



INS2I -2016

## APPEL A PROJETS

## PEPS INS2I 2016

## Identification

Nom du porteur du projet	Jean-Julien AUCOUTURIER
Adresse e-mail du porteur	<a href="mailto:aucouturier@gmail.com">aucouturier@gmail.com</a>
Titre long du projet	Warping Affectif de la Rétroaction Multimodale
Acronyme du projet	WARMER

## Résumé du projet (10 lignes maxi)

Cette proposition vise à mettre en place une nouvelle collaboration entre deux équipes du périmètre thématique de l'INS2I, l'une spécialiste de traitement du signal audio (UMR9912 à Paris) et l'autre de signal visuel (UMR6164 à Rennes), autour de la conception d'un «miroir émotionnel déformant» : l'observateur interagit en parlant devant le miroir ; son signal visuel est capturé par une caméra, et sa parole par un microphone. A son insu, son visage reflété par le miroir est déformé algorithmiquement en temps-réel pour le faire paraître e.g. plus souriant, et sa parole entendue au casque est déformée pour la faire paraître e.g. plus joyeuse. D'un point de vue théorique, cette collaboration permettra de concevoir un modèle de déformation de forme audiovisuelle à partir de points de contrôle, unifiant pour la première fois des modèles développés indépendamment dans les communautés signal visuels et sonores. D'un point de vue pratique, le projet ambitionne des applications cliniques pour la prise en charge non-médicamenteuse des pathologies émotionnelles comme la dépression et l'anxiété

Nom partenaire	Qualité / Titre	e-mail partenaire	Unité de recherche
Jean-Julien Aucouturier (*)	CR CNRS	<a href="mailto:aucouturier@gmail.com">aucouturier@gmail.com</a>	STMS UMR9912
Marco Liuni	CR Ircam	<a href="mailto:marco.liuni@ircam.fr">marco.liuni@ircam.fr</a>	STMS UMR9912
Renaud Segulier	PU Supelec	<a href="mailto:renaud.seguier@centralesupelec.fr">renaud.seguier@centralesupelec.fr</a>	IETR UMR6164
Catherine Soladié	MdC Supelec	<a href="mailto:catherine.soladie@centralesupelec.fr">catherine.soladie@centralesupelec.fr</a>	IETR UMR6164

**Objectifs scientifiques et technologiques :**

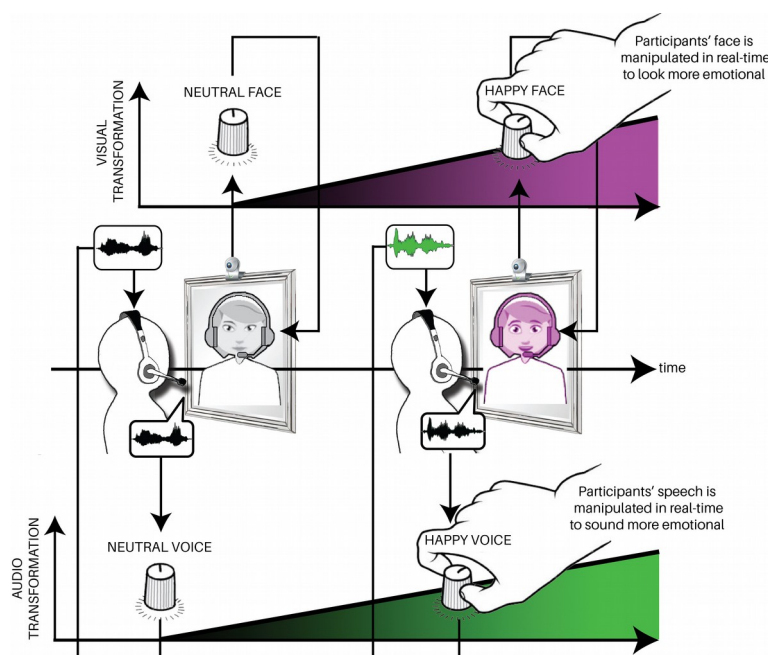
L'objectif de cette proposition est de mettre en place une nouvelle collaboration entre deux équipes du périmètre thématique de l'INS2I, l'une spécialiste de traitement du signal audio (STMS UMR9912 à Paris, dont le porteur est issu) et l'autre spécialiste du traitement du signal visuel (IETR UMR 6164 à Rennes).

**L'objectif scientifique** de cette collaboration est (1) d'avancer sur des problématiques algorithmiques liées à la synthèse temps-réel d'expressions émotionnelles dans le domaine audiovisuel (i.e., voix et visage) et en particulier d'étudier l'interaction algorithmique entre ces deux modalités ; (2) de permettre une application de ces technologies à des problématiques cliniques importantes dans nos sociétés occidentales (i.e., le besoin de prise en charge non-médicamenteuse de désordres affectifs comme la dépression et l'anxiété).

**L'objectif technologique** de cette collaboration est de développer un prototype de miroir déformant audiovisuel, où l'observateur se voit et s'entend s'exprimer de façon de plus en plus positive, grâce à une manipulation temps-réel de son *feedback* audio et visuel (Illustration 1) : l'observateur interagit en parlant devant le miroir ; son signal visuel est capturé par une caméra, et sa parole par un microphone. A son insu, son visage reflété par le miroir est déformé algorithmiquement pour le faire paraître e.g. plus souriant (avec des techniques de *piece-wise linear warping* – voir plus loin), et sa parole entendue au casque est déformée pour la faire paraître e.g. plus joyeuse (avec des techniques de *phase vocoder warping* – voir plus loin).

**Ambition :** Il existe un état de l'art important en traitement du signal facial et vocal visant à synthétiser des expressions émotionnelles réalistes (Vinciarelli, Pantic et Bourlard, 2009), motivé par des applications dans le domaine des interfaces homme-machine avec avatars, tutoriels intelligents et autres *serious games*. Malgré ce contexte, notre proposition reste particulièrement ambitieuse car elle se donne des contraintes nouvelles et mal couvertes par les algorithmes existants : d'une part, nous visons des transformations visage/voix si réalistes que l'observateur lui-même puisse s'y tromper. Il ne s'agit pas seulement d'éviter les indices classiques d'artificialité (voix robotique, mauvaise reconstruction du visage), mais de produire des expressions audiovisuelles qui fassent

partie de l'espace individuel de l'observateur, et pas d'une autre personne. Cette contrainte est évidemment différente de celle de la synthèse d'un avatar impersonnel, fut-il photo-réaliste (Stoiber, Segulier & Breton, 2009). D'autre part, le contexte de «rétroaction continue» (i.e., alors que le locuteur est en train de s'observer parler) nous contraint à un temps-réel radical qui est lui aussi peu traité par la littérature. Par exemple dans le domaine vocal, la lecture à voix haute devient inconfortable dès que la latence de la rétroaction est supérieure à quelques dizaines de millisecondes seulement (Fairbanks & Guttman, 1958). ; en comparaison, les algorithmes actuels de conversion de voix nécessitent souvent une fenêtre d'analyse de plusieurs centaines de millisecondes pour permettre une résolution fréquentielle suffisante, voire l'accumulation de données encore plus longue pour des modèles prédictifs (voir par ex. Toda, Muramatsu & Banno, 2012)



*Illustration 1: Schéma de principe de la boucle de rétroaction visuelle et vocale modifiée: l'observateur se voit dans un miroir et s'entend au casque alors que des algorithmes de traitement du signal visuel et sonore modifient en temps-réel son visage et sa voix afin de les déformer dans le sens d'une plus grande expression émotionnelle, par ex. ici un visage plus souriant et une voix plus joyeuse.*

**Nouveauté :** le principal facteur de nouveauté du projet consiste en l'application de ces technologies de traitement du signal visuel et sonore dans le domaine clinique. Il existe un très fort besoin de nouveaux moyens de prise en charge non-médicamenteuse des pathologies émotionnelles liées par ex. à la dépression et à l'anxiété, qui sont endémiques dans nos sociétés occidentales. Dans un récent article publié dans PNAS (Aucouturier et al, 2016), nous avons démontré que la seule partie «audio» du dispositif ci-dessus était capable de créer une induction émotionnelle forte chez l'adulte sain (les personnes s'entendant lire avec une voix modifiée pour être plus positive deviennent plus joyeuses, et celles qui s'entendent lire avec une voix plus négative deviennent plus tristes). Ce premier résultat, qui montre la faisabilité du dispositif, a attiré l'attention de plusieurs collègues cliniciens (Pitié-Salpêtrière, Cochin, Service de Santé des Armées) car il n'existe actuellement aucun dispositif clinique permettant d'intervenir sur la boucle sensorimotrice vocale, et a fortiori multimodale voix/visage. Cette proposition permettra donc d'installer un nouveau dialogue entre chercheurs informatique et partenaires cliniciens, autour de thématiques fortes de l'INS2I (traitement du signal visuel et sonore), et dans un domaine à fort impact sociétal.

### **Pertinence du projet :**

L'objectif de cette proposition est de mettre en place une nouvelle collaboration entre deux équipes CNRS spécialisées dans le traitement du signal sonore/vocal (équipe CREAM de l'UMR9912) et le traitement du signal visuel/facial (équipe FAST de l'UMR6164). Cette collaboration s'inscrit au cœur des thématiques de l'INS2I, en mettant en interaction les aspects *Traitement du signal* et *Synthèse de signaux temporels* pour l'aspect voix avec *l'Informatique graphique et la Réalité augmentée* pour l'aspect visage, avec un soucis commun de *Système temps-*

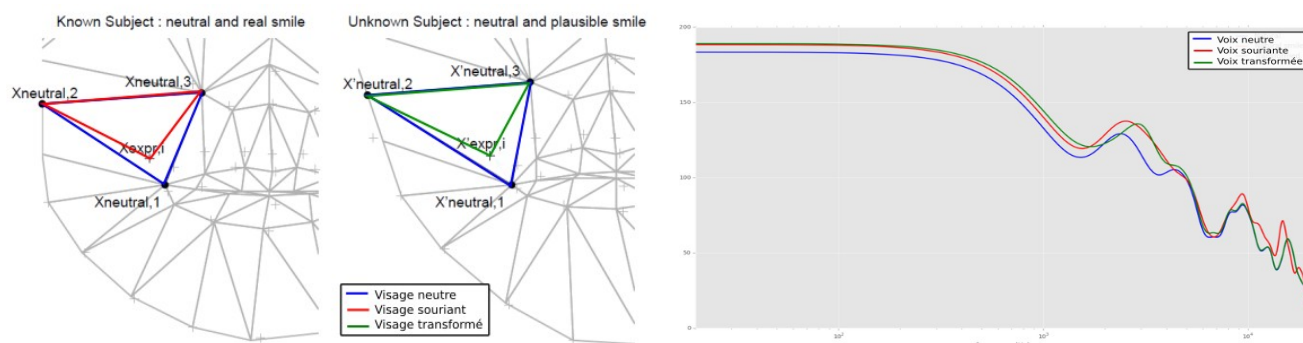
*réel* (tous mots-clé de la section7). Enfin, notre ambition clinique pour le projet s'inscrit en plein dans la volonté de l'INS2I de porter des nouvelles applications des sciences de l'information aux sciences du vivant.

Les **retombées attendues** de ce travail en interaction entre la modalité visuelle et sonore sont à la fois théoriques et applicatives. D'un point de vue **théorique**, ce travail constituera à notre connaissance le premier système temps-réel de manipulation multimodale des propres signaux d'un observateur, et permettra de concevoir un modèle unique de déformation de forme audiovisuelle à partir de points de contrôle (visuels ou sonores), qui étendra l'état de l'art spécifique à chacune des modalités. Cette avancée permettra en particulier d'élucider les constantes de temps nécessaires à la manipulation synchrone dans les deux modalités (existe-t'il une latence commune acceptable pour la voix et le visage ? Une manipulation doit-elle précéder l'autre ? Sur quels indices bas-niveau faut-il synchroniser les modalités?), un problème important dans le domaine des interfaces homme-machine et de la réalité virtuelle/augmentée (Oviatt, 1999). D'un point de vue **applicatif**, nous avons déjà fait la preuve de concept de l'efficacité de la rétroaction modifiée dans le seul domaine vocal pour l'induction d'émotion (Aucouturier et al, 2016), et l'effet psychologique de la rétroaction faciale est connu depuis les années 1970 – quoi qu'implémentée de façon très artisanale (par ex. un stylo coincé entre les dents pour forcer à sourire - Strack, Martin, Stepper, 1988). Notre système sera la première plate-forme algorithmique capable d'effectuer ces deux types de manipulation de façon conjointe, et nous nous attendons à de nombreuses applications dans le domaine de la recherche en psychologie/neuroscience. Pour catalyser ce potentiel d'application, la proposition prévoit l'organisation fin 2016 d'une journée d'étude autour du prototype construit dans le projet, et réunissant des cliniciens français autour de l'idée de rétroaction multimodale pour les pathologies de dépression (Pitié-Salpêtrière) et d'anxiété (Cochin, Service de Santé des Armées). Enfin, outre les applications cliniques ambitionnées par le projet, le système développé pourra également se prêter à d'autres applications dans le domaine des interfaces homme-machine. Par exemple, parce qu'il est temps-réel et réaliste, il pourra s'insérer dans une boucle vidéo de type *Skype* entre deux utilisateurs humains, ou permettre l'adaptation en temps-réel de l'expressivité d'avatars photo-réalistes dans des tutoriels intelligents ou des *serious games*.

### **Programme de recherche :**

**Verrou scientifique :** En cohérence avec le cadre (exploratoire et court-terme) du financement PEPS, le principal verrou scientifique de la proposition est celui de l'intégration de technologies déjà existantes chez les deux partenaires : d'une part, l'équipe-projet Cream de l'UMR9912 STMS développe depuis 2 ans dans le cadre d'un projet ERC une technique de manipulation émotionnelle de la voix, basée en partie sur le principe du *vocoder de phase* (voir ci-dessous), dans le but de construire une boucle de rétroaction sonore. D'autre part, l'équipe FAST de l'UMR6164 IETR travaille depuis 2 ans sur une technique de reconnaissance automatique d'expressions faciales qui utilise en première étape une technique de génération automatique d'expressions basique (sourire, peur ...) sur le flux vidéo de la personne observée, basée sur le principe du *piece-wise linear warping* (voir ci-dessous). Notre proposition est d'intégrer ces deux composantes dans une plate-forme commune, en résolvant les problèmes de synchronisation et d'interopérabilité des deux processus temps-réel, et d'assurer un haut degré de réalisme des deux types de transformation pour rendre le prototype de miroir crédible pour l'utilisateur.

**Méthodologies :** De façon indépendante, les deux technologies de manipulation émotionnelle proposées dans le domaine visuel par l'UMR6164 et sonore par l'UMR9912 ont convergé vers un principe de déformation (*warping*) paramétrique à partir de points de contrôle (Illustration2). Dans le domaine visuel, l'enveloppe convexe des visages est segmentée par triangulation, et la déformation de la forme du visage causée par une expression particulière (par ex. un sourire) est paramétrée de façon linéaire par rapport à des points de contrôle définis par cette triangulation. Cette approche, dite de *piece-wise linear warping*, permet à la fois d'apprendre à partir d'une base de visages neutres et expressifs un modèle de déformation indépendant de la morphologie d'un individu, et de l'appliquer à n'importe quelle morphologie au prix du simple calcul de la segmentation (Soladié, 2013). Dans le domaine sonore, l'enveloppe spectrale du signal est extraite dans une architecture de *vocoder de phase* (Liuni&Roebel, 2013), et la déformation de l'enveloppe causée par une expression particulière (par ex. le même sourire) est paramétrée par rapport à la fréquence et l'amplitude des formants (bosses) de la voix, qui servent de points de contrôle. La déformation est ensuite appliquée à n'importe quelle enveloppe spectrale de voix au prix du simple calcul des formants. La mise en interaction de ces deux processus permettra la conception d'un modèle unique de déformation de forme audiovisuelle à partir de points de contrôle, avancée théorique qui enrichira l'expertise de chacune des deux équipes.



*Illustration 2: Méthodologies de déformation (warping) à partir de points de contrôle dans la modalité visuelle (gauche) et sonore (droite). A gauche: la déformation du visage liée à une expression (ici, le sourire) est apprise sur une base de visages connus (neutres – bleu - et souriants -rouge), et paramétrée de façon linéaire par rapport à trois points de contrôle obtenus par triangulation du visage neutre. Cette déformation est ensuite appliquée en temps-réel sur les points de contrôles de visages inconnus, pour recréer une expression comparable (vert). A droite: même processus pour la déformation de l'enveloppe spectrale du signal vocal due à la parole souriante. Illustrations adaptées de Soladié et (2013) et Rachman et al. 2016.*

**Programme prévisionnel :** La collaboration sera structurée autour de la conception d'un prototype de miroir déformant intégrant les deux processus de déformation visuelle et sonore décrits ci-dessus. *Été/Automne 2016 :* Le projet permettra l'achat d'un dispositif matériel (PC+carte son+carte graphique dédié+écran+caméra), dupliqué dans les deux laboratoires, pour lequel sera développée une plate-forme logicielle commune (possiblement basée sur le langage Max/Mitter) intégrant les technologies des deux équipes. Deux réunions de travail (l'une à Paris, l'autre à Rennes) seront prévues pour coordonner ce travail. *Hiver 2016 :* Le projet permettra l'organisation d'une journée d'étude à Paris, réunissant des cliniciens français autour de l'idée d'applications cliniques du dispositif.

**Livrables :** Une plate-forme logicielle open-source implémentant le dispositif. Actes de la journée d'étude sur les applications cliniques de la rétroaction audiovisuelle modifiée.

### Cohérence du projet :

**Complémentarité :** la complémentarité des partenaires de la proposition est très lisible : il s'agit de construire une plate-forme multimodale visage et voix basée sur l'expertise de 2 équipes de traitement du signal, l'une spécialisée en visage (FAST) et l'autre en voix (CREAM). Cette collaboration est nouvelle – car il n'existe pas de précédent de collaboration entre les deux laboratoires, et originale – car les deux communautés de traitement visuel (VCIP, ICIP) et sonore (ICASSP, ISMIR) sont encore peu perméables. Elle est d'autre part réaliste, car de façon indépendante, les deux disciplines ont convergé vers un modèle identique de « déformation à partir de points de contrôle » qui n'attend qu'à être uniformisé. Enfin, les deux équipes ont la même ambition d'aller jusqu'à des applications cliniques, et le projet permettra ainsi de mettre en commun leurs différents partenaires cliniciens au cours d'une journée d'étude co-organisée à Paris à la fin du projet.

**L'équipe-projet CREAM de STMS UMR9912 (IRCAM/CNRS/UPMC),** financée initialement par le projet ERC StG CREAM (335536) porté par Jean-Julien Aucouturier, a pour but de développer des techniques de manipulation de stimuli vocaux et musicaux pour étudier comment le cerveau crée des émotions. A l'image du porteur du projet (PhD informatique à Paris 6, postdoctorat en neuroscience au Riken Brain Science Institute à Tokyo), l'équipe est composée pour moitié de traiteurs de signaux, et pour moitié de neuroscientifiques, et propose à la fois de nouveaux outils algorithmiques pour la communauté psychologie/neuroscience (Rachman et al, 2016) et des applications théoriques de ces outils dans le domaine des sciences du vivant (Aucouturier et al , 2016). L'équipe collabore de façon rapprochée avec l'Institut du Cerveau et de la Moelle Épinrière (ICM) de l'Hôpital Pitié-Salpêtrière à Paris, notamment pour la prise en charge de la dépression et de l'anxiété.

**L'équipe FAST de l'IETR UMR 6164 (CNRS/Centrale-Supélec)** travaille depuis une quinzaine d'années en analyse de visage. L'équipe a proposé de nouveaux algorithmes d'optimisation rapide des modèles déformables associés aux visages, en particulier pour en analyser les expressions. Elle a été coordinatrice du projet ANR Immemo qui a remporté deux challenges internationaux : la première place en détection de micro-expressions sur le visage


(FERA 2011) (Sénéchal, 2012) et les deux premières places en reconnaissance d'émotions (AVEC 2012) sur des signaux audio-vidéo (Soladié, 2012). L'équipe FAST a co-fondé deux startups : 3D Sound Labs et Dynamixyz. Cette dernière commercialise un outil de performance capture visant à transférer les expressions d'un comédien sur le visage d'un personnage 3D. Dynamixyz propose également des plugins temps réel donnant la possibilité d'animer des visages 3D parlants et expressifs. L'équipe travaille depuis quelques années sur la contagion d'émotion entre les personnes (Soladié, 2013), en collaboration avec le Service de Santé des Armées.

### **Budget demandé :**

Matériel	PC + Carte son + Carte Graphique + écran + Caméra (x2)	7000€
Missions	2 réunions de travail (Paris, Rennes)	3000€
Journée d'étude clinique	Invitation de cliniciens, démonstration	5000€
	<b>Total</b>	<b>15000€</b>

### **Avis et visa du directeur d'unité**

Cette collaboration entre deux UMR responsables de deux grands aspects de la perception multimodale est de nature à faire progresser la technique d'induction d'émotion sur les deux plans expressifs de la parole et des mimiques faciales avec des applications potentielles de grande importance pour les deux labos. Nous soutenons donc cette demande sans réserve.



**Ircam - CNRS**  
**UMR 9912 STMS**  
 3, Place Igor Stravinsky  
 F 75004 PARIS

Gérard Assayag

### **Références**

- Aucouturier, J.J., Johansson, P., Hall, L., Segnini, R., Mercadié, L. & Watanabe, K. (2016) Covert Digital Manipulation of Vocal Emotion Alter Speakers' Emotional State in a Congruent Direction. *Proceedings of the National Academy of Sciences*, doi: 10.1073/pnas.1506552113
- Fairbanks, G., & Guttman, N. (1958). Effects of delayed auditory feedback upon articulation. *Journal of Speech & Hearing Research*.
- Luini, M & Roebel, A. Phase vocoder and beyond. *Musica/Tecnologia*, [S.l.], p. 73-89, ago. 2013
- Oviatt, S. (1999). Ten myths of multimodal interaction. *Communications of the ACM*, 42(11), 74-81.
- Rachman, L., Liuni, M., Arias, P., Lind, A., Johansson, P., Hall, L., Richardson, D., Watanabe, K., Dubal, S. and Aucouturier, J.J. (2016) DAVID: An open-source platform for real-time emotional speech transformation. <http://biorxiv.org/content/early/2016/01/28/038133>
- Stoiber, N, Segulier, R. and Breton, G. Automatic design of a control interface for a synthetic face, *Proc. 13Th international conference on Intelligent user interfaces*, 2009, pp. 207-216
- Strack, S, Martin, L. and Stepper, S., Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis," *Journal of Personality and Social Psychology*, vol. 54, no. 5, pp. 768-777, 1988
- Toda, T., Muramatsu, T., & Banno, H. (2012, September). Implementation of Computationally Efficient Real-Time Voice Conversion. In *INTERSPEECH* (pp. 94-97).
- Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12), 1743-1759.
- Sénéchal, T., Rapp, V., Salam, H., Segulier, R., Bailly, K. and Prevost, L. *Facial Action Recognition Combining Heterogeneous Features via Multi-Kernel Learning*. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, Institute of Electrical and Electronics Engineers, 2012.
- Soladié, C., Salam, H., Pelachaud, C., Stoiber, N. and Séguier, R. Multimodal Fuzzy Inference System Using a Continuous Facial Expression Representation for Emotion Detection, *ACM International Conference on Multimodal Interaction (ICMI)*, 2nd International Audio/Visual Emotion Challenge and Workshop - AVEC 2012, Santa Monica, California, U.S.A., Oct. 2012.
- Soladié, C., Salam, H., Stoiber, N. and Séguier, R. Continuous Facial Expression Representation for Multimodal Emotion Detection. *International Journal of Advanced Computer Science*, 2013.