Date de début et durée

01/09/2023 - 3 ans

Master(s) EIPHI en lien avec le projet :

GREEM (Control for Green Mechatronics)
VIBOT (VIsion & roBOTics)
MAIA (Master Degree in Medical Imaging and Applications)

Résumé (1500 caractères)

L'intéroception, notre capacité à percevoir les signaux de notre propre corps (voix, cœur, muscles), est considérée en psychologie et en neurosciences comme cruciale pour réguler nos émotions, et son altération serait impliquée dans de nombreuses pathologies comme l'autisme ou la dépression. Pourtant, ces communautés scientifiques manquent d'outils technologiques permettant de contrôler expérimentalement la perception des signaux corporels et valider ainsi ces théories. En réunissant des compétences bourguignonnes et franc-comtoises en traitement du signal et automatique, notre projet propose (1) de développer de nouvelles méthodes pour contrôler en temps-réel les expressions faciales (ex. faire sourire notre visage) et (2) d'intégrer les algorithmes de déformations faciales dans la plateforme de transformation vocale existante à FEMTO-ST pour aboutir au 1er système temps réel à même d'étudier l'effet de l'altération de l'intéroception sur une personne.

Description scientifique du projet (6300 caractères)

Les expressions faciales et la voix sont deux "signaux corporels" fondamentaux du répertoire expressif humain. Ils influencent ce que nous pensons d'une personne (par exemple, en la rendant plus attirante [ODoherty2003]) et également la façon dont nous nous comportons à son égard (par exemple, en faisant preuve d'une plus grande empathie [Surakka1998]). De manière intéressante, la perception de nos propres signaux corporels a aussi une influence sur nos émotions et nos comportements. Il existe plusieurs modèles théoriques décrivant nos émotions comme résultant de la perception de nos propres signaux corporels (e.g. [Prinz2004]). Par exemple, la perception de notre cœur qui s'emballe et de nos muscles qui se contractent peut être identifiée à la *peur*. À partir de là, de nombreux travaux ont étudié l'impact de l'altération de la perception de nos propres signaux corporels sur l'émergence des émotions [Aucouturier2016], du sentiment de stress ou même de troubles de la régulation émotionnelle [Zamariola2019] ou de l'autisme [Dubois2016].

Pourtant, il manque aujourd'hui les outils technologiques permettant de contrôler expérimentalement la perception des signaux corporels et valider ainsi ces théories. Dans le projet ASIMOV, nous proposons de développer un algorithme de transformation visuelle capable de manipuler un flux vidéo entrant en *temps réel* afin de contrôler de façon paramétrique les expressions faciales et d'intégrer ces méthodes dans un dispositif de contrôle boucle-fermée à même d'étudier leur effet sur une personne. Ce dispositif se doit d'être innovant vis à vis de l'état de l'art sur deux points : d'une part, étant donné la sensibilité des êtres humains à de petites désynchronisations entre les mouvements des lèvres et la voix, la contrainte de temps réel est critique dans ce projet et va nécessiter le développement d'algorithmes optimisés. D'autre part, le réalisme des transformations est

également un second point critique qu'il faudra considérer tout au long des développements algorithmiques.

Les transformations seront basées dans un premier temps sur la détection et le suivi de points caractéristiques du visage comme les coins des yeux et des lèvres et une manipulation de leurs positions en utilisant un modèle paramétrique prédéfini. Ces stratégies d'altération offrent des perspectives très intéressantes pour aboutir à des transformations réalistes et temps réel. En effet, les techniques de détection des points caractéristiques faciaux ont été améliorées très significativement ces dernières années et offrent maintenant des performances remarquables en termes de précision et en termes de temps de calcul (e.g. avec mediapipe). Cependant, la dépendance de cette méthode à un modèle constitue une limitation de cette approche et va nécessiter de déterminer de manière expérimentale, par exemple via des techniques basées sur la corrélation inverse (i.e. reverse correlation [Ponsot2018]), les modèles de transformations correspondant aux représentations mentales associées à des expressions faciales cibles.

Une seconde piste intéressante va porter sur l'utilisation de stratégies inspirées des réseaux antagonistes génératifs (Generative Adversarial Networks ou GANs). Ces réseaux sont capables de générer des images ou de l'audio de très grande qualité mais offrent bien souvent assez peu de contrôle sur les caractéristiques des données générées. Il existe tout de même des travaux très intéressants, comme par exemple GanSpace [Harkonen2020], où Härkönen et al. recherchent des directions de contrôles interprétables pour des GANs existants. Ils ont montré que les composantes principales, obtenues en appliquant une simple réduction de dimensions ACP dans l'espace latent, du GAN correspondaient finalement à des contrôles *interprétables*. Ils permettent de modifier les attributs de l'image, qui vont de propriétés simples comme la pose et la forme de l'objet, à des propriétés plus sophistiquées comme l'éclairage ou les attributs du visage (e.g. intensité du sourire). On peut aussi mentionner des architectures comme CycleGAN [Zhu2017] qui permettent de transformer une image d'un domaine à un autre (par exemple appliquer la texture d'une image sur une seconde image) ou GANimation [Pumarola2019] qui propose de conditionner un GAN par des annotations des unités d'action (AU - Action Units). Cette seconde classe de méthode, orientée données et non plus modèle, offre une piste très intéressante pour les transformations des expressions faciales mais nécessite également une optimisation très importante pour viser les applications temps réelles et à faibles latences. De plus, ces réseaux génératifs modifient des images et leur extension au domaine vidéo va très certainement apporter des challenges associées aux limitations bien connues des GANs où les transitions dans l'espace latent souffrent de discontinuité, produisant parfois des changements abrupts dans les sorties et donc dans notre cas des discontinuités entre des images successives d'une vidéo.

Les algorithmes développés seront intégrés dans la plateforme actuellement en cours de développement à FEMTO-ST portant les transformations temps réelles de la voix pour aboutir au 1er système temps réel à même d'étudier l'effet de l'altération de l'intéroception sur une personne.

Bibliographie

[ODoherty2003] J. ODoherty et al. "Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness," Neuropsychologia, 2003.

[Surakka1998] V. Surakka and J. K. Hietanen, "Facial and emotional reactions to duchenne and non-duchenne smiles," International Journal of Psychophysiology, 1998.

[Prinz2004] J.J. Prinz, “Gut reactions: A perceptual theory of emotion”, Oxford University Press, 2004.

[Zamariola2019] G. Zamariola et al., “Relationship between interoception and emotion regulation: new evidence from mixed methods”. Journal of Affective Disorders, 2019.

[Dubois2016] D. DuBois et al., “Interoception in autism spectrum disorder: A review”, International Journal of Developmental Neuroscience, 2016.

[Harkonen2020] E. Harkonen et al., “GANSpace: Discovering Interpretable GAN Controls”, NeurIPSystems, 2020

[Zhu2017] Zhu et al., “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”, ICCV 2017.

[Pumarola2019] A. Pumarola at al. , “GANimation: One-Shot Anatomically Consistent Facial Animation”, IJCV, 2019.

[Aucouturier2016] J-J Aucouturier et al., Covert digital manipulation of vocal emotion alter speakers’ emotional states in a congruent direction, PNAS, 2016.

[Ponsot2018] E. Ponsot, “Uncovering mental representations of smiled speech using reverse correlation”, Journal of the Acoustical Society of America, 2018,

1 ou 2 illustrations

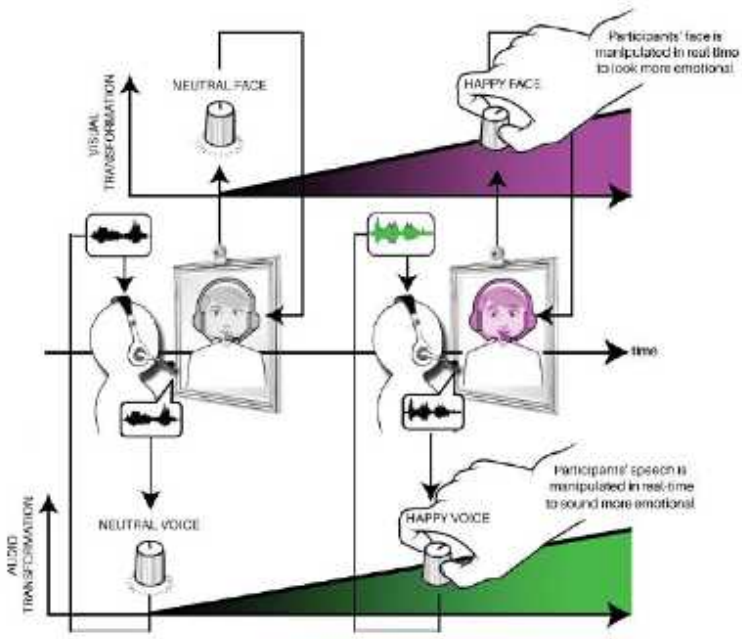

Schéma expérimental visé de l’étude de l'altération de l’intéroception.

Synergie formation-recherche, ancrage dans les formations (1000 caractères)

La thématique du projet entre dans celles des masters GREEM (pour l'aspect systèmes intelligents), VIBOT et MAIA (pour les aspects traitement du signal, vidéo et IA). Les ressources technologiques et humaines de ASIMOV permettront la mise en place d'équipes projet avec des étudiants de ces différentes formations intéressés par le traitement du signal et de l'image et aussi de l'intelligence artificielle. Nous prévoyons bien sûr de recruter en priorité des étudiants venant des Masters EUR EIPHI identifiés et de proposer des projets étudiants dans le cadre de modules pédagogiques ou même hors maquette (pour les M2 MAIA et VIBOT). Nous souhaitons mettre en place un système de tutorat où les étudiants en M2 auraient la possibilité d'encadrer les projets étudiants et où le doctorant recruté participerait à l'encadrement des étudiants en stage M2. Idéalement, les projets étudiants prépareraient au stage et le stage à la thèse de doctorat.

Proposition(s) de stage(s) ; pour quel(s) master(s) ? (1000 caractères)

Le projet prévoit de recruter un stagiaire (M2, 6 mois) par an.

- 1 sujet master en 1ere année du projet sur la détermination des modèles de transformations visuelles des expressions faciales correspondant aux représentations mentales des utilisateurs via des méthodes de reverse correlation.
- 1 sujet master en 2nd année du projet portant sur le conditionnement du GAN pour les transformations des expressions faciales.
- 1 sujet master M2 en 3ème année de projet sur l'intégration des algorithmes de transformations visuelles sur la plateforme Labview de vocal feedback existante à Femto-ST/AS2M.

Positionnement du projet par rapport à l'état de l'art (1000 caractères)

Il existe quelques travaux portant sur la modification de flux audiovisuel altérant les expressions faciales. Dans le domaine visuel, certaines techniques contrôlent les paramètres morphologiques d'un visage (par exemple, le soulèvement des joues, ouverture de la bouche [Arias2020]) ou même des techniques basées sur des approches d'apprentissage profond pour apprendre les transformation expressives à partir de pairs d'images de visages [Xiao2018]. Cependant, il est important de mentionner que ces méthodes ne fonctionnent bien souvent pas en temps-réel.

[Arias2020] P. Arias et al., "Realistic Transformation of Facial and Vocal Smiles in Real-Time Audiovisual Streams", IEEE Trans. on Affective Computing, 2020

[Xiao2018] T. Xiao et al., "Elegant: Exchanging latent encodings with gan for transferring multiple face attributes", ECCV 2018.

Rupture, prise de risque (1000 caractères)

Si les progrès observés ces dernières années sur les techniques de manipulation des visages sont très importants, ceux-ci concernent en général le réalisme des images

générées. La tendance actuelle étant une course en termes de degré de liberté où les réseaux utilisant plusieurs centaines de milliards de paramètres proposés par les entreprises GAFA (e.g. GPT-3) obtiennent des performances remarquables. Les outils de transformation d'images, basés sur les GANs, ne peuvent bien souvent pas fonctionner en temps réel et leur optimisation est une tâche délicate car réentrainer ces réseaux est très complexe. Nous proposons ici de rechercher le meilleur compromis entre le réalisme des images générées et le temps d'inférence. La manipulation des positions des landmarks faciaux en utilisant un modèle paramétrique prédéfini reste une alternative crédible aux modèles basés sur les GANs avec l'avantage d'être plus facilement optimisable pour fonctionner en temps réel.

Multi ou pluri-disciplinarité (1000 caractères)

La technologie de transformation visuelle intégrée à la plateforme de vocal feedback de Femto-ST va représenter un outil sans équivalent ouvrant la voie à de nombreuses expérimentations interdisciplinaires avec les communautés en psychologie et neurosciences.

Structuration à l'échelle régionale (dont collaborations avec laboratoires, industriels ou autres entités régionales), structuration intra-EIPHI et/ou inter Graduate Schools (1000 caractères)

Nous proposons de développer des algorithmes de transformations visuelles temps réel et de les intégrer dans une plateforme existante à FEMTO-ST/AS2M. Cette plateforme, sans équivalent actuellement, résulterait donc de travaux issus de Dijon pour les transformations dans le domaine visuel et Besançon pour les transformations de la voix.

Internationalisation et/ou ouverture nationale (dont collaborations avec labos, industriels ou autres entités nationales ou internationales) (1000 caractères)

Les 3 masters EIPHI concernés par le projet sont des masters internationaux ou Erasmus+. Cette ouverture favorise grandement le recrutement d'étudiants internationaux pour les projets et stages associés au projet, le futur parcours à l'international de nos étudiants passés par le projet et continuant en thèse dans des laboratoires à l'étranger, ainsi que le montage de projets bilatéraux ou européens impliquant des collaborateurs à l'étranger.

Connexion avec le monde industriel et valorisation (1000 caractères)

Le projet est propice à créer des entreprises ou à vendre des licences d'exploitation sur les technologies créées. À titre d'exemple, la licence accordée récemment à la société Unissey

(anciennement deepsense) suite à un projet de maturation conduit par imViA/SAYENS sur une technologie de mesure physiologique par analyse vidéo. Il permettra de proposer des nouvelles prestations vers les industriels en termes de collaboration et de possibilité de bourse de thèse CIFRE.

Impact sur le tissu socio-économique régional et/ou les grands enjeux sociétaux (1000 caractères)

Le projet s'intègre dans le bassin régional des systèmes intelligents. Par ses visées applicatives, il s'insère également dans l'écosystème biomédical de la région notamment dans TEMIS Santé.
D'un point de vue sociétal, le projet permettra d'initier des travaux portant par exemple sur l'autisme ou les troubles de la régulation émotionnelle ou du stress.

Effet levier (dont cofinancements acquis et/ou envisagés) (1000 caractères)

ASIMOV pourrait bien sûr servir de support pour répondre à d'autres AAP à l'échelle nationale ou internationale. Il est important de préciser ici que le projet ASIMOV porte sur les développements technologiques. L'utilisation de ces nouveaux outils fera l'objet d'autres travaux collaboratifs et interdisciplinaires qui pourront porter sur l'étude de la dépression, de l'autisme ou de la régulation émotionnelle par exemple.
De plus, nous visons un modèle de dissémination des algorithmes développées sous la forme de logiciels open-source qui pourront être réutilisés par d'autres chercheurs en neuroscience ou en psychologie.

Culture scientifique, technique et industrielle (impératif pour les thèses) (1000 caractères)

L'équipe projet sera impliquée dans des actions à destination du grand public comme la Fête de la science, la nuit des chercheurs, les 24h du Temps et les journées portes ouvertes des établissements UBFC. Le lien formation recherche sera développé à travers, entres autres, la participation aux journées du GDR ISIS (e.g. la journée du GDR Visage, geste, action et comportement). Le colloque consacré à l'Enseignement des Technologies et des Sciences de l'Information et des Systèmes constitue aussi une relation privilégiée entre collègues académiques de l'enseignement supérieur et du secondaire.

Contribution de chaque partenaire et justification si le projet est monopartenaire (1500 caractères)

L'équipe d'ImViA impliquée dans ce projet travaille maintenant depuis plusieurs années sur l'analyse du visage pour la reconnaissance d'expression faciale ou l'estimation des

mouvements et changements colorimétriques associés aux signaux vitaux (rythmes respiratoires ou cardiaques par exemple). L'étudiant recruté travaillera donc principalement sur le développement des transformations visuelles par deep learning et modèles paramétriques prédéfinis avec l'encadrement de l'équipe d'ImViA.
Les travaux sur la corrélation inverse pour déterminer les modèles de transformations correspondants aux représentations mentales associées à des expressions faciales cibles seront menés avec l'équipe de FEMTO-ST/AS2M.
Enfin, l'intégration des algorithmes développés dans la plateforme fera l'objet également d'un travail conjoint entre les deux équipes.

Planning de déroulement du projet (1000 caractères)

Un état de l'art pour identifier les différents réseaux GAN existants sera réalisé tout au long du projet. Lors de la 1ere année, l'étude de reverse correlation sera menée pour déterminer les modèles de transformation adaptées et les premières approches de transformations visuelles seront développées.
Lors de la seconde année, les travaux porteront plus particulièrement sur les méthodes de transformations visuelles basées sur les GANs. Ces deux classes de méthodes seront comparées expérimentalement au regard de leur performance en termes de réalisme des transformations et rapidité d'exécution en 3ème année du projet et les méthodes sélectionnées seront intégrées dans la plateforme de transformations vocales existantes à FEMTO-ST/ AS2M.

Plan de financement

ImViA (129700€)
1 thèse : 115k€
2 stages M2 : 3600*2 = 7200
déplacement : 5000€
matériel : 2500€

FEMTO-ST : 15600
1 stage M2 : 3600€
matériel : 10k€
déplacement inter site : 1000e
petits matériels : bons d'achat pour participants aux exp : 1000e