse du robot ne sera pas négligeable et qu'il devra aussi être étudié selon la variabilité émotionnelle de la musique. Dans notre problème d'interaction, étant donné qu'une seconde de musique est suffisante p