Interaction multimodale : Faites un geste, parlez lui!

Jean Emmanuel Viallet

Orange Labs, France Télécom Recherche & Développement

Travaux effectués avec Sébastien Carbini

sebastien.carbini@ifremer.fr

Travail soutenu par le Network of Excellence SIMILAR financé par l'UE.

Introduction

•Définition

–Multimodalité : combinaison de plusieurs modalités de communication (geste, parole, écriture, ...)

•Intérêt de la multimodalité en Interface Homme-Machine (IHM)

–Interactions homme-machine plus transparentes, plus flexibles, plus efficaces et plus expressives [Oviatt 2002]

•Efficacité de la multimodalité → Kirusa et France Télécom R&D Boston
Périphérique PDA : stylet + voix
Tâche : trouver un hôtel sur les Champs Elysées avec les Pages Jaunes

Stylet seul : 50 s
Voix seule : 2 min
Multimodalité : 30 s
→ **Multimodalité + rapide**

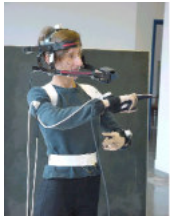
Contexte grand écran : geste + parole

Grand écran

Visualisation	Interaction classique
à distance	contact
plusieurs personnes	une seule personne
avec déplacements	sans déplacement

•**But** :
Une interface homme-machine multimodale pour interagir à plusieurs par le geste et la parole avec un grand écran.

– Geste et parole → interaction à distance
– "Personne libre" → positions libres et gestes non contraints



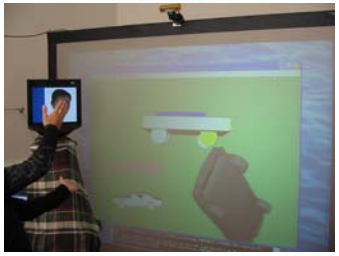
•**Problèmatique** :

– Détecter et suivre les utilisateurs et leurs parties du corps
– Quels gestes pour une IHM ? Comment interpréter ces gestes ?
– Synchroniser et fusionner le geste et la parole avec le contexte applicatif

Contexte

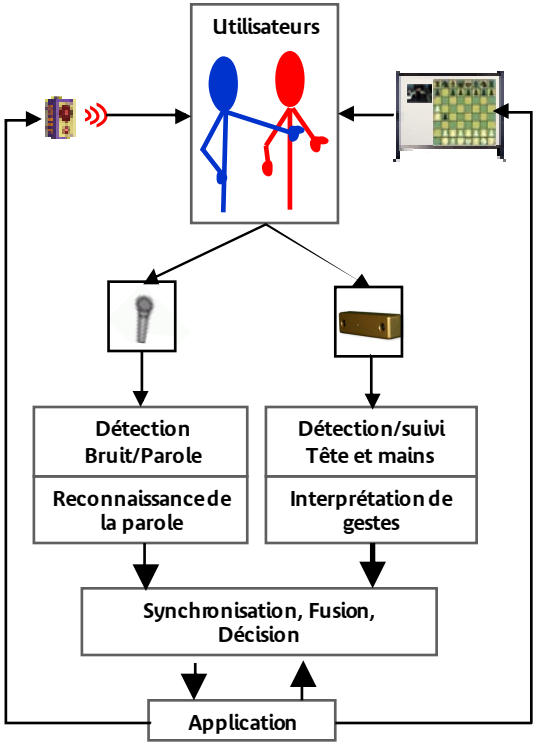
France Télécom R&D → PRP Multimodalité

Projets européens



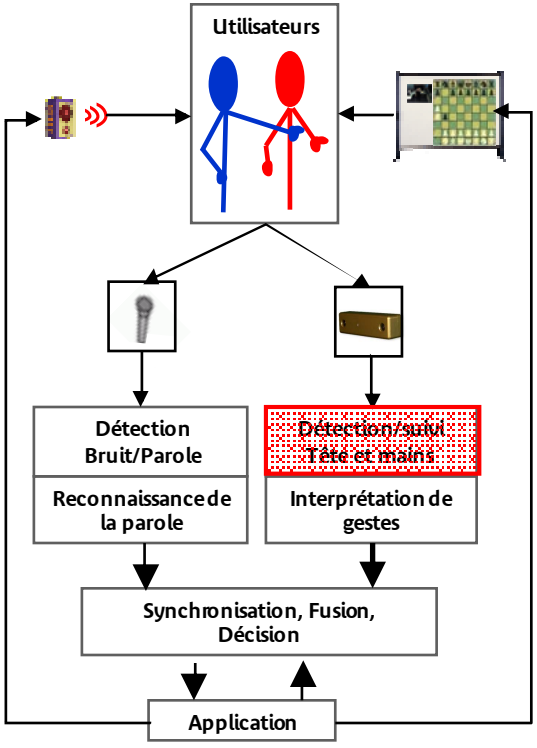
Plan

- Introduction
- Détection des gestes
- Interprétation de gestes
- Reconnaissance de la parole
- Synchronisation des modalités
- Evaluation de la souris oro-gestuelle
- Fusion des modalités avec le contexte
- Conclusion

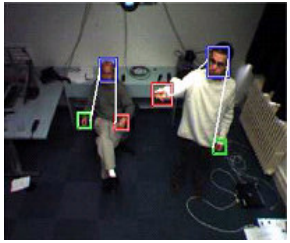


Plan

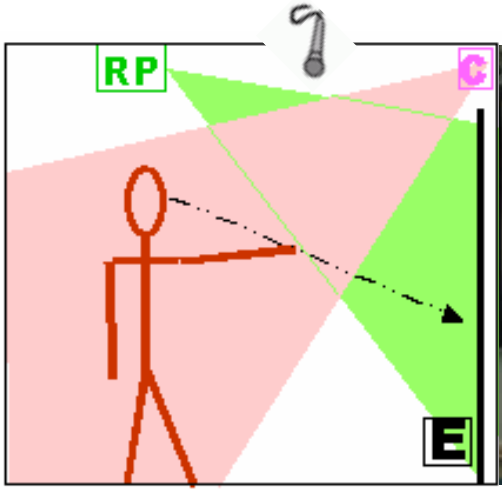
- Introduction
- Détection des gestes
- Interprétation de gestes
- Reconnaissance de la parole
- Synchronisation des modalités
- Evaluation de la souris oro-gestuelle
- Fusion des modalités avec le contexte
- Conclusion



Détection des gestes : Dispositif expérimental



- Caméra **stéréo** Bumblebee **(C)**
  - Champ 70° assez large pour deux personnes
- Microphone omnidirectionnel ou HF
- Utilisateur devant une image retro-projetée (2,1x 1,8m) **(E)**
- Rétro-projecteur courte focale **(RP)**



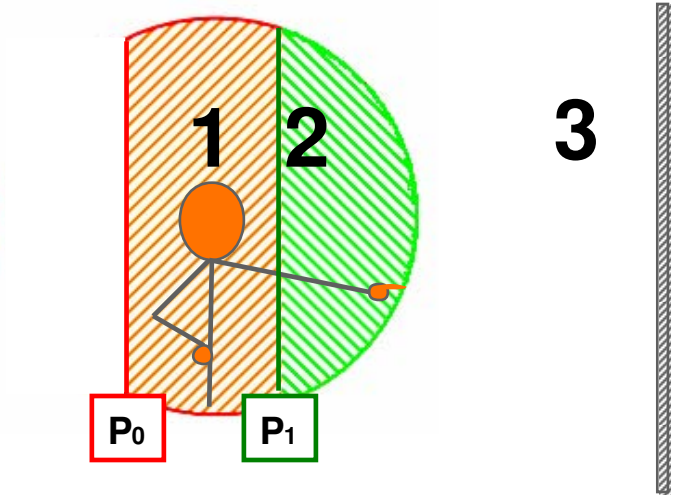
Détection des gestes : Détection des parties du corps 2/3

1. Détection de la tête par réseau de neurones

orienté vers les zones chair en mouvement de distance connue

2. Détection des mains

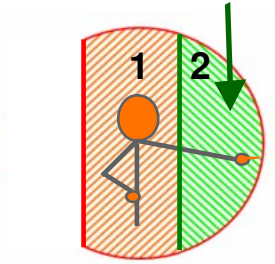
- 1 : **La zone de repos**
  - permet de ne pas interagir en permanence avec le système
- 2 : **La zone d'action**
  - permet à l'utilisateur de traduire son intention d'interagir
- 3 : **La zone de non-détection**



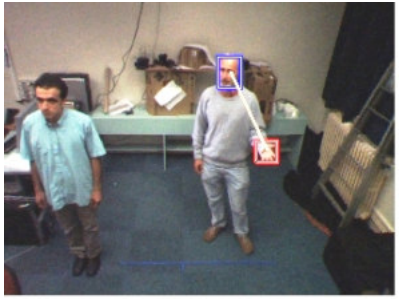
Détection des gestes : Détection des parties du corps 3/3

3. Détection de la main

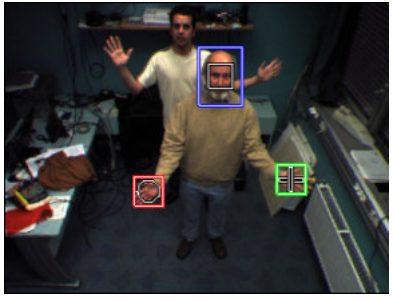
- Difficulté: forme variable, faible résolution
- Hypothèses
  - Zone de teinte-chair en mouvement dans la zone d'action (2)
  - 1ère main avancée → main de pointage
  - 2ème main → main de commande



Erreur possible

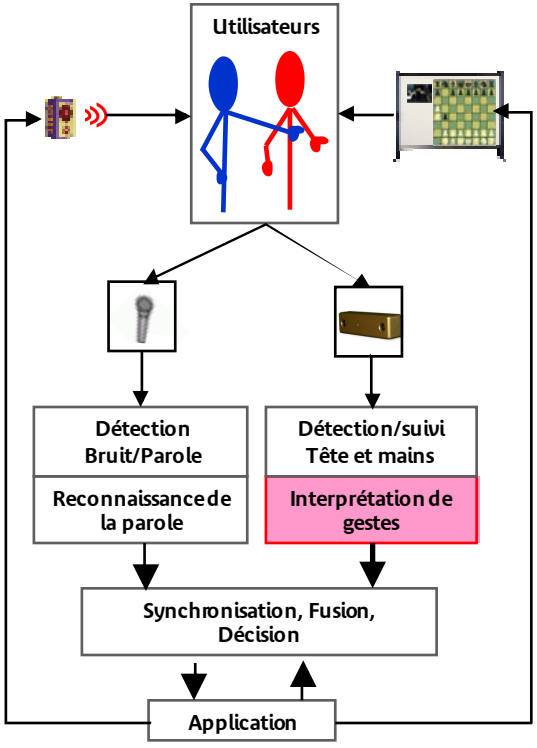


Sans erreur



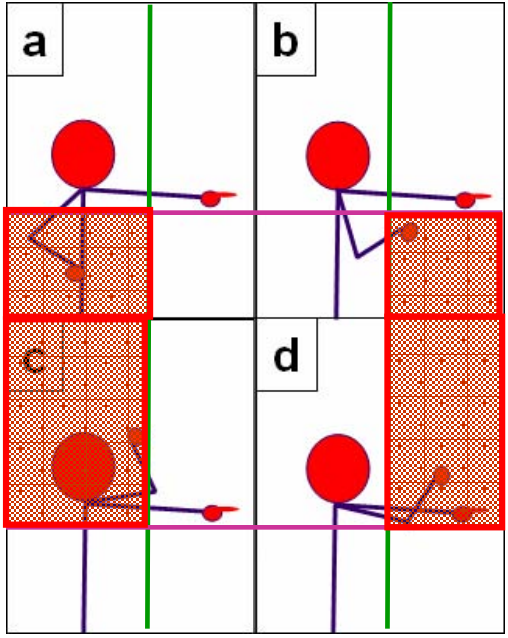
Plan

- **Introduction**
- Détection des gestes
- Interprétation de gestes
- Reconnaissance de la parole
- Synchronisation des modalités
- Evaluation de la souris oro-gestuelle
- Fusion des modalités avec le contexte
- **Conclusion**

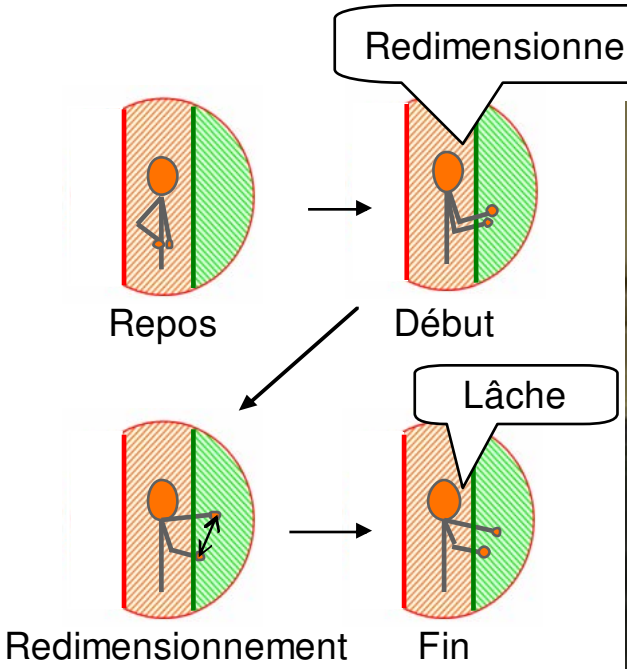


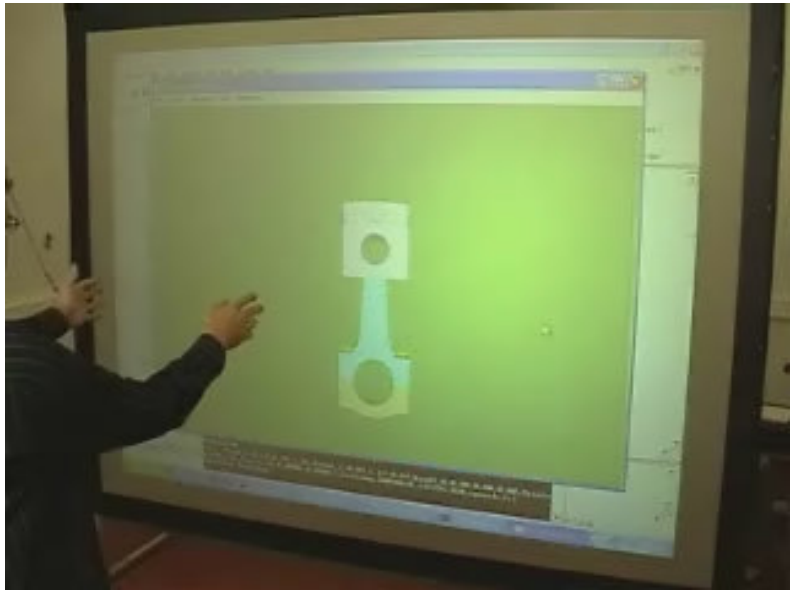
Interprétation de geste :pointage/zoom

(a) 2nd main Reculée→ Non Détectée Baissée→ Inactive	(b) 2nd main Avancée→ Détectée Baissée→ Inactive
(c) 2nd main Reculée→Zoom arrière Levée→Active	(d) 2nd main Avancée→Zoom avant Levée→Active

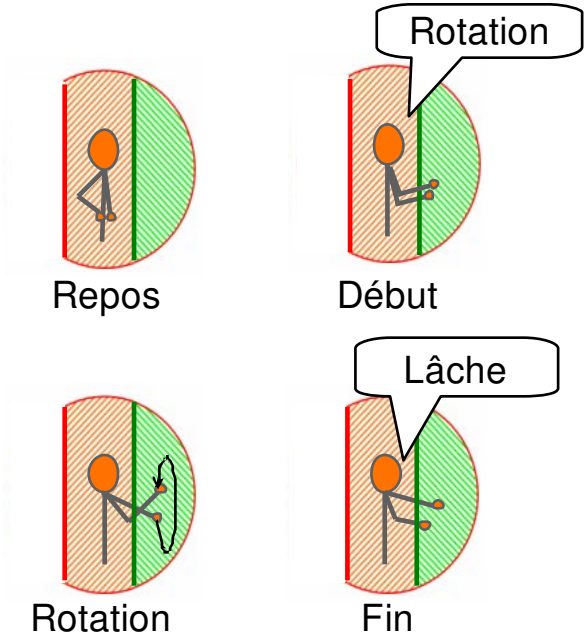


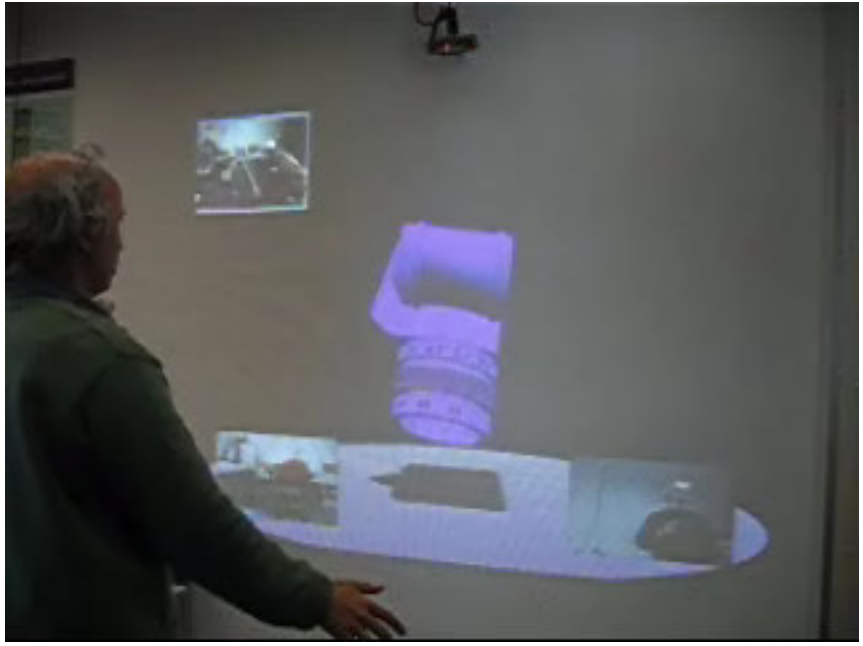
Interprétation de gestes : redimensionnement





Interprétation de gestes : rotation

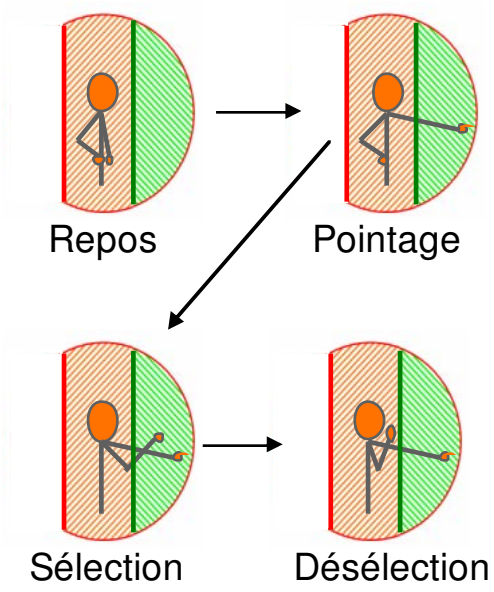


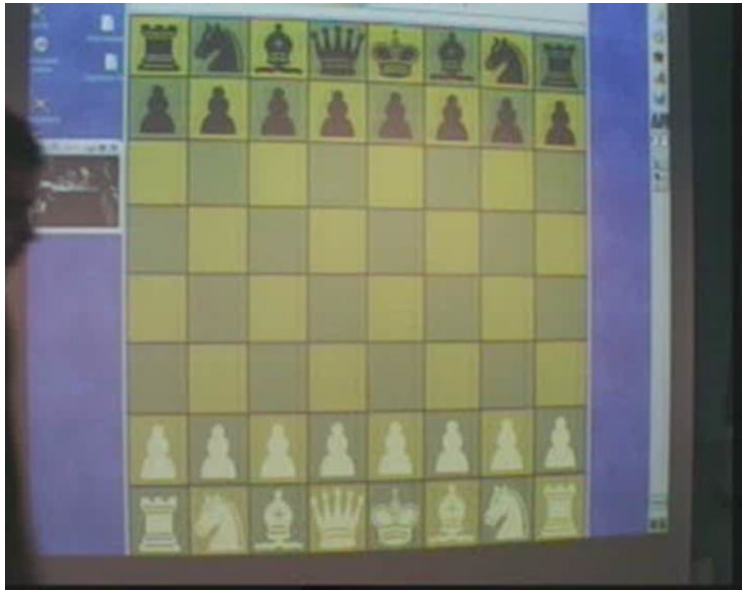


Interprétation de gestes : pointage / sélection

1ère main avancée → pointage (convention axe tête-main)

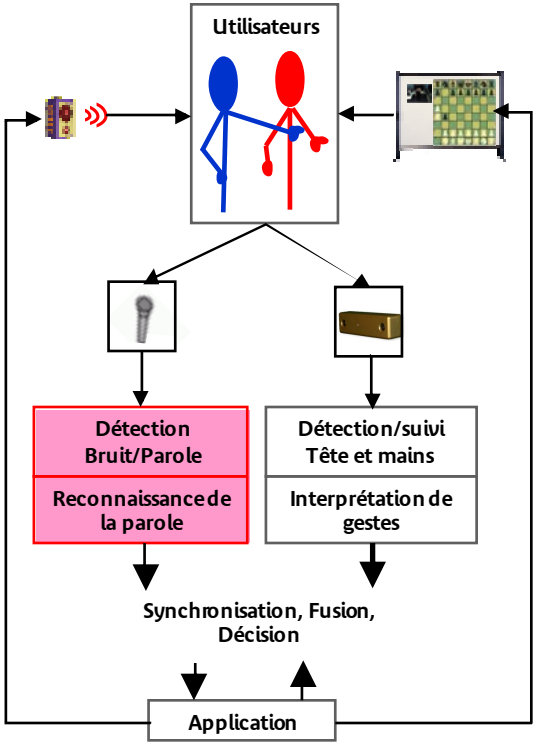
2nd main avancée → sélection (convention main avancée = sélection)



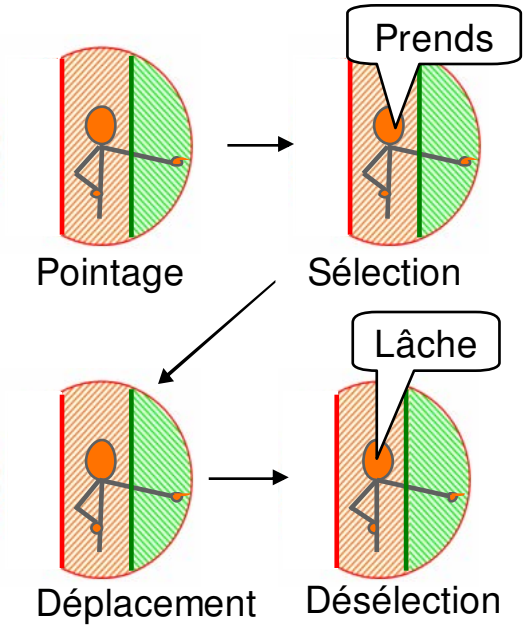


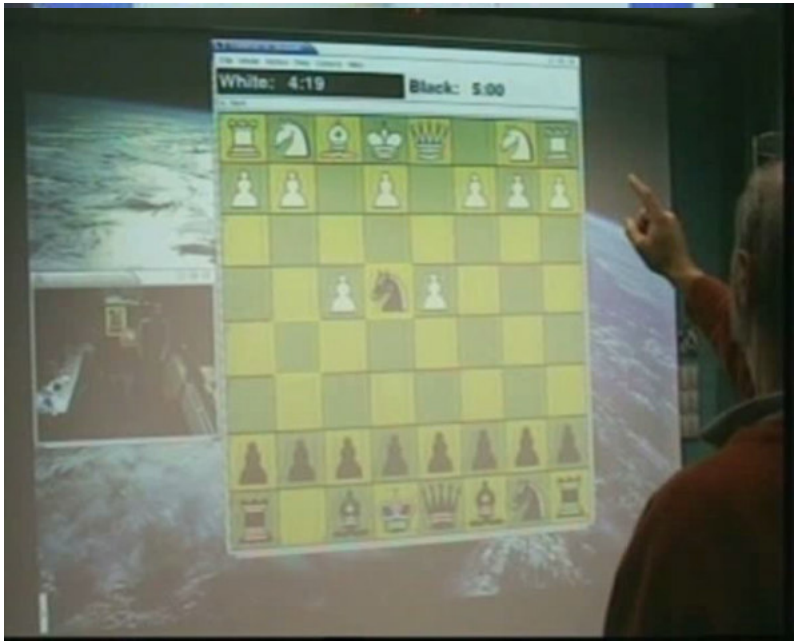
Plan

- **Introduction**
- Détection des gestes
- Interprétation de gestes
- Reconnaissance de la parole
- Synchronisation des modalités
- Evaluation de la souris oro-gestuelle
- Fusion des modalités avec le contexte
- Conclusion



Souris oro-gestuelle





Reconnaissance de parole

Reconnaissance de parole déjà disponible à France Télécom R&D

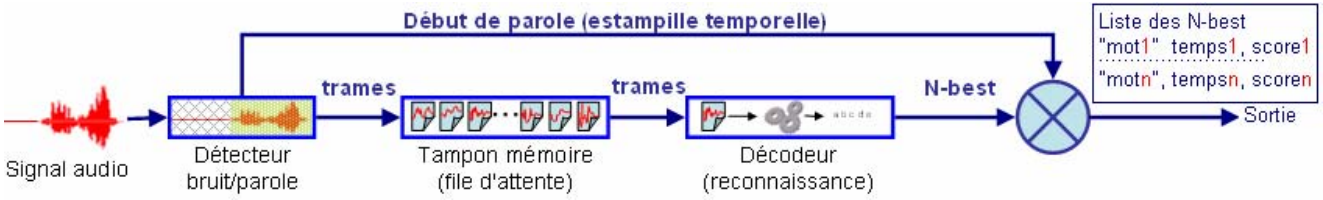
→Collaboration avec le laboratoire TECH/SSTP

Détecteur bruit/parole

- Fournit au plus tôt les estampilles temporelles de début et fin de parole en sortie
- Ne transmet au décodeur que les trames de parole

Décodeur

- Traite les trames + vite que le temps réel (rattrape son retard)
- Délai de 240 ms (détection du retour au silence)
- Calcul de n-best (temps faible par rapport 240 ms)



Reconnaissance de parole

- Exemple de N-best obtenu

20060628_104931.001 **speech_start**
20060628_104932.281 **speech_end**

Estampilles temporelles de début et fin de parole

!parole/tst_mkvnb Version 4.1 Juin 2004
!CMD: @20060628_104930 parole/tst_mkvnb -sol=10 -in -out -bdl -ver parole/echec_20050107_F15x27c_g08_R0p.gkz
!EXP: Src=si.inl;Type=f10.r10;File=1;

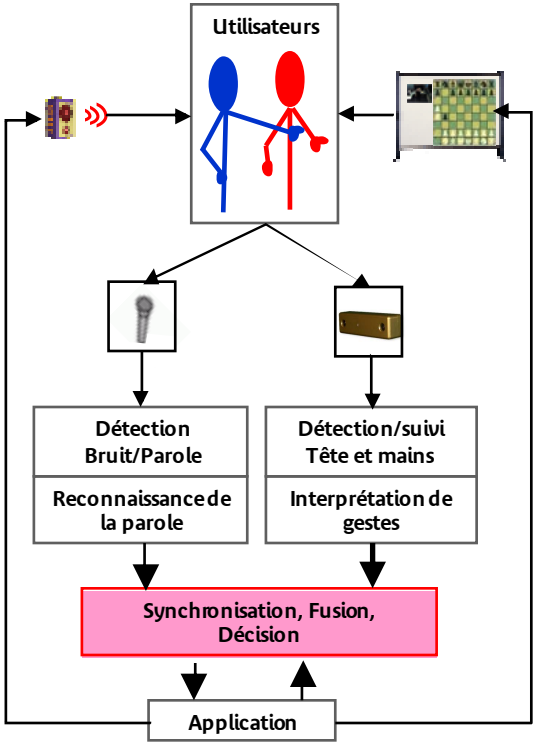
Sol=1 ; Dec="**prend la reine**" ;	Score=**1967** ; ...
Sol=2 ; Dec="**prend la reine haut**" ;	Score=**1896** ; ...
Sol=3 ; Dec="**prend la reine du haut**" ;	Score=**1817** ; ...
Sol=4 ; Dec="**haut prend la reine**" ;	Score=**1793** ; ...
Sol=5 ; Dec="**prend la reine en haut**" ;	Score=**1732** ; ...
Sol=6 ; Dec="**fou prend la reine**" ;	Score=**1723** ; ...
Sol=7 ; Dec="**haut prend la reine haut**" ;	Score=**1722** ; ...
Sol=8 ; Dec="**fou prend la reine haut**" ;	Score=**1652** ; ...
Sol=9 ; Dec="**haut prend la reine du haut**" ;	Score=**1643** ; ...
Sol=10 ; Dec="**REJET**" ;	Score=**1642** ; ...

N-best

Rang **Phrases reconnues** **Score**

Plan

- **Introduction**
- Détection des gestes
- Interprétation de gestes
- Reconnaissance de la parole
- Synchronisation des modalités
- Evaluation de la souris oro-gestuelle
- Fusion des modalités avec le contexte
- Conclusion



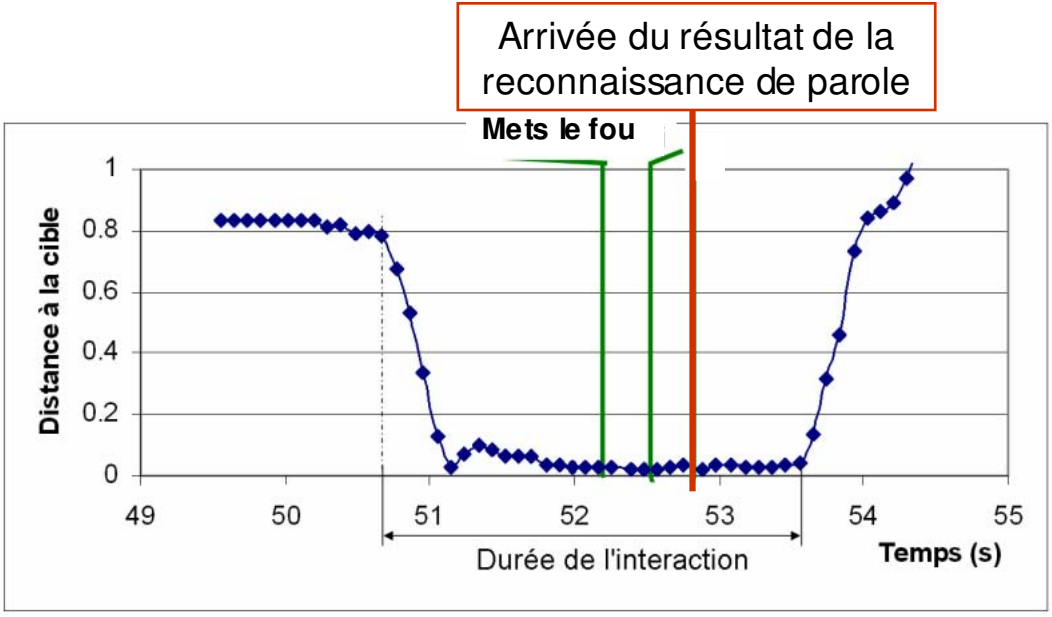
Synchronisation des modalités geste et parole

Sans synchronisation

Prise en compte du geste au moment de l'arrivée du résultat parole

Problème

Déplacement du curseur avant arrivée du résultat de reco de parole



Synchronisation des modalités geste et parole

Problème

Déplacement du curseur avant arrivée du résultat de reco de parole

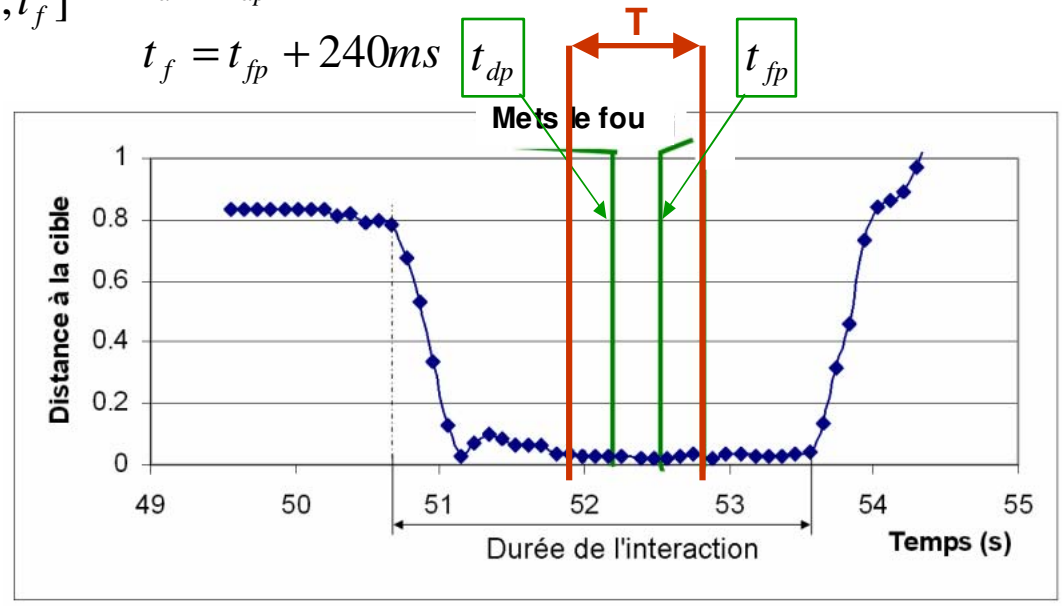
Synchronisation des modalités geste et parole

- Moyenne de la direction pointée pendant un intervalle de temps comprenant la parole [Stiefelhagen et al. 2004]
- Hypothèse supplémentaire : la main marque une pause sur la cible
  →pondération par la vitesse

$$T = [t_d, t_f]$$

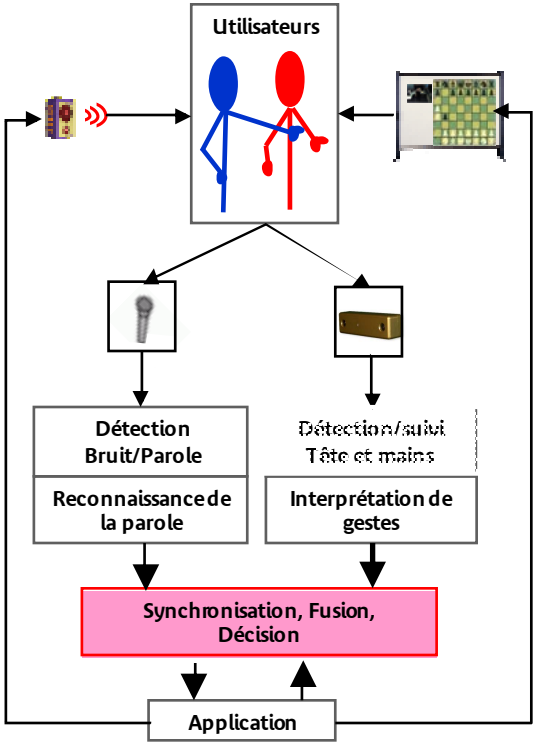
$$t_d = t_{dp} - 240ms$$

$$t_f = t_{fp} + 240ms$$



Plan

- **Introduction**
- Détection des gestes
- Interprétation de gestes
- Reconnaissance de la parole
- Synchronisation des modalités
- Evaluation de la souris oro-gestuelle
- Fusion des modalités avec le contexte
- Conclusion



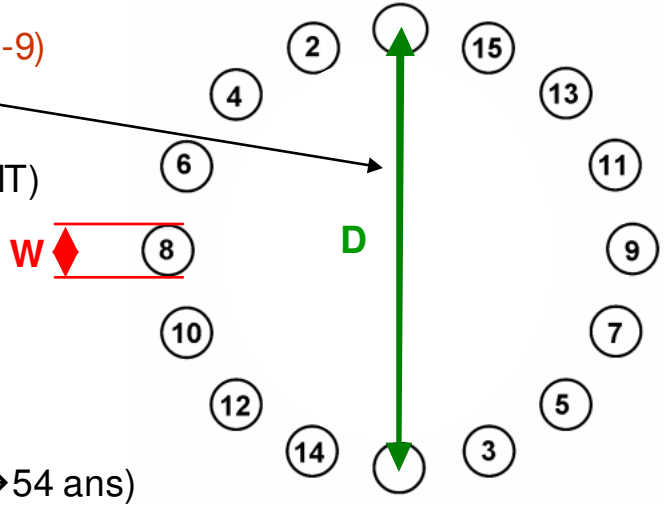
Evaluation de la souris oro-gestuelle (ISO-9241-9)

Loi de Fitts [Fitts 1954] → difficulté d'une tâche pointage/sélection

Indice de difficulté (*formulation de Shannon*) : $ID = \log_2\left(\frac{D}{W} + 1\right)$

Protocole expérimental (norme ISO-9241-9)

- Sélectionner 16 cibles dans l'ordre
- Mesure :
  - Temps de mouvement inter-cible (MT)
  - Distance au centre de la cible
  - % d'erreur de sélection



Paramètres

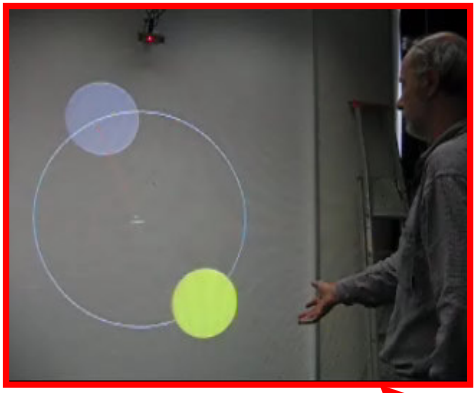
- 10 adultes (7 hommes, 3 femmes), (24→54 ans)
- 3 dispositifs testés (**souris oro-gestuelle**, **souris gestuelle**, **souris**)
- 13 conditions taille-distance (W, D)
- 15 trajectoires inter-cibles

→ 10x3x13x15 = **5850** essais

Evaluation de la souris oro-gestuelle (ISO-9241-9)

Fitts' Law [Fitts 1954] → MT = a + b ID, ID: difficulté d'une tâche de pointage sélection.

Indice de difficulté (*formulation* de *Shannon*) : $ID = \log_2\left(\frac{D}{W}+1\right)$



Parametres :

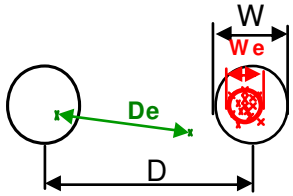
- 10 adultes (7 hommes, 3 femmes), (age entre 24 et54)
- 3 dispositifs testées (**souris oro-gestuelle**, **souris gestuelle**, **souris ordinaire**)
- 13 conditions sur les cibles (taille W-distance D) (5 indices de difficulté)
- 15 trajectoires intercibles

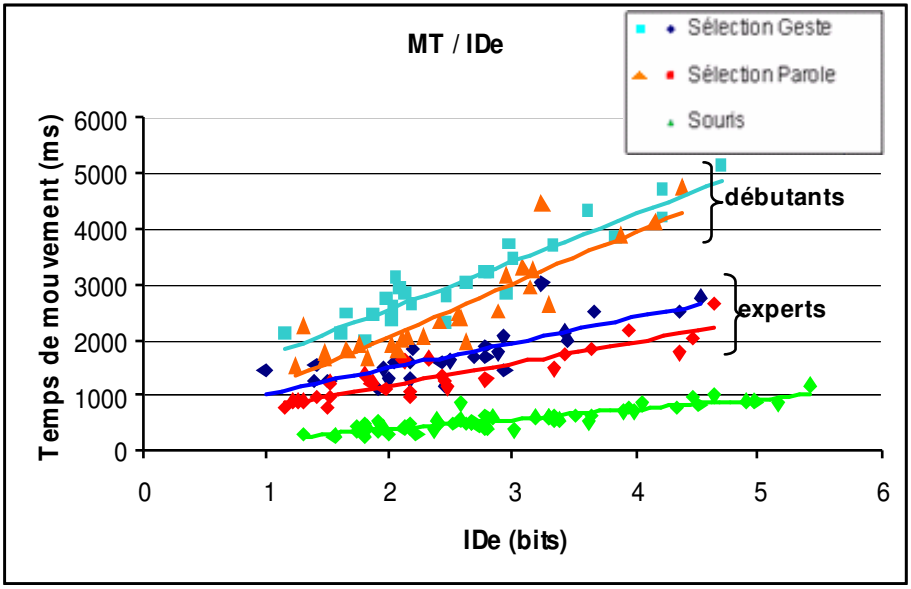
→ 10x3x13x15 = **5850** essais

Distance D / Width W	32	64	128	256
96	2.00	1.32	(0.81)	(0.46)
192	2.81	2.00	1.32	(0.81)
384	3.70	2.81	2.00	1.32
768	4.64	3.7	2.81	2.00

Evaluation de la souris oro-gestuelle (ISO-9241-9)

- Indice de difficulté (*formulation de Shannon*): $ID = \log_2\left(\frac{D}{W}+1\right)$
- Indice de difficulté effectif : $IDe = \log_2\left(\frac{De}{We}+1\right)$





Comparaison avec d'autres études

Study	Year	Methodology ID	TP	Layout	Task Type	Learning	Adj. Acc.	ID Range (bits)	TP (bits/s)	Error (%)	Dispositif
		Shannon	Mean	Circular	Serial	Practice	Yes	1.32-4.64	1.06	14.6	Geste
		Shannon	Mean	Circular	Serial	Practice	Yes	1.32-4.64	1.20	17.8	Parole
		Shannon	Mean	Circular	Serial	Practice	Yes	1.32-4.64	[illegible]	[illegible]	Souris

notre étude

[Soukoreff et MacKenzie 2004]

Study	Year	Methodology ID	TP	Layout	Task Type	Learning	Adj. Acc.	ID Range (bits)	TP (bits/s)	Error (%)	Notes
Douglas, Kirkpatrick & MacKenzie	1999	Shannon	Mean	Horizontal	Serial	Yes	Yes	2.3–6.3	2.1	17.5	Isometric Joystick
		Shannon	Mean	Circular	Discrete	Yes	Yes	2.3–6.3	2.2–2.3	3.4	Isometric Joystick
									1.8	25.0	Touchpad
Isokoski									1.7–1.9	3.8	Touchpad
									4.4–4.5	6.9–9.4	Six Mice
Keates, Hwang, Langdon, Clarkson & Robinson	2002	Shannon	Mean	Circular	Serial	No	Yes	—	1.77	69	Mouse (motion impaired)
									4.88	11	Mouse (able-bodied)
MacKenzie & Orsnoczak	1993	Shannon	Mean	Horizontal	Serial	Yes	Yes	1.0–4.1	0.99	4.07	Touchpad
					Serial	Yes	Yes	1.0–4.1	1.43	9.92	Tactile Touchpad
					Serial	Yes	Yes	1.0–4.1	1.07	5.76	Lift & Tap Touchpad
MacKenzie					Serial	No	Yes	1.0–4.1	3.7	2.4	Mouse
		Shannon	Mean	Horizontal	Serial	No	Yes	1.0–4.1	4.1	3.5	GyroPoint
		Shannon	Mean	Horizontal	Serial	No	Yes	1.0–4.1	1.4	1.6	RemotePoint
MacKenzie, Kauppinen & Silfverberg	2001	Shannon	Mean	Circular	Serial	Yes	Yes	3.8	4.9	9.6	Mouse
		Shannon	Mean	Circular	Serial	Yes	Yes	3.8	3.0	8.6	Trackball
		Shannon	Mean	Circular	Serial	Yes	Yes	3.8	1.8	9.0	Joystick
		Shannon	Mean	Circular	Serial	Yes	Yes	3.8	2.9	7.0	Touchpad
Oh & Stuerzlinger	2002	Shannon	Mean	Circular	Serial	Yes	Yes	2.58–3.75	3.04	32	Laser Pointer
		Shannon	Mean	Circular	Serial	Yes	Yes	2.58–3.75	4.09	20	Mouse
Silfverberg et al.	2001	Shannon	Mean	Circular	Discrete	Practice	Yes	1.6–5.1	1.6–2.55	1.8–9.6	Isometric Joystick
Poupyrev, Okabe & Maruyama	2004	Shannon	Mean	Horizontal	Serial	No	Yes	2.3–6.3	7.14		Stylus Tapping
		Shannon	Mean	Horizontal	Serial	No	Yes	2.3–6.3	4.8–5.9		Stylus Dragging

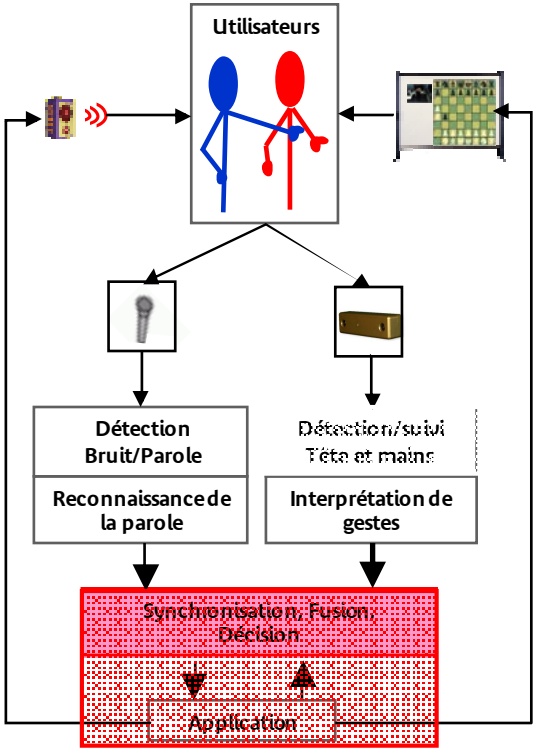
Performances supérieures à notre interface

Performances proches de notre interface

Performances souris

Plan

- **Introduction**
- Détection des gestes
- Interprétation de gestes
- Reconnaissance de la parole
- Synchronisation des modalités
- Evaluation de la souris oro-gestuelle
- Fusion des modalités avec le contexte
- Conclusion

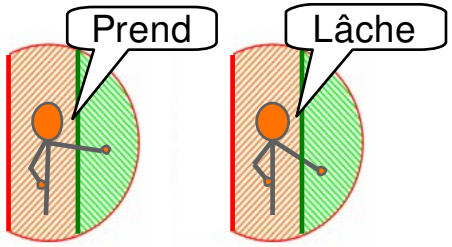


Fusion des modalités : contexte applicatif

Souris oro-gestuelle

→ générique

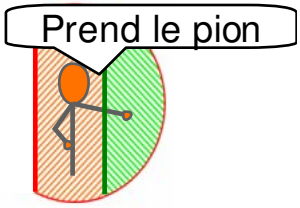
→ n'exploite pas la richesse de la parole



Pourquoi utiliser le contexte applicatif pour la fusion ?

→ exploiter la richesse de la parole

→ commandes plus spontanées, plus courtes

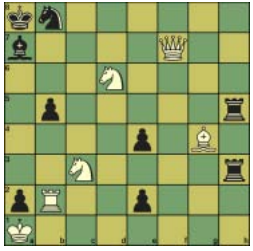


Pourquoi le contexte d'un jeu d'échec ?

→ jeu de commandes simples et variées

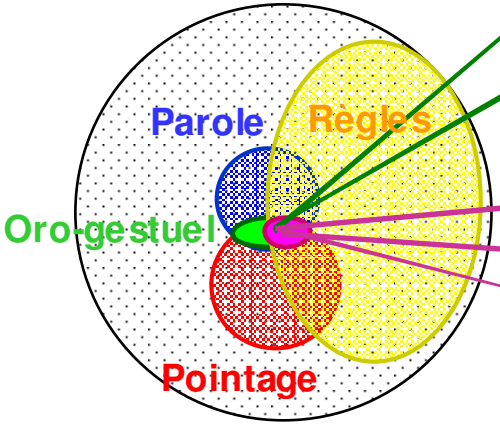
→ contexte simple et restreint

→ un ou deux joueurs



Fusion des modalités : décision

- Complémentarité



Un coup → effectue le coup

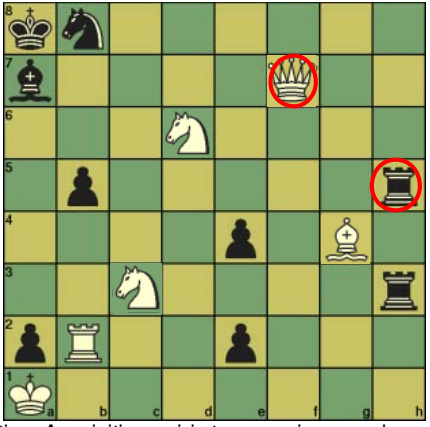
Plusieurs coups → ambiguë

Hypothèse : le joueur veut effectuer un coup légal

Ensemble oro-gestuel légal

Un coup → effectue le coup

Plusieurs coups → montre l'ambiguïté



"Prends le pion"

"Cavalier prend pion"

"fou prend"

Résultat

vidéo

Résumé : 3 interactions possibles

A : 'drag and drop' oro-gestuel

2 pointages gestuels

+ sélection/désélection à la voix

3,52 secondes

B : 'drag and drop' gestuel

2 pointages gestuels

+ sélection/désélection gestuelle

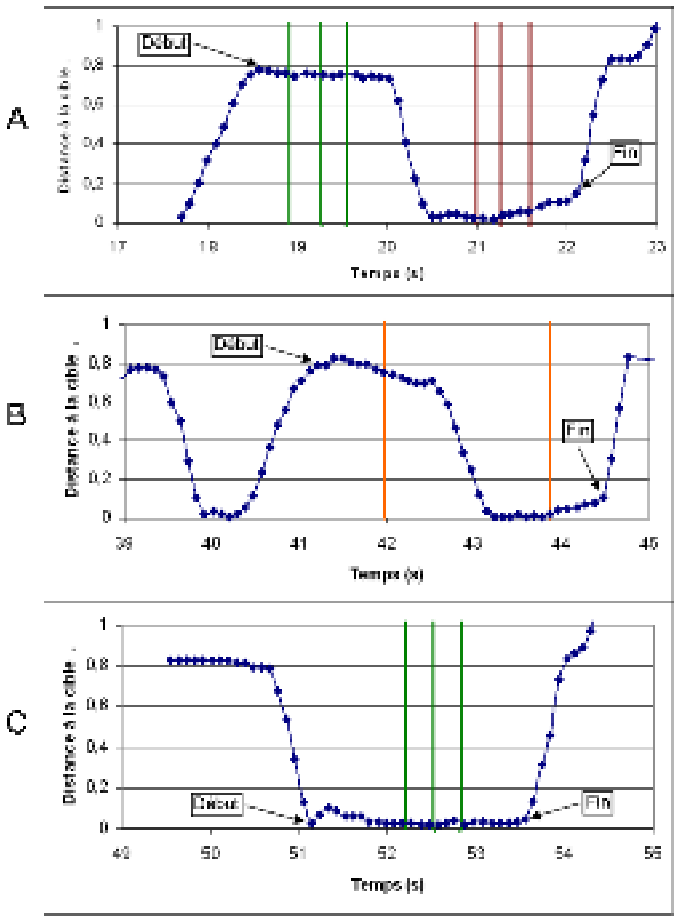
3,26 secondes

C : interaction multimodale avec utilisation du contexte

1 pointage gestuel

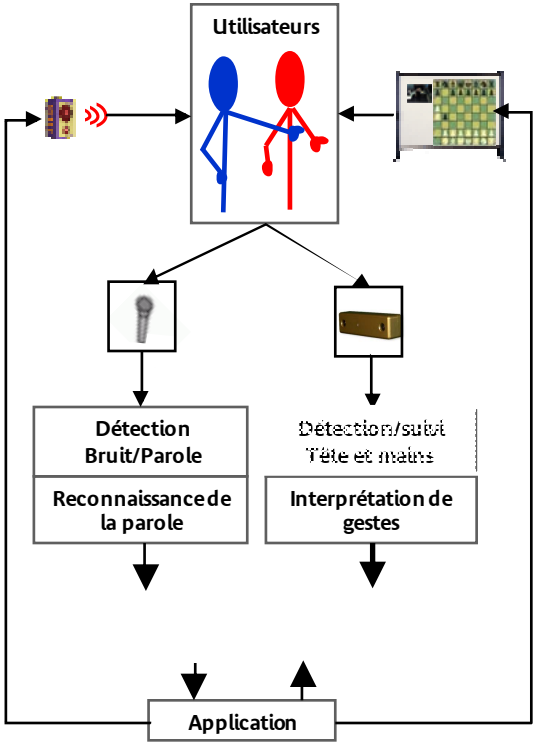
+ "met la reine"

2,41 secondes



Plan

- **Introduction**
- Détection des gestes
- Interprétation de gestes
- Reconnaissance de la parole
- Synchronisation des modalités
- Evaluation de la souris oro-gestuelle
- Fusion des modalités avec le contexte
- **Conclusion**



Conclusion

Interaction grand écran

•"Personne libre"
•Pas de calibrage ou d'apprentissage spécifique à l'utilisateur
•Gestes spontanés

Détection et Suivi des parties du corps (vision par ordinateur)

•Temps réel (multi-personnes)
•Robustesse: aux distracteurs avant-plan et arrière plan

Multimodalité

•Expérimentation selon norme ISO → caractérisation interface
•Souris oro-gestuelle → générique
•Fusion avec le contexte → commandes multimodales plus spontanées

FIN

Merci de votre attention

- Message personnel : Proposition de thèse
- Intitulé **:** Fusion multimodale pour terminaux mobiles et contenus interactifs
- Laboratoire d'accueil : FT Orange Labs, division Technologies à Lannion.et à Grenoble
- Période : Automne 2007- Automne 2010. Thèse CIFRE.
- Statut : CDD-FR de 36 mois. Salaire mensuel de l'ordre de 1680 € net
- Co-encadrants à FT : Jean Emmanuel VIALLET (TECH/IRIS) et Eric PETIT (TECH/IDEA)
- Directeur de thèse universitaire : à déterminer
- jeanemmanuel.viallet@orange-ftgroup.com