

# Meeting Room : Interactive Systems Laboratories Project

Seminaire de Master - SH

Skultety Christophe - Université de Fribourg  
christophe.skultety@unifr.ch

Janvier 2005

## Introduction

Ce document a été écrit dans le cadre d'un travail de séminaire de Master à l'Université de Fribourg dans le groupe de recherche DIVA (Documents Image Voice Analysis). L'objectif du séminaire est de faire un état de l'art complet sur les projets actuels d'enregistrement et d'analyse de réunion. Le projet présenté ici porte le nom de "Meeting Room : an Interactive Systems Laboratories Project" et à été développé a l'Université de Canergie Mellon conjointement avec l'Université de Karlsruhe.

## 1 Le Projet "Meeting Room"

Les réunions interviennent pratiquement quotidiennement dans n'importe quel environnement de travail. Un individu absent peut se poser beaucoup de questions sur une réunion : qui était présent ? quels sont les sujets qui ont été abordés ? quelles décisions ont été prises ou encore quelles réactions ont eu les participants ? etc. Les informations permettant de répondre à ces questions se trouvent au niveau de la vidéo (gestes, expression du visage, langage corporel) ou du son (paroles, intonation).

Le but du projet "Meeting Room" est de développer un système d'enregistrement de réunion multimodal et automatisé, permettant de répondre à la plupart des questions que peut se poser une personne sur la réunion. Ce système permet de visualiser l'enregistrement de la vidéo et du son, ainsi que la transcription de la parole. Il possède les caractéristiques suivantes : il est non-intrusif, les participants de la réunion peuvent se déplacer et parler librement dans la pièce. Il doit résoudre le problème d'assignement, c'est-à-dire être capable d'identifier l'orateur à tout moment. Le système est automatisé au maximum, c'est un aspect essentiel pour un déploiement efficace et une utilisation presque systématique du système. L'identification dans le système est multimodale. Un identificateur unique échouera forcément dans une situation particulière, par exemple la reconnaissance basée sur le visage ne fonctionnera pas si l'orateur tourne le dos à la caméra. Une identification multimodale (basée sur plusieurs modes d'identification) donnera de meilleurs résultats. L'ensemble du système fonctionne pratiquement en temps réel.

Ce document est organisé comme suit : la section 2 présente une vue générale du système et de son architecture avec ses différents composants, les sections 3 à 5 présentent les parties

les plus importantes du système, la reconnaissance de la parole, l'identification de personnes multimodale et le browser de réunion, et finalement la section 6 offre une conclusion ainsi qu'un aperçu du travail futur.

## 2 Vue générale du système

Une salle de conférence a été spécialement équipée pour le développement du système. Celui-ci est capable d'identifier automatiquement jusqu'à six orateurs. L'architecture du système est présentée à la figure 1. Les enregistrements des flux audio et vidéo sont transmis au composant d'identification de personnes multimodale. Celui-ci identifie les participants et l'orateur. Les flux audio et vidéo sont ensuite transmis au browser de réunion et le flux audio est également transmis au composant de reconnaissance de la parole Janus. Le système a été étendu pour supporter plusieurs systèmes de reconnaissance Janus simultanément afin de satisfaire aux exigences du temps réel. Janus produit des hypothèses de dialogue qui sont transmises au serveur de dialogue. Il est en charge de transformer ces hypothèses en format de dialogue. La transcription est transmise au browser de réunion pour qu'elle puisse y être visualisée.

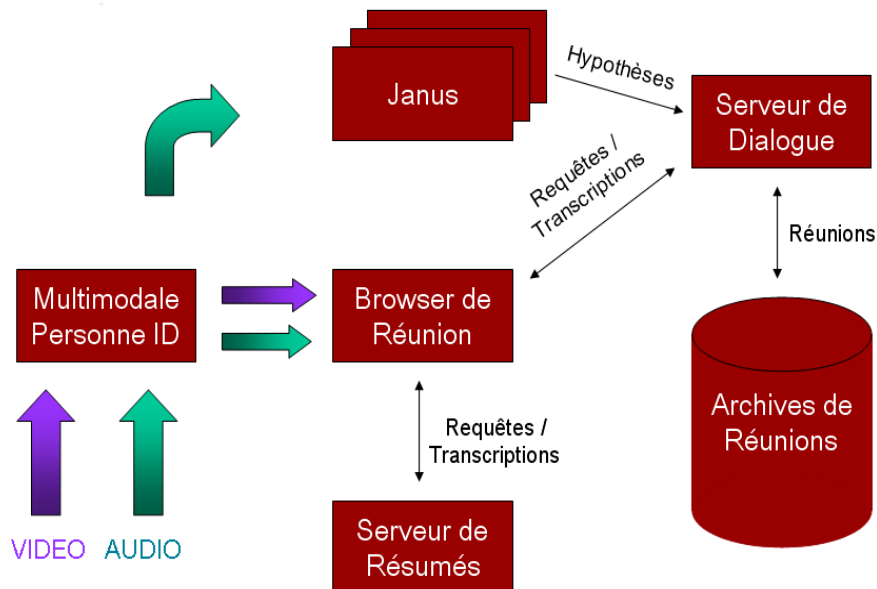


FIG. 1 – Architecture du Système

Il y a également un composant qui porte le nom de serveur de résumés. A la demande d'un utilisateur, ce serveur va produire un résumé de la taille spécifiée, qui pourra être visualisé dans le browser. Au moment où une réunion se termine, elle est immédiatement archivée, avec ses résumés s'il y en a, dans une base de données. Ceci se fait à travers le serveur de dialogue.

## 3 Reconnaissance de la Parole

Le but du système de reconnaissance de la parole est de produire une transcription automatique du dialogue. Comme il a été dit précédemment, le système est basé sur le toolkit de

reconnaissance Janus (JRTk). Le travail se fait sur un enregistrement audio continu avec de multiples orateurs pouvant utiliser de multiples micros.

Le principe de la reconnaissance de la parole est le suivant : il faut partitionner les données en segments homogènes et assigner à chacun un label orateur. Un segment est homogène s'il ne contient que les paroles d'un orateur. Il peut se produire un phénomène appelé sur-segmentation (le segment ne contient par l'ensemble d'un tour de parole d'un orateur, il est coupé) ou un phénomène de sous-segmentation (il y a plus qu'un orateur sur le même segment). Une fois la segmentation effectuée, deux techniques nécessitant un seul orateur par uttérance, sont utilisées pour réaliser l'affectation : VTLN (Vocal Tract Length Normalization) et Speaker Adaptation.

## 4 Reconnaissance des Personnes

Le but du module d'identification des personnes est de traquer et d'identifier de manière continue les participants d'une réunion dans une pièce. Pour augmenter la robustesse et l'efficacité du processus d'identification, une approche multimodale a été adoptée avec l'intégration de plusieurs systèmes de reconnaissance. Le schéma du système d'identification des personnes est présenté à la figure 2.

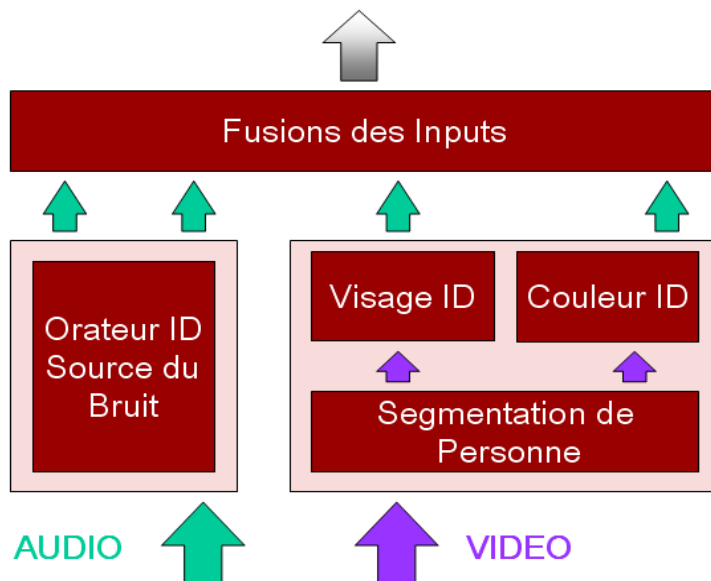


FIG. 2 – Schéma de l'Identificateur de Personnes Multimodale

Le flux vidéo passe d'abord par une étape de segmentation des personnes qui précède l'identification du visage et l'identification de la couleur. L'espace de recherche dans l'image étant trop grand pour satisfaire à l'exigence du temps réel, il est nécessaire de le réduire. Le but de cette étape de pré-calcul est de segmenter les personnes par rapport au fond. La technique utilisée est basée sur la soustraction du fond. Pour extraire les personnes, les différences entre la nouvelle image et une image du fond prise précédemment sont calculées. La qualité de cette segmentation est liée à la qualité de la représentation du fond. Le système ne peut pas se baser uniquement sur une image du fond fixe, mais doit garder un modèle évolutif car des objets comme des chaises, une lampe ou une table peuvent être déplacés au cours de la réunion.

Une fois les personnes segmentées, l'identification peut débuter. Pour le flux vidéo, elle se fait sur la base de la couleur et du visage. Pour l'identification de la couleur un modèle est créé pour chaque participant en utilisant des histogrammes de couleur. L'avantage de l'identification par la couleur est que cette technique ne souffre pas de problèmes d'occlusion ou de changement de vue. Par contre la couleur est modifiée par la variation de l'illumination. Ce sont donc des histogrammes de teinte-saturation qui sont utilisés en combinaison avec le calcul d'une fonction de distribution. Un test a été réalisé avec des personnes circulant dans un vestibule. Les résultats, avec seize modèles distincts testés sur environ 5000 images, on permit d'obtenir 67.8% de taux de reconnaissance avec une taille d'histogramme optimale de 75x75.

Egalement basé sur le flux vidéo, le système identifie les personnes en se basant sur le visage. Au moment où l'article a été écrit, cette détection était développée mais pas encore intégrée au système. Le travail se fait sur plusieurs visages de différentes tailles dans différentes positions. La localisation du visage dans l'image se fait avec la couleur de la peau à l'aide d'un système développé à l'ISL. L'identification du visage est basée sur les techniques d'EigenFaces et DSW (Dynamic Space Warping).

Le système doit être capable d'identifier l'orateur à tout moment à partir du flux audio. Cette identification se fait sur la base du spectre de la parole de chaque orateur. Il est nécessaire d'entraîner le système au préalable. Cela se fait offline. L'estimation de la source du bruit permet de combiner l'audio et la vidéo. Détecter la provenance du bruit permet de sélectionner la caméra la plus appropriée pour afficher l'enregistrement dans le browser.

Une fois que chaque module a effectué sa tâche, les informations sont passées à la fusion des inputs. Le but est de trouver la configuration la plus probable de l'identité des participants ainsi que de l'orateur. Un simple test a été réalisé avec trois personnes, deux micros et deux caméras avec un dialogue d'environ quatre minutes. Un résultat a été calculé avec la fusion des inputs des différents identificateurs et un autre résultat sans la fusion, dans ce cas une configuration est considérée comme erronée, si un seul des identificateurs échoue. Le taux d'erreur avec la fusion des input est de 10.67%, tandis que sans la fusion des inputs, le taux d'erreur est de 12.51%.

## 5 Browser de Réunion

Une partie importante des systèmes de reconnaissance de réunions est la capacité à capturer, manipuler et visualiser tous les aspects de la réunion. C'est à cette fin qu'un browser a été développé. Il supporte les possibilités suivantes :

- Fournir un accès rapide aux enregistrements (vidéo, audio et transcription), pour la visualisation
- Archivage des réunions dans une base de données
- Création automatique de résumés audio, vidéo et de la transcription
- Possibilité aux participants d'apporter des corrections ou des annotations au terme de la réunion

La figure 3 montre l'aspect de la fenêtre principale du browser. Il y a trois parties distinctes : celle du haut montre le déroulement de la réunion et les participants. Celle du bas à gauche permet la visualisation de la transcription et celle située en bas à droite permet d'afficher la vidéo ou un résumé.

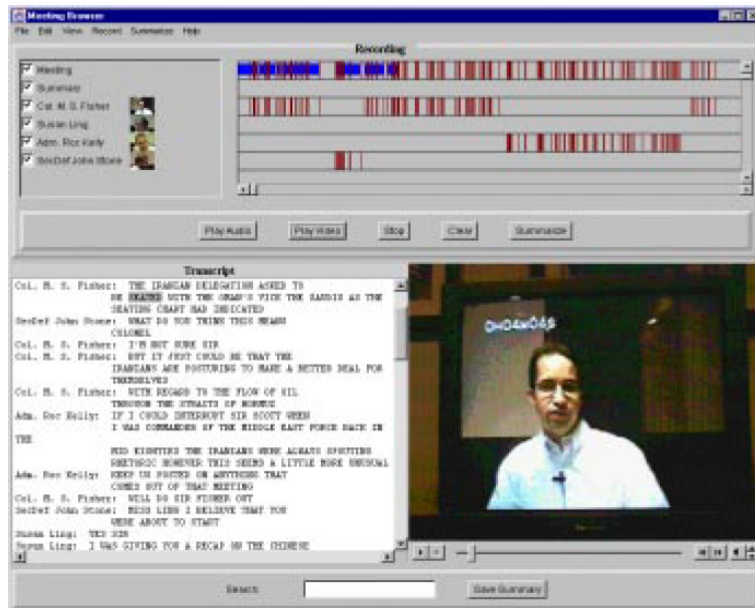


FIG. 3 – Fenêtre principale du Browser

Au terme de la réunion, l'enregistrement ainsi que les résumés s'il y en a, sont stockés dans une base de donnée pour la visualisation future. Les réunions y sont représentées dans un format en arbre avec tous leurs attributs. L'utilisateur peut effectuer une recherche par n'importe quelle combinaison de participants à la réunion, par sujets discutés, par mots-clés, par la durée ou encore la date et l'heure de la réunion. S'il existe un résumé pour la réunion, il est possible de le visualiser dans le browser sans devoir charger la réunion en entier.

Il est possible de réaliser un résumé audio, vidéo et de la transcription de la réunion. Un résumé est toujours créé sur la base de la transcription, ensuite les portions vidéo et audio correspondantes y sont attachées. L'utilisateur peut spécifier la taille ainsi que le sujet central de son résumé. Cette information est envoyée au serveur de résumés qui analyse le dialogue et retourne le résumé. La création du résumé se fait sur la base de l'algorithme suivant :

1. Elimination des mots-outils "stopwords", de la transcription. (exemples de mots-outils : The, an, etc.)
2. Phase de stemming (recherche de la racine des mots) et identification du stem le plus fréquent. Pour la simplification, ce ne sont que les quatre premières lettres de chaque mot qui sont considérées pour la suite de l'algorithme
3. Chaque tour de parole est pondéré en fonction du nombre d'apparition du stem le plus fréquent
4. Le tour de parole "le plus lourd", c'est-à-dire celui où apparait le plus de fois le stem le plus fréquent, est inclus dans le résumé
5. Afin d'éviter la redondance, les mots contenant le stem le plus fréquent seront désormais ignorés
6. Si la taille du résumé est inférieure à celle spécifiée par l'utilisateur, alors on retourne à l'étape 2.

Il existe également un module de détection des émotions qui est incorporé au browser de réunion. Il permet de donner à l'utilisateur des informations sur l'état d'un orateur : énervé,

calme, joyeux, etc. La détection d'émotion se fait à l'aide de deux systèmes développés à l'ISL : un détecteur d'émotion et un système de détection des caractéristiques du discours. Ces deux systèmes fonctionnent offline sur les données de la réunion, mais les résultats sont incorporés à l'enregistrement et visualisables dans le browser.

## 6 Conclusion

Il y a encore beaucoup de travail qui peut être réalisé sur le projet "Meeting Room". L'ISL prévoit pour le travail futur :

- une amélioration des différents systèmes de reconnaissance
- l'intégration au système du module de reconnaissance du visage
- le développement d'un module de détection de l'attention (qui regarde où?)
- la réalisation de tests du système en dehors des laboratoires ISL
- la possibilité de réaliser un résumé sur plusieurs réunions
- de réaliser le "Meeting Room" sur un Laptop

La principale force du système est le fait qu'il soit automatisé au maximum. Comme il a été dit précédemment, pour que le système soit fréquemment utilisé, c'est une condition nécessaire. Particulièrement en ce qui concerne la transcription de la parole. Si celle-ci se fait manuellement, le système restera inévitablement au stade de projet académique. Egalement, la non-intrusivité du système le rend agréable à utiliser. Par contre, il n'y a pas d'informations sur les inputs manuels nécessaires pour lancer le système. On peut aussi s'interroger sur le taux d'erreur à environ 10% avec la fusion des inputs, cependant le travail va probablement faire baisser ce pourcentage de taux d'erreur. L'aspect documents n'est pas du tout pris en compte dans ce projet, mais par contre il est traité dans le cadre d'autres projets développés à l'ISL.

## Références

- [1] Michael Bett, Ralph Gross, Hua Yu, Xiaojin Zhu, Yue Pan, Jie Yang and Alex Weibel, *Multimodal Meeting Tracker*, Proceedings of RIAO2000, Paris, 2000.
- [2] Jie Yang, Xiaojin Zhu, Ralph Gross, John Kominek, Yue Pan and Alex Weibel, *Multimodal People ID for a Multimedia Meeting Browser*, Proceedings of ACM Multimedia, 1999.
- [3] A. Waibel, M. Bett, M. Finke, R. Stiefelwagen, *Meeting Browser : Tracking and Summarizing Meetings*, Broadcast News Transcription and Understanding Workshop, Lansdowne, Virginia 1998.
- [4] Matthew A. Turk and Alex P. Pentland, *Face Recognition Using Eigenfaces*, Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pages 586-591, 1991.
- [5] Klaus Zechner, *Automatic Generation of Concise Summaries of Spoken Dialogues in Unrestricted Domains*, Proceedings of the 24th ACM-SIGIR International Conference on Research and Development in Information Retrieval, New Orleans, LA, 2001
- [6] Thomas S. Polzin and Alex H. Waibel, *Detecting Emotions in Speech*, Proceedings of the CMC, 1998