ARTUS : synthèse et tatouage audiovisuel des mouvements d'un personnage animé virtuel pour l'accessibilité d'émissions télévisuelles aux téléspectateurs sourds comprenant la Langue Française Parlée Complétée

Gérard Bailly[1*], Virginie Attina[1], Cléo Baras[7], Patrick Bas[2], Séverine Baudry[3], Denis Beautemps[1], Rémi Brun[4], Jean-Marc Chassery[2], Frank Davoine[5], Frédéric Elisei[1], Guillaume Gibert[1], Laurent Girin[1], Denis Grison[6], Jean-Pierre Léoni[6], Joël Liénard[2], Nicolas Moreau[7], Philippe Nguyen[3]

(1) Institut de la Communication Parlée, CNRS/INPG, 46, av. Félix Viallet Grenoble – France
(2) Laboratoire Image et Signaux, CNRS/INPG, BP. 46, 38402 Saint Martin d'Hères– France
(3) Nextamp, Les Lanthanides, bâtiment G2, 12, square du Chêne Germain, 35510 Cesson-Sévigné - France
(4) Attitude Studio, Bât. 126 - 50 avenue du Président Wilson , 93214 St Denis-la-Plaine – France
(5) Heudiasyc, CNRS/UTC, Centre de Recherches de Royallieu, BP 20529, 60205 Compiègne – France
(6) ARTE, 8, Rue Marceau, 92785 Issy-les-Moulineaux Cedex 9 – France
(7) ENST, 46, rue Barrault - 75013 Paris – France
* Adresser toute correspondance électronique à gerard.bailly@icp.inpg.fr

***Abstract*. The ARTUS Project aims at watermarking hand and face gestures of a virtual animated agent in a broadcasted audiovisual sequence. For deaf televiewers that master cued speech, the animated agent can be then incrusted - on demand and at the reception - in the original broadcast as an alternative to subtitling.**

I. INTRODUCTION

Le projet ARTUS [9] a pour objectif d'insérer dans les émissions télévisées des informations imperceptibles permettant d'animer à la réception un personnage animé virtuel codant la Langue Française Parlée Complétée (LFPC). Dans le cadre de ce projet, financé par le Réseau National des Télécommunications (RNRT), les mouvements de ce personnage sont calculés à partir d'un télétexte pré-existant, codés et insérés de manière indélébile et imperceptible dans l'émission audiovisuelle originale en utilisant des techniques de tatouage. Le clone ARTUS est incrusté à la demande et à la réception dans l'émission télévisuelle tatouée (cf. Fig. 1).

Alors que, dans le système télétexte actuel, l'information textuelle est transmise pendant les retours trames du balayage vidéo, cette procédure – par ailleurs gourmande en bande passante - est incompatible avec la généralisation d'une diffusion sous la forme de flux MPEG. En transmettant cette information (ou toute autre transformée : ici sa prononciation encodée par des gestes) par tatouage, elle peut suivre le chemin de diffusion, indépendamment du format de diffusion ou de stockage du signal (MPEG ou analogique) et sans traitement spécifique.

Cet article décrit les diverses composantes d'un premier système complet de doublage réalisé en décembre 2005 et en cours d'évaluation auprès d'un public de téléspectateurs sourds ayant un bon niveau de pratique de la LFPC.

Fig. 1 : Incrustation du clone ARTUS dans une émission d'ARTE.

Fig. 2. Codage des consonnes

Fig. 3. Codage des voyelles

II. LA LANGUE FRANÇAISE PARLEE COMPLETEE

Le Langage Parlé Complété (LPC, en anglais Cued Speech) est un langage gestuel inventé pour les sourds en 1967 par Orin Cornett aux Etats-Unis [11]. Il a été depuis adapté à plus de 50 langues [12]. Contrairement à la langue des signes, le LPC se greffe sur un langage parlé existant et permet, grâce à un segment du corps additionnel bien visible (la main) de transmettre des traits phonétiques articulés par des segments du corps peu ou pas visibles et donc inaccessibles à la lecture labiale seule (le larynx, le vélum et la langue notamment). Ce système de communication complétée rend donc les sons « plus visibles » et supplée l'insuffisance des indices phonétiques transmis par le seul mouvement des organes visibles du langage (lèvres) par l'ajout de clés de main pointant sur diverses parties du visage.

Facile à apprendre par les parents d'enfant sourd, il facilite la communication de l'enfant dans son environnement quotidien, son insertion dans le monde des entendants ainsi que l'apprentissage de la lecture et de l'écriture [17].

La structure syllabique Consonne-Voyelle étant largement majoritaire dans les langues du monde, le LPC encode conjointement une consonne et la voyelle subséquente au sein d'un même geste consistant à pointer avec une certaine clé de doigts, commune à un certain sous-groupe de consonnes (cf. Fig. 2), une partie du visage, position commune à un sous-groupe de voyelles (cf. Fig. 3). Les voyelles et consonnes isolées utilisent un codage par défaut du segment manquant. La version française du LPC, la LFPC comporte 8 clés et 5 positions de visage différentes.

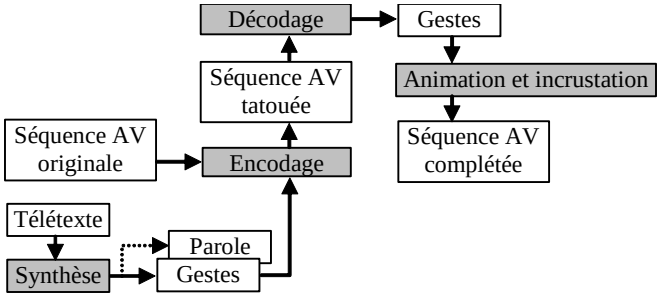

Fig. 4. Synopsis du système. Les composants grisés ont été développés spécifiquement pour cette application. Notre codeur virtuel est oraliste : le système de synthèse génère aussi la parole prononcée… non transmise pour l'instant par le système proposé en télédiffusion.

III. DESCRIPTION DU SYSTEME PROPOSE

Le système proposé (cf. Fig. 4) relève de multiples défis technologiques et scientifiques :

- Analyse des gestes de la LFPC produits par plusieurs codeurs et codeuses [2] et, plus particulièrement notre codeuse cible, dont les mouvements et leur synchronisation avec la production de parole ont été plus particulièrement étudiés [15].
- Transduction du télétexte en LFPC comprenant une adaptation importante d'un système de synthèse de parole multimodale à partir du texte COMPOST [3] aux spécificités du télétexte (absence de ponctuations, noms propres, etc) et au LPC (ajout de la modalité gestuelle au synthétiseur opérant par concaténation d'unités stockées, prosodie spécifique, etc…).
- Codage conjoint des gestes faciaux et des gestes LPC de manière à atteindre un débit de l'ordre de 200 bits par seconde. Un codage par quantification matricielle [similaire à 16] a été employée.
- Tatouage des flux vidéo et audio disponibles
- Animation vidéo-réaliste d'un clone de notre codeuse-cible incluant le développement de modèles 3D de forme et d'apparence du visage et de la main susceptibles d'être synthétisés et incrustés en temps-réel sur un téléviseur intelligent.

Les spécificités de ces divers modules sont données ci-après.

Fig. 5. Capture des mouvements du visage et de la main lors de l'énonciation de corpus lus. Les phrases, parfois difficiles à prononcer, avaient été préalablement enregistrées par une locutrice et étaient diffusées par haut-parleur. Ces consignes acoustiques évitent une situation trop proche de la lecture où le décodage du texte prend le pas sur l'expression.

A. Analyse des gestes LPC

Les gestes de la tête et de la main, les mouvements de la face et des doigts de notre codeuse oraliste cible ont été enregistrés avec un système VICON® à 12 caméras opérant à 120Hz (cf. Fig. 5), le son étant enregistré en synchronie. Des marqueurs demi-sphériques ont été collés sur le visage et la main de la codeuse (respectivement 63 et 50). Outre des gestes élémentaires, la codeuse a énoncé 239 phrases permettant de récupérer en moyenne deux versions de chaque diphone du français.

L'analyse de ces mouvements confirme les schémas de production mis en évidence par Attina et al [1, 2] :

- Déploiement de la clé de main pratiquement synchrone avec l'approche de la main du visage
- Atteinte de la cible vers le milieu acoustique de la consonne dans une séquence consonne-voyelle et au début acoustique du segment concerné dans le cas de segment isolé.

Par contre, notre codeuse non professionnelle mais ayant une longue et intense pratique de la LFPC dans son milieu familial a des mouvements de tête importants : si le bras assure l'essentiel du transport de la main au visage, la tête elle-même y participe de manière significative (en moyenne 7.7%). Cet emploi plus important des gestes posturaux a aussi été signalé chez les signeurs natifs [10].

B. Modélisation des gestes LPC

Une première réduction de dimensionnalité des gestes LPC est effectuée en réduisant les 3*113 degrés de liberté originaux à 9 composantes élémentaires pour représenter les gestes de main, 7 pour le visage, 12 pour le bras et la tête. Cette réduction effectuée par analyse en composantes principales sur les coordonnées brutes (visage) ou angulaires (main, bras, tête) permet en outre de régulariser les données de capture de mouvements, souvent incomplètes ou bruitées.

Fig. 6. Capture et modélisation des gestes de main.

C. Synthèse des gestes LPC

Plusieurs composants spécifiques ont été ajoutés au système de synthèse à partir du texte COMPOST [14].

Traitements linguistiques et marquage. Le télétexte étant privé de ponctuations de fin de phrase, un algorithme spécifique considère tous les débuts de tronçons de télétexte comme des débuts de phrase potentiels et élimine les candidats peu vraisemblables. Des marques de synchronisation égales aux « time-codes » de début d'affichage des tronçons correspondants sont alors insérées dans le texte produit. Le modèle rythmique décrit ci-dessous adapte la durée des pauses entre les phrases afin de respecter ces rendez-vous avec le contenu de l'émission.

Prosodie. Même si les codeurs arrivent à minimiser l'impact de l'ajout de la modalité gestuelle sur le débit de parole, le codage des consonnes isolées codées sur le côté, l'inertie du bras impose un rythme plus lent et une élocution plus hyperarticulée des syllabes complexes. L'intonation est aussi plus marquée. Un module prosodique spécifique a donc été entraîné sur les phrases du corpus en utilisant le système SFC développé par Bailly et Holm [5]. Dans les trois émissions que nous avons traitées, seules quatre phrases n'ont pu être prononcées dans le temps qui leur était imparti avec un retard moyen de 120ms. Notons d'ailleurs que les personnes en charge du marquage télétexte n'ont aucun moyen systématique de vérifier si les tronçons apparaissent suffisamment longtemps pour être lus (et donc prononcés…) !

Synthèse par unités stockées. La synthèse de parole et la génération des gestes de la main et du visage sont produites par sélection, concaténation et lissage de segments pré-stockés. Deux types de segments sont considérés : les « polysons » qui mémorisent le signal et les gestes faciaux allant d'une cible acoustique d'un allophone à la suivante (certains sons – notamment les glides - n'ont pas de cibles et sont donc emprisonnés dans un segment plus large) ; et les « diclés » qui mémorisent les gestes de main, les mouvements du bras et de la tête allant d'une cible de clé à la suivante. Pour la synthèse, les frontières des diclés sont alignées avec celles des polysons suivant les règles de synchronisation parole-gestes édictées à la section A. Les unités multi-représentées sont sélectionnées par une programmation dynamique utilisant des coûts de sélection et de concaténation spécifiques.

Génération. La génération du son est effectuée par TD-PSOLA et celle des gestes de la main et du visage est effectuée en étirant/compressant les gestes pré-stockés en préservant au mieux les parties transitoires. Un lissage anticipateur [4] est enfin effectué de manière à préserver la continuité des mouvements.

D. Codage des gestes LPC

Une deuxième réduction de dimensionnalité des gestes LPC est alors effectuée avant transmission afin d'atteindre le débit autorisé par les techniques de tatouage estimé à quelques centaines de bits par seconde. La technique utilisée est celle de la quantification matricielle. Le principe de ce codage est simple et consiste en un découpage des gestes synthétisés en blocs de 80 ms auxquels on associe un index dans une table de blocs représentatifs transmis à l'avance au récepteur. La métrique utilisée tient compte de la contribution de chaque paramètre à l'explication de la variance des données originales. Cette technique a été appliquée sur l'ensemble des gestes des segments pré-stockés. Ce choix peut poser problème dans une application d'interprétation temps-réel où les mouvements d'un codeur sont capturés et analysés à la volée. Dans le cas d'une synthèse, le codage peut œuvrer à l'intérieur d'un ensemble relativement restreint de réalisations connues à l'avance. Au final, le débit imposé par le tatouage, de l'ordre de 200 bits/s pour l'ensemble des paramètres à transmettre, a été respecté avec une qualité de codage satisfaisante.

E. Tatouage

Les techniques de tatouage permettent d'insérer à l'intérieur d'un signal (vidéo ou audio) une information additionnelle de manière indélébile et imperceptible (invisible en vidéo et inaudible en audio). Si cette information est souvent un numéro identifiant l'auteur pour les applications de protection de la propriété intellectuelle, elle peut aussi être un flux continu d'informations synchrones avec le document audiovisuel, e.g. les paramètres d'animation du clone ARTUS. L'application est alors dite de « contenu augmenté » ou de « canal caché ».

Tatouer un signal revient à transmettre de l'information (une séquence de bits) dans un canal numérique, aux principes largement étudiés mais aux conditions de transmission très particulières. Les principes sont les mêmes que lorsque, par exemple, l'on surfe sur Internet grâce à une connexion ADSL: rendre le débit d'information (le nombre de bits transmis par seconde) le plus élevé possible tout en maintenant le Taux d'Erreur Binaire (TEB) (le nombre de bits erronés sur le nombre de bits émis) le plus faible possible. Les conditions de transmission sont par contre très différentes : une ligne ADSL peut transmettre plusieurs Mbits/s avec des taux d'erreur très faibles, parce que le bruit de fond qui perturbe la ligne est relativement faible. Dans un contexte de tatouage, ce bruit n'est autre que l'image ou le son eux-mêmes. Ces signaux sont obligatoirement à des niveaux de puissance beaucoup

plus élevés que le signal utile, le tatouage qui véhicule l'information, et qui ont des caractéristiques qui varient à chaque instant dans des proportions qui peuvent être considérables. Il s'agit alors de réussir à transmettre une centaine de bits par seconde avec un TEB de l'ordre d'une erreur tous les 1000 bits émis.

145	100	103	105	137	133	110	120
128	113	142	105	138	100	147	133
112	143	130	119	139	148	108	101
133	127	95	72	80	68	74	67
130	129	71	75	71	69	92	86
126	129	94	82	85	98	65	99
119	129	69	75	98	76	66	79
141	129	69	70	77	81	71	83

Moyenne : 103,25

144	99	102	104	136	132	109	119
127	112	141	104	137	99	146	132
112	143	130	119	139	148	108	101
133	127	95	72	80	68	74	67
130	129	71	75	71	69	92	86
126	129	94	82	85	98	65	99
119	129	69	75	98	76	66	79
141	129	69	70	77	81	71	83

Moyenne : 103

Fig. 7. Codage d'un bit par la technique QIM sur un bloc d'image de 8x8 et un pas de quantification de 0.25. Ceci se fait simplement en retranchant une unité de luminance aux 16 premiers pixels du bloc.

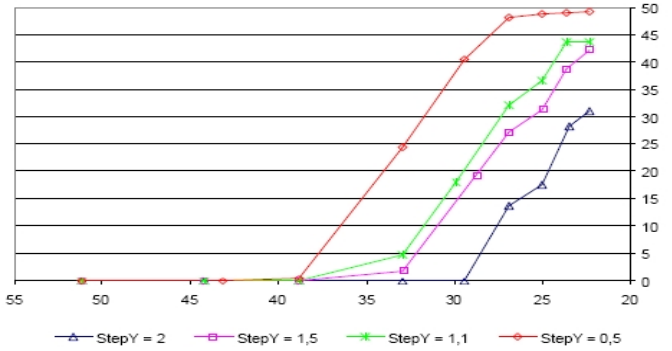

Fig. 8. Taux d'erreur au décodage en fonction du PSNR de l'image (plus le PSNR est important, plus la distorsion visuelle est faible) et pour divers pas de quantification. Taille des blocs: 32x32 pixels.

Tatouage vidéo

Une des techniques de tatouage adoptée dans le projet Artus est un tatouage substitutif [8]. La technique utilisée est une technique de quantification par modulation d'index (QIM), la plus populaire pour ce type de tatouage. De plus, un signal de synchronisation a été ajouté pour pouvoir décoder le tatouage même si la vidéo subit des transformations géométriques (changements 4/3 16/9, fenêtrage, …).. Les données sont codées dans l'image par quantification de la luminance moyenne de blocs de l'image (cf. Fig. 7). La taille choisie est de 32x32 pixels. Afin que le décodage puisse s'effectuer même si la vidéo subit une transformation valumétrique (changement de la luminance l de l'image par une fonction non-linéaire $f(l)$), nous avons adopté un pas de quantification flottant qui est proportionnel à la luminance des blocs voisins de l'image. Pour assurer une robustesse au bruit de transmission qui soit constante quelque soit le pas de quantification, nous avons choisi une grille de quantification fractale (cf. Fig. 9).

La robustesse du schéma de tatouage par QIM face aux transformations géométriques, à l'ajout de bruit et à la compression MPEG-2 a été également évaluée (cf. Fig. 8). Avec 432 bits/images, cela nous offre un débit de 1350 octets/sec, ce qui est largement suffisant pour transmettre les informations d'animation du clone et les codes correcteurs d'erreur.

L'implémentation de cette technique s'est effectuée en utilisant le logiciel K!TV qui permet une implémentation souple de traitements vidéo en temps réel sous windows.

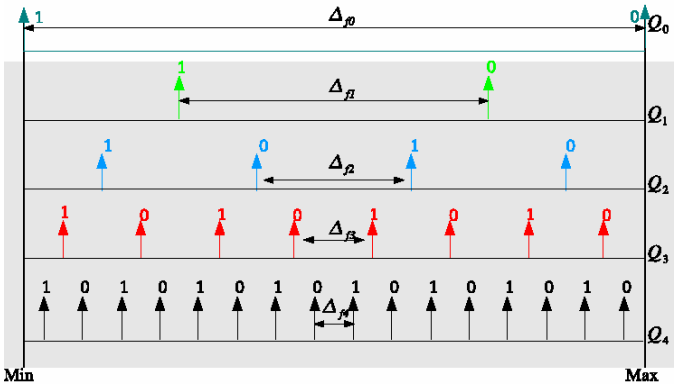

Fig. 9. Grille de quantification fractale permettant d'assurer une robustesse face aux transformations géométriques en tatouage vidéo [7].

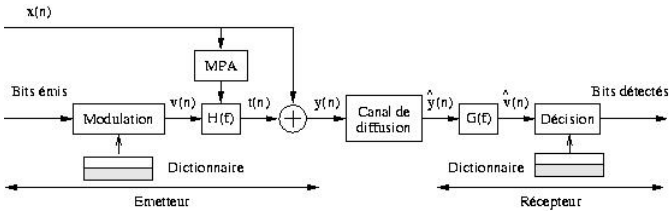

Fig. 10. Schéma de principe du système de tatouage audio.

Tatouage audio

Outre les enjeux en matière d'inaudibilité, de débit et de fiabilité de détection, le tatouage inséré doit de plus satisfaire à deux types de contraintes. La première est une contrainte de robustesse aux perturbations apportées par le transfert de la séquence audiovisuelle dans le réseau de diffusion ; le signal audio peut alors subir une opération de compression /reconstruction MPEG, des opérations de filtrage, de changement de formats (comme le passage d'une représentation numérique vers un signal analogique) ou des modifications de l'échelle des temps particulièrement perturbantes, etc. La seconde est liée à la complexité des algorithmes mis en jeu : le système doit être implantable en temps réel sur des machines actuelles, par exemple des PC standards.

Pour satisfaire ces contraintes, deux approches ont été envisagées dans cette étude. La première [18] détermine, pour chaque débit de transmission, un tatouage qui garantisse une transmission sans erreur. Le débit d'information admissible et le plus élevé possible est ensuite choisi en testant la qualité auditive du signal tatoué. La seconde [6] se propose de développer un système qui garantisse en premier lieu la contrainte d'inaudibilité. Différentes techniques sont ensuite mises en œuvre pour augmenter le débit d'information tout en minimisant le taux d'erreur binaire. Cette dernière, donnant lieu au schéma de tatouage Fig. 10, est présentée par la suite. Emettre une suite de bits (caractérisant les paramètres d'animation du clone après codage) requiert une opération de « modulation », qui assigne de façon bijective à chaque bit un signal représentable par un vecteur de dimension N

appartenant au « dictionnaire ». Le signal modulé v(n) est donc construit par concaténation de ces vecteurs à la cadence Fe/N définissant le débit. Pour rendre le signal de tatouage t(n) imperceptible, on exploite des résultats de psycho-acoustique, en particulier le phénomène de masquage : l'oreille accepte que l'on rajoute au signal original un autre signal sans déceler ce signal pourvu que celui-ci ait des propriétés fréquentielles particulières [20]. La détermination d'un « seuil de masquage » (extrait d'un modèle psycho-acoustique classiquement utilisé en compression des signaux audio) donne la limite supérieure de la puissance d'un tatouage inaudible pour l'oreille en présence du signal original pour chacune des fréquences audibles (20 Hz - 20 kHz). Ce seuil de masquage est pris en compte sous la forme d'un filtre H(f) qui permet de « mettre en forme spectralement » le signal modulé. Le tatouage qui en résulte est ensuite directement additionné au signal audio original x(n) pour fournir le signal tatoué y(n), qui sera transmis dans le canal de diffusion télévisuel. Au récepteur, la théorie des communications numériques indique qu'une stratégie raisonnable consiste à d'abord rendre le signal en sortie du filtre noté G(f) sur la figure le plus ressemblant possible au signal v(n) en jouant sur les caractéristiques de ce filtre [19]. Il ne reste plus qu'à réaliser l'opération de « détection », c'est-à-dire à se demander tous les N échantillons quel est le vecteur reçu le plus ressemblant aux vecteurs du « dictionnaire» utilisé pour construire le signal modulé. Des simulations sur des signaux audio variés (différentes voix et divers types de musique) montrent qu'à un débit de quelques centaines de bit/s, on obtient un TEB de l'ordre de 0.001 en assurant l'inaudibilité du tatouage.

La robustesse du système aux perturbations apportées par le réseau de diffusion télévisuel a également été considérée. La compression MPEG se révèle peu contraignante si l'on apporte quelques aménagements au dispositif (choix des vecteurs du dictionnaire à bande limitée). Beaucoup plus gênantes sont les perturbations désynchronisantes (lié à la modification de l'échelle des temps). Le mécanisme suivant a donc été adopté : toutes les secondes approximativement, le signal de tatouage est précédé par un signal prédéfini de synchronisation de durée suffisamment longue pour être facilement détectable au récepteur (existence d'un pic de corrélation) mais pas trop pour éviter d'être trop pénalisante en terme de débit. La mesure du nombre d'échantillons entre les différents pics de corrélation permet d'en déduire la position des bits dans le signal reçu.

Deux programmes distincts, le premier construisant le signal tatoué à l'émetteur et le second détectant les bits au récepteur avec tout son mécanisme de synchronisation, ont été écrits (en langage C). Ils fonctionnent tous les deux en temps réel (exploitation du logiciel libre PortAudio).

F. Animation vidéoréaliste

Divers modèles de la forme articulée 3D du haut du tronc et de la main et de leur apparence ont été réalisés soit en partant de données photogrammétriques de la codeuse soit en adaptant un modèle générique à sa morphologie (cf. Fig. 12). A partir de moulage en plâtre, un modèle de forme de la main spécifique a été développé par la technique de skinning » (cf. Fig. 11).

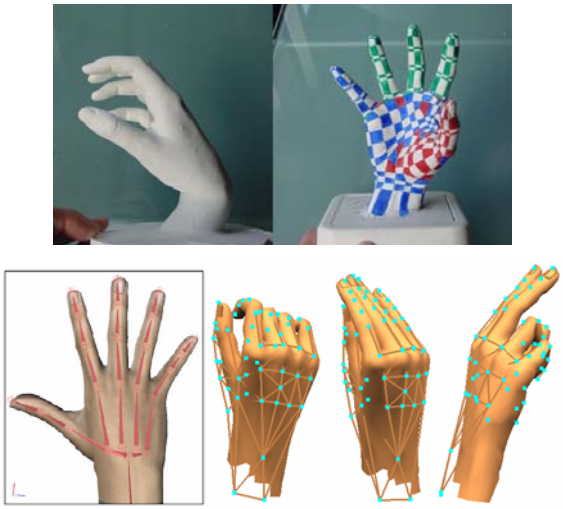

Fig. 11. En haut : Moulage de la main de la codeuse et définition du maillage de référence. En bas : Ossature, résultat après « skinning » et contrôle par le modèle de forme construit à partir des données de capture de mouvement.

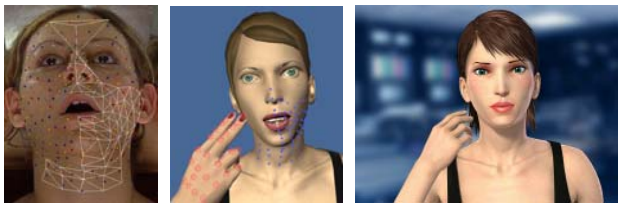

Fig. 12. Pilotage d'un modèle de surface vidéo-réaliste par le modèle de forme issu des données de capture de mouvement. A gauche, projection du modèle de forme sur une photo de la codeuse. A droite, deux alternatives à la représentation choisie dans la Fig. 1, utilisant une adaptation d'avatars sont en cours d'évaluation.

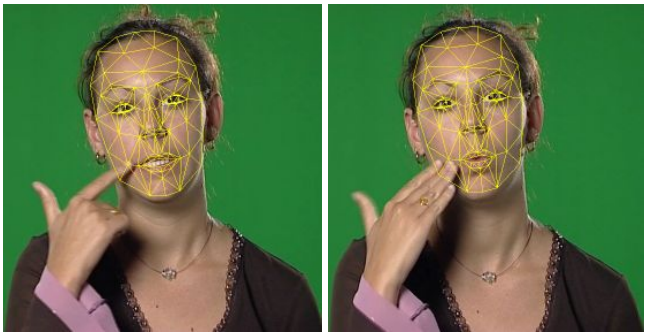

Fig. 13. Suivi des mouvements tridimensionnels de la tête et des gestes faciaux d'un interprète (ici d'une codeuse) de la LFPC.

IV. SUIVI DES GESTES D'UN(E) INTERPRETE DE LA LFPC

Un des objectifs du projet ARTUS est également, dans un cadre plus prospectif, de proposer une méthode automatique permettant de suivre les mouvements tridimensionnels de la tête et les gestes du visage non maquillé d'un(e) interprète de manière à tatouer une émission en temps-réel. Des travaux ont été menés dans ce sens, exploitant un modèle adaptatif de l'apparence faciale du codeur. La robustesse des méthodes de suivi aux occultations locales du visage ainsi qu'aux fortes rotations hors plan est augmentée par l'utilisation de techniques de filtrage stochastique [13]. Le suivi se fait

aujourd'hui à partir d'un code C/C++ non optimisé à une cadence moyenne de 4,5 images par secondes sur un PC équipé d'un processeur Intel cadencé à 3,6 GHz. Les gestes considérés sont ceux des sourcils et des lèvres. Le suivi des gestes de main suivant la même méthode est en cours d'étude.

V. CONCLUSIONS

ARTUS propose un nouveau service télévisuel à l'intention des malentendants qui consiste à adjoindre à des documents télévisuels un ensemble d'informations indélébiles et imperceptibles utiles à leur compréhension. Un premier démonstrateur utilisant les divers composants décrits dans cet article a été développé et est en cours d'évaluation tant du point de vue de l'intelligibilité de la codeuse virtuelle que de la compréhension globale de l'émission permise par ce système en comparaison avec le télétexte standard. Un examen plus poussé de la charge cognitive induite par ces deux systèmes d'aide à la compréhension doit être conduit de manière à évaluer l'impact sur l'attention des téléspectateurs ciblés, lors de longues expositions à des documents audiovisuels augmentés.

Ce système et l'ensemble des techniques développées peuvent être appliqués à d'autres systèmes de communication langagière (langue des signes) ou non (insertion d'icônes ou fonctions de mise en transparence de parties de l'image).

Notons finalement que ce système peut être étendu à d'autres situations (codage temps-réel d'émissions, etc.), à d'autres supports (CDROMS interactifs, etc.) et à d'autres applications de loisirs numériques (apprentissage de la LFPC, etc.).

REMERCIEMENTS

Nous tenons à remercier notre codeuse Yasmine Badsi pour s'être prêtée à ces nombreuses et importantes collectes de données. Nos remerciements s'adressent aussi à Ilarion Pavel du MENRT et Jean-Jacques Rigoni d'ElanSoft qui ont soutenu ce projet.

REFERENCES

[1] Attina, V. (2005) *La Langue Française Parlée Complétée : production et perception*. PhD Thesis. Institut National Polytechnique: Grenoble - France.

[2] Attina, V., Beautemps, D., Cathiard, M.-A., and Odisio, M. (2004) *A pilot study of temporal organization in Cued Speech production of French syllables: rules for a Cued Speech synthesizer*. Speech Communication, **44**: p.197-214.

[3] Bailly, G. and Alissali, M. (1992) *COMPOST: a server for multilingual text-to-speech system.* Traitement du Signal, **9**(4): p.359-366.

[4] Bailly, G., Gibert, G., and Odisio, M. (2002) *Evaluation of movement generation systems using the point-light technique*. in *IEEE Workshop on Speech Synthesis*. Santa Monica, CA. p.27-30.

[5] Bailly, G. and Holm, B. (2005) *SFC: a trainable prosodic model.* Speech Communication, **46**(3-4): p.348-364.

[6] Baras, C. (2005) *Tatouage informé de signaux audio numériques*. PhD thesis. Ecole Nationale Supérieure des Télécommunications: Paris.

[7] Bas, P. (2005) *A quantization watermarking technique robust to linear and non-linear valumetric distortions using a fractal set of floating quantizers*. in *Information Hiding Workshop 2005*. Barcelonna, Spain

[8] Bas, P., Chassery, J.-M., and Macq, B. (2002) *Image Watermarking: an evolution to content based approaches.* Pattern Recognition, **35**(3): p.545-561.

[9] Bas, P., Lienard, J., Chassery, J.-M., Beautemps, D., and Bailly, G. (2003) *Artus: animation réaliste par tatouage audiovisuel à l'usage des sourds*. in *Journée Nationale sur « Image et Signal pour le Handicap »*. Paris

[10] Boyes Braem, P. (1999) *Rhythmic temporal patterns in the signing of early and late learners of German Swiss Sign Language.* Language and Speech, **42**: p.177-208.

[11] Cornett, R.O. (1967) *Cued Speech.* American Annals of the Deaf, **112**: p.3-13.

[12] Cornett, R.O. and Daisey, M.E. (1992) *The cued speech resource book for parents of deaf children*. Raleigh, NC: The National Cued Speech Association, Inc.

[13] Dornaika, F. and Davoine, F. (2005) *Simultaneous facial action tracking and expression recognition using a particle filter*. in *IEEE International Conference on Computer Vision*. Beijing, China. p.1733-1738.

[14] Gibert, G. (2006) *Mise en oeuvre d'un synthétiseur 3D de Langage Parlé Complété*. Institut National Polytechnique: Grenoble.

[15] Gibert, G., Bailly, G., Beautemps, D., Elisei, F., and Brun, R. (2005) *Analysis and synthesis of the 3D movements of the head, face and hand of a speaker using cued speech.* Journal of Acoustical Society of America, **118**(2): p.1144-1153.

[16] Girin, L. (2004) *Joint matrix quantization of face parameters and LPC coefficients for low bit rate audiovisual speech coding.* IEEE Transactions on Speech and Audio Processing, **12**(3): p.265-276.

[17] Leybaert, J. (2003) *The role of Cued Speech in language processing by deaf children: an overview*. in *Auditory-Visual Speech Processing*. St Jorioz - France. p.179-186.

[18] LoboGuerrero, A. (2004) *Etude de techniques de tatouage audio pour la transmission de données*. PhD Thesis. Institut National Polytechnique: Grenoble - France.

[19] Proakis, J. (2001) *Digital communications*. 4th edition ed. New York: McGraw-Hill.

[20] Zwicker, E. and Feldtkeller, E. (1981) *Psycho-acoustique, l'oreille récepteur d'information*. Collection technique et scientifique des télécommunications. Paris: Masson.