LISTEN et visiophonie personne libre

Jean Emmanuel Viallet

LISTEN et visiophonie personne libre

J. E.Viallet, O. Bernier, M. Collobert, R. Féraud, D. Collobert,
A. Gilloire*, Y. Mahieux* et G. Le Tourneur*
DTL/DLI/TNT, DIH/CMC*
France Télécom, Centre National d'Études des Télécommunications
Technopole Anticipa, 2 avenue Pierre Marzin, BP 40,
22307 Lannion Cedex, France
email : viallet@lannion.cnet.fr

Résumé

La fonction visiophonique "personne libre" et le terminal de démonstration LISTEN, développés au CNET à Lannion, s'appuient sur une détection de visage pour contrôler automatiquement une caméra motorisée et une antenne acoustique. La prise de vue et la prise de son suivent en permanence la personne qui est ainsi libre de se déplacer.

1. Introduction : le visiophone, une technologie actuelle contraignante.

Les visiophones actuels se caractérisent par une prise de vue et une prise de son fixes.

La personne doit d'abord se rendre à son terminal avant de démarrer une communication ce qui réduit les occasions de communication.

La personne doit ensuite se placer dans le champ de la caméra et ne plus en sortir. Ce champ fixe restreint la gestuelle. Les interlocuteurs sont figés et la communication perd en naturel.

Une prise de son par un téléphone filaire introduit d'autres contraintes : elle lie physiquement la personne à son terminal. Un téléphone mobile tenu à la main limite la gestuelle. Ces inconvénients peuvent être réduits par le port d'un casque réunissant les fonctions de microphone et de haut parleur. Le casque réduit le champ sonore perçu par son porteur et le microphone celui perçu par son interlocuteur.

2. LISTEN : Une communication médiatisée inspirée de la communication en vis à vis

Nous analysons brièvement la communication en vis à vis afin de nous en inspirer pour améliorer la fonction visiophonique.

L'image participe pleinement à la communication interpersonnelle. La dimension non verbale de la communication fournit des indices de plusieurs natures [1] :

- des indices cognitifs qui complètent le contenu du message. Par un geste, un regard une personne signifie son accord ou son incompréhension.

- des indices affectifs et sociaux qui renseignent sur l'état d'esprit et sur les réactions affectives du correspondant.

- des indices agissant sur l'établissement du tour de parole. Par un geste, un regard une personne demande ou refuse la parole.

Les aspects visuel et langagier de la communication sont étroitement associés. La plupart des gestes sont produits lorsqu'une personne parle [2]. Gestuelle et parole ne doivent pas être dissociés : la fréquence et l'amplitude des gestes traduisent l'appartenance culturelle.

La nature et le contenu de la communication sont fortement influencés par le contexte. Dans un contexte contraignant de communication comme celui des visiophones classiques, l'individu consacre une part notable de ses ressources à se plier à l'environnement au détriment de la communication : les interlocuteurs doivent souvent vérifier qu'ils ont bien été entendus et compris, qu'ils sont cadrés et vus.
Afin que les interlocuteurs puissent se consacrer au contenu même de l'échange, il est souhaitable que la technologie ne fasse pas écran à la communication .

Lors d'une communication en vis à vis, l'oeil et l'oreille s'orientent naturellement vers la source d'information. Cette source d'information est constituée du visage, principal vecteur des informations visuelles et acoustiques de la communication et aussi du reste du corps (gestes et posture).

L'image joue donc un rôle essentiel une fois la communication établie. Mais il a d'abord fallu prendre conscience de la présence de l'autre puis établir un contact visuel ou sonore avant de pouvoir communiquer en vis à vis.

La présence peut être permanente. Elle favorise les occasions de communication mais risque également d'être perçue comme une violation d'un espace privé. Une présence visuelle permanente en champ large (comme quelqu'un vu de loin, au bout du couloir) respecterait mieux le droit à l'image et assurerait aussi la confidentialité des échanges sonores avant que la communication médiatisée ne soit établie.

3. Localisation et suivi de la personne, source d'information.

Pour retrouver la situation de communication en vis à vis, il faut que les capteurs de prise de vue et de prise de son suivent constamment la personne lorsqu'elle se déplace. Le problème est donc d'abord de détecter et de localiser rapidement une personne.

Une personne et son cadre de travail forment une scène visuelle et sonore a priori complexe. Cependant, cette personne est toujours caractérisée par son image même lorsqu'elle ne parle pas. S'adressant à son interlocuteur, le visage de l'utilisateur est plus ou moins de profil.

La détection acoustique est complexe et sujette à erreur (sources et chemins acoustiques multiples dues aux réflexions, ...)[3].

C'est pourquoi, dans LISTEN, nous privilégions la détection visuelle.

La plupart des techniques de détection de visages ne fonctionnent qu'avec un visage de face en gros plan [4]. Le détecteur de visage par réseau de neurones que nous avons développé repère un visage distant et de profil (son image peut être aussi petite que 20*15 pixels de large, ce qui correspond à un visage situé à 3,8 mètres (pour une image CIF et un champ de 60°). Ce détecteur possède des performances [5], tant en terme de taux de détection que de taux de fausses alarmes, comparables à l'état de l'art.

Cependant, il n'est pas possible d'appliquer le réseau de neurones, en temps réel, à toute l'image CIF traitée, sans informations a priori. Nous avons donc développé des modules de détection de mouvement et de teintes chairs [6] pour identifier des régions d'intérêt où sont susceptibles de se trouver des visages. Ces régions correspondent principalement aux couples visages/cous et mains/bras.

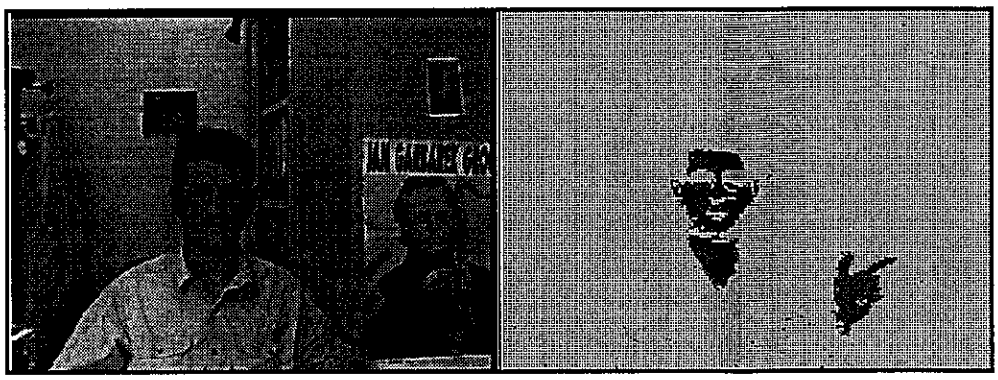

Figure 1 : Détection de la teinte chair : image d'origine et image après détection.

Plusieurs personnes pouvant être présentes, nous avons choisi de détecter et de suivre la personne dont l'image du visage est la plus grande. Une heuristique permet de scruter une région d'intérêt, afin d'y détecter un visage, en typiquement 0,25 seconde.

Après détection, la région d'intérêt correspondant au visage est suivie en permanence. LISTEN peut démarrer en champ large puis faire un plan plus serré sur la personne détectée. Des développements sont en cours pour suivre en parallèle plusieurs personnes et traiter la situation où deux personnes se croisent (zones recouvertes et découvertes).

4. Prises de vue et de son automatiques : incidence sur la communication

Une caméra motorisée et une antenne acoustique se comportent comme un oeil et une oreille électroniques. LISTEN [7] permet de les orienter automatiquement vers le visage détecté.

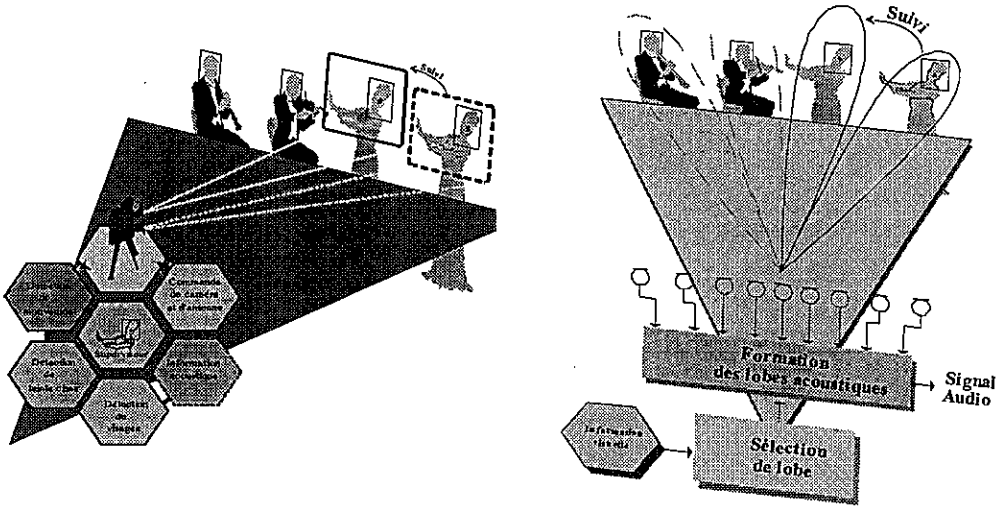

Figure 2 : Principe de fonctionnement : prise de vue et prise de son.

Un champ large initial permet d'obtenir un compromis entre une présence visuelle permanente, destinée à à favoriser les occasions de communication, et une définition réduite. L'image est ainsi suffisamment neutre pour respecter l'environnement de la personne.

La personne est libre de ses mouvements et de ses déplacements et de tout équipement portatif (casque, localisateur). La personne dialogue alors avec l'image de son interlocuteur et n'est plus obligée de s'adresser à la caméra ou au microphone, comme en présence de capteurs fixes. La communication gagne en naturel : LISTEN est un visiophone "personne libre".

En contrôlant automatiquement la focale de la caméra, le système permet de restituer pleinement la dimension gestuelle qui accompagne la parole de la personne qu'elle soit proche ou éloignée de la caméra (Figure 3). L'image du visage demeure à peu près de la même taille. Ainsi, une même distance apparente est conservée entre les interlocuteurs.

La commande simultanée de la caméra en site et en azimut autorise des mouvements de caméra souples, sans à-coups, confortables du point de vue de l'observateur. LISTEN fonctionne actuellement sans compression de l'image et du son. Remarquons que c'est la même caméra qui assure la prise de vue destinée d'une part à l'analyse de l'image et d'autre part à la personne distante. L'influence d'une compression avec compensation du mouvement sera prochainement évaluée.

Le système connaît en permanence la position du visage. Cette information est communiquée à l'antenne acoustique. L'antenne acoustique est formée de neufs microphones unidirectionnels [8]. Par réglage électronique des retards sur chaque microphone, le lobe acoustique de l'antenne peut être orientée immédiatement dans l'une des treize directions uniformément réparties du demi plan horizontal situé devant l'antenne.

La directivité du lobe acoustique est comprise entre 20 et 30 degrés à la fréquence de 1 kHz. Cette directivité autorise une prise de son centrée sur la personne et qui atténue les sons en provenance de sources situées à l'extérieur du lobe acoustique.

L'antenne acoustique permet une prise de son sur une profondeur importante, même lorsque l'utilisateur est à quelques mètres de l'antenne (bien que l'antenne, de 40 cm de large, ait été conçue pour un terminal individuel).

L'observateur perçoit correctement le signal de parole ainsi que les variations d'intensité associées aux déplacements d'avant en arrière du participant distant. On constate que le niveau subjectif de perception de la parole semble supérieur à l'amplitude physique (inversement proportionnelle au carré de la distance). Ce phénomène peut s'expliquer par le fait que l'observateur compense partiellement la baisse du signal en voyant l'utilisateur s'éloigner. Ainsi, l'image et le son sont subjectivement perçus comme cohérents.

Figure 3 : La prise de vue réalisée par LISTEN au cours d'une communication visiophonique " personne libre ".

5. Conclusion

Les techniques présentées s'inspirent de la communication en vis à vis : la prise de vue et la prise de son s'orientent automatiquement vers la source d'information. .

Ces techniques sont en cours d'adaptation pour la visioconférence VARÈSE (projet CNET) dans lequel la restitution de l'image et du son s'inspirent de la situation en vis à vis : les interlocuteurs sont vus à l'échelle 1 sur un écran géant et le son est spatialisé.

Couplées avec la reconnaissance de la parole, les techniques de suivi, utilisées par LISTEN, permettront le développement d'interfaces homme-machine innovantes.

6. Références

[1] S. Whittaker, " Rethinking video as a technology for interpersonal communications : theory and design implications ", *Int. J. Human-Computer Studies*, 42, pp 501-529, 1995.

[2] J. Cassell, "What You Need to Know about Natural Gesture", Keynote Address *Second International Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, USA, 14-16 Octobre 1996,

[3] U. Bub, M. Hunke, A. Waibel, "Knowing who to listen to in speech recognition : visually guided beamforming", *Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing Conference*, Detroit, Michigan, 1995.

[4] R. Chellappa, C. L. Wilson, S. Sirohey, "Human and Machine Recognition of Faces : A Survey", *Proceedings of the IEEE*, Vol 2615, pp89-98, 1995.

[5] R. Feraud, "A Modular face detection system", *Proceedings of NEURAP Conference*, Marseille, Mars 1997.

[6] M. Hunke, A. Waibel, "Face Locating and Tracking for Human Computer Interaction", *Proceedings of the 28th Asimolar Conference on Signals, Systems, and Computers*, Pacific Grove, California, 1994.

[7] M. Collobert, R. Féraud, G. Le Tourneur, O. Bernier, J.E. Viallet, Y. Mahieux and D. Collobert, "LISTEN : a System for Locating and Tracking Individual Speakers", *Proceedings of Second International Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, USA, 14-16 Octobre 1996, pp 283-288.

[8] Y. Mahieux, G. Le Tourneur, A. Saliou, "A microphone array for Multimedia Workstations", *Journal of the AES*, Vol. 44, n° 5, pp 365-372, 1996.