Vers une prise en compte des contraintes de présentation multimodale lors du calcul de réaction d'un système de dialogue personne-machine

Meriam Horchani
France Télécom R&D / CLIPS-IMAG
2 avenue Pierre Marzin
22307 Lannion, France
(+33) 2 96 05 32 95
meriam.horchani@orange-ftgroup.com

Laurence Nigay
CLIPS-IMAG
385 rue de la bibliothèque
38041 Grenoble, France
(+33) 4 76 51 44 40
laurence.nigay@imag.fr

Franck Panaget
France Télécom R&D
2 avenue Pierre Marzin
22307 Lannion, France
(+33) 2 96 05 28 52
franck.panaget@orange-ftgroup.com

RESUME

Alors que l'ère est à la personnalisation des interfaces-utilisateurs, celle-ci reste trop souvent superficielle : elle porte uniquement sur la présentation et n'influence que peu le contenu sémantique des réactions des systèmes. Travaillant sur des systèmes d'information grand-public, nous visons une communication personne-machine la plus naturelle possible. Ceci passe, d'une part, par une coopération accrue du système – réalisée notamment via une surinformation pertinente – et, d'autre part, par la prise en compte de l'influence des contraintes de présentation sur la réaction du système – en particulier pour identifier les cas où le système a trouvé trop de solutions à une requête donnée pour une présentation efficace et sans surcharge cognitive. Par contraintes de présentation, nous entendons les contraintes inhérentes aux modalités utilisées, les contraintes induites par l'environnement d'utilisation et les contraintes imposées par l'utilisateur lui-même. Dans le cadre de nos travaux, la prise en compte de ces contraintes et de leur influence sur la sélection du contenu passe par la proposition d'un composant de choix qui fait le pont entre le composant de dialogue et ceux de présentation et d'interaction des IHM inspirées du méta-modèle ARCH. Considérant un composant de dialogue capable de fournir un ensemble de réponses possibles au composant de choix que nous proposons, celui-ci utilise les contraintes de présentation pour affiner la réaction du système et indiquer au composant de présentation la sortie à générer. Notre article décrit plus précisément le composant proposé ainsi que son articulation par rapport aux autres composants du méta-modèle ARCH.

Categories and Subject Descriptors

D.2.2 [**Software Engineering**]: Design Tools and Techniques – *user interfaces;* H.5.m [**Information Interfaces and Presentation (e.g., HCI)**]: Miscellaneous.

General Terms

Design, Human Factors.

Keywords

IHM grand-public, dialogue naturel, multimodalité en sortie, contraintes de présentation, choix du contenu

1. INTRODUCTION

Face au succès de l'informatique grand-public, les attentes des utilisateurs sont plus grandes que jamais. En effet, la diversité des usagers, la multiplication des terminaux et des contextes d'interaction, la masse et la variété des données stockées, entrainent de nouveaux "besoins". Parmi eux, l'accessibilité des informations et la personnalisation des interactions. En IHM, ces défis mobilisent divers champs de recherche, telles que la plasticité, l'adaptativité, la visualisation, etc. Pour nous, cette accessibilité et cette personnalisation passent par le dialogue naturel multimodal.

Le dialogue naturel ne peut se résumer à la communication en langue naturelle. Derrière ce terme, se cachent un certain nombre de propriétés, dont la capacité à générer des réponses coopératives et la flexibilité de l'interaction [6]. Si nous considérons que, lorsque le système communique avec l'utilisateur, la génération de réponses coopératives passe par le choix d'une stratégie de dialogue, la flexibilité de l'interaction passe, elle, par une présentation multimodale de la réponse.

Généralement, le choix de la réponse du système est clairement dissocié de la présentation de cette réponse. Par conséquent, les contraintes de présentation ne peuvent influencer la réaction du système. Par contrainte de présentations, nous entendons, entre autres, les contraintes imposées par l'utilisateur, les contraintes inhérentes aux modalités utilisées et les contraintes liées aux caractéristiques des terminaux. Pourtant, la personnalisation et l'accessibilité devraient passer non seulement par une adaptation de la présentation mais aussi par une adaptation du contenu : elles devraient être permises à la fois par la production de réponses coopératives multimodales et par une prise en compte des contraintes de présentation pour le choix du contenu.

Dans cette optique, nous proposons un composant de choix de stratégie de dialogue qui s'intègre dans une architecture modulaire d'IHM à la ARCH [8]. Avant de décrire ce composant, son fonctionnement et son articulation avec les autres modules logiciels, nous précisons notre positionnement et nos motivations.

2. POSITIONNEMENT ET MOTIVATIONS

Généralement, tout l'enjeu de la multimodalité en sortie est de déterminer quelles modalités doivent être utilisées pour présenter une réponse, et comment ces modalités doivent coopérer entre elles. Pour que ce choix soit possible, il est nécessaire de formaliser la notion de modalité ainsi que les types de multimodalités possibles.

Selon notre approche, une modalité est caractérisée par un couple <d, l>, où d représente un dispositif physique et l un langage d'interaction [2]. Un dispositif physique est tout artefact

permettant au système de "percevoir" un message de l'utilisateur (dispositif physique d'entrée) ou de lui en présenter un (dispositif physique de sortie). Clavier et pointeur sont des dispositifs physiques d'entrée alors qu'écran et haut-parleur sont des dispositifs physiques de sortie. Notons que chaque terminal possède au moins un dispositif physique d'entrée et un de sortie. Un langage d'interaction, quant à lui, est un système conventionnel structuré – composé d'une grammaire et d'éléments terminaux – dont les expressions sont porteuses de sens pour le système. Le langage naturel est un langage d'interaction, au même titre qu'un formulaire de récupération ou de présentation d'informations.

Même dans une approche orientée système de la multimodalité en sortie, l'humain ne peut être négligé. En effet, il est primordial d'identifier les sens (ou modalités *sensorielles*) permettant de percevoir une modalité donnée. Par exemple, si l'on considère la modalité <haut-parleur, langage naturel>, l'ouïe permet à l'utilisateur de percevoir le message. Par contre, dans le cas de la modalité <écran, langage naturel>, le sens humain qui intervient est la vue. Les capacités de perception de l'utilisateur, de par leur dépendance au(x) dispositif(s) physique(s) utilisé(s), doivent d'influencer les choix de(s) modalité(s). Cette contrainte est triple car, aux contraintes de présentation imposées par l'utilisateur et à celles dues aux terminaux, s'ajoutent celles propres aux modalités sélectionnées dont doit tenir compte tout système qui se veut utilisable : on ne peut afficher plus de x items sur un écran de PC ou oraliser plus de y items sans risquer une surcharge cognitive et une désorientation de l'utilisateur.

Notre approche nous fait également prendre en considératio les coopérations entre modalités du point de vue du système. Plus précisément, nous adoptons les propriétés CARE [Nigay, 1997] selon lesquelles : (a) deux modalités peuvent être Complémentaires pour une seule tâche de présentation (b) une modalité peut être Assignée à une tâche de présentation (c) deux modalités peuvent être Redondantes pour une même tâche de présentation (d) deux modalités peuvent être Equivalentes pour une même tâche de présentation. Auquel cas, il doit être décidé si une seule des modalités est utilisée ou si les deux sont utilisées de façon redondante.

Notre positionnement ne concerne pas seulement l'appréhension de la multimodalité, mais également celle de la communication personne-machine. Défendant un dialogue le plus naturel possible, nous mettons particulièrement l'accent sur le comportement coopératif du système. Ce comportement se traduit notamment par des réponses suggestives – quand aucune solution n'est trouvée ou quand le nombre de solutions est trop important – et par des réponses complétives – pour compléter une solution donnée [7]. Le choix des suggestions dans le premier cas et celui des surinformations dans le deuxième dépendent de la stratégie de dialogue adoptée par le système. Pour une requête donnée, nous identifions quatre stratégies de dialogue possibles:

- Stratégie de dialogue 1 – il n'y a pas de solution : le système suggère à l'utilisateur des réponses alternatives ;
- Stratégie de dialogue 2 – il y a une unique solution : le système présente la réponse et d'éventuelles informations supplémentaires ;
- Stratégie de dialogue 3 – il y a quelques solutions : le système présente la liste des réponses possibles sans aucune information supplémentaire ;
- Stratégie de dialogue 4 – il y a trop de solutions : le système suggère à l'utilisateur un critère possible pour restreindre l'ensemble des réponses.

Si le choix de réponses ou de critères de recherche alternatifs ainsi que les informations supplémentaires semblent dépendre en grande partie d'une pertinence relevant des sciences humaines, nous sommes convaincus que, quand il y a plus d'une solution, la distinction entre *quelques* et *trop* de solutions doit être décidée en fonction des modalités utilisées pour présenter la réponse, elles-mêmes dépendantes des contraintes de présentation imposées par l'utilisateur, par les terminaux utilisées et par les spécificités propres aux modalités. Ce qui revient à dire que la présentation doit influencer la réponse du système.

Prenons l'exemple d'un annuaire d'entreprise. Cet annuaire comprend 4 Carole. Si l'utilisateur l'interroge via un système d'information classique et que sa requête porte sur le numéro de téléphone de Carole, le système lui indique les noms et numéros de téléphone de ces 4 Carole. Imaginons que l'utilisateur puisse demander au système d'avoir une réponse orale : idéalement, le système ne devrait pas donner la liste de solutions sous peine de surcharger cognitivement l'utilisateur ; il devrait plutôt lui proposer de préciser sa requête en utilisant un autre critère. Ceci revient à considérer qu'étant donné les contraintes de présentation, il y a trop de solutions, et qu'il faut donc passer de la troisième à la quatrième stratégie de dialogue. Il en va de même si l'utilisateur souhaite une réponse visuelle et que le terminal utilisé est un mobile ne pouvant afficher plus de 3 solutions sans ascenseur.

Cette influence de la présentation sur la stratégie de dialogue a rarement été prise en compte jusqu'à présent, et encore plus rarement mise en œuvre. Pourtant, une meilleure coordination entre présentation et choix du contenu était déjà soulignée dans [4] concernant les présentations multimédia intelligentes. Cette idée est reprise dans le modèle de référence pour les IMMPS (*Intelligent MultiMedia Presentation Systems*) [1] qui considère plusieurs étapes de génération multimedia/multimodale. L'une d'elle est centrée sur le contenu et comprend aussi bien l'affinement des buts de présentation, la sélection du contenu, l'ordonnancement des informations que l'allocation des modalités. Mais ces travaux demeurent conceptuels et le composant que nous présentons par la suite a pour objectif de concrétiser l'influence de la présentation sur le contenu. Le modèle WWHT (*What, Which, How, Then*) [5] vise, quant à lui, à organiser la génération de présentation multimodale selon quatre questions : "que présenter ? Avec quelles modalités ? Comment ? Quelles évolutions ?" Notre travail, qui se concentre sur la question "que présenter ?", est complémentaire, dans la mesure où le contenu est fixé dans la plate-forme WWHT proposée.

Pour pallier au manque en la matière, nous proposons donc un composant de choix de stratégie de dialogue, qui, comme nous l'expliquons dans la partie suivante, s'intègre dans une architecture d'IHM à la ARCH.

3. UN COMPOSANT DE CHOIX DE STRATEGIE DE DIALOGUE

Comme nous venons de l'exposer, notre motivation première est de proposer une architecture permettant de concevoir des systèmes dont la stratégie de dialogue est déterminée en fonction des contraintes de présentation. A cela, s'ajoute le besoin de pouvoir changer rapidement et simplement les stratégies de dialogue. En effet, il n'y a pas d'études ergonomiques portant sur l'adéquation des stratégies aux contraintes de présentation. Aussi les règles proposées doivent-elles être facilement mise en œuvre afin d'effectuer de tels tests. Ceci nous a poussés à choisir une architecture d'IHM à la ARCH plutôt qu'une architecture de dialogue classique : la modularité du méta-modèle ARCH garantit la réutilisabilité de ses composants. Rappelons qu'ARCH distingue, d'un côté, les composants du domaine (CDo) spécifiques à l'application et, de l'autre, les composants de présentation et d'interaction (CPI) pour la gestion de l'IHM. Entre les deux, le composant de dialogue (CD) gère le dialogue entre l'utilisateur et l'application.

La séparation entre le contenu, géré par le CD, et la présentation, du ressort des CPI, est nette dans ce méta-modèle. Or nous souhaitons moduler cette distinction. C'est pourquoi nous proposons l'introduction d'un composant intermédiaire, appelé composant de choix de stratégie de dialogue (CCSD). Comme le montre la figure 1, ce composant fait le pont entre le CD et les CPI. Il poursuit la logique du méta-modèle, car il permet de réduire la trop grande distance entre les CPI, qui sont dépendants des modalités, et le CD, qui en est indépendant – tout comme le composant d'interaction a permis, par le passé, de découper le composant de présentation et d'en diminuer la complexité [3].

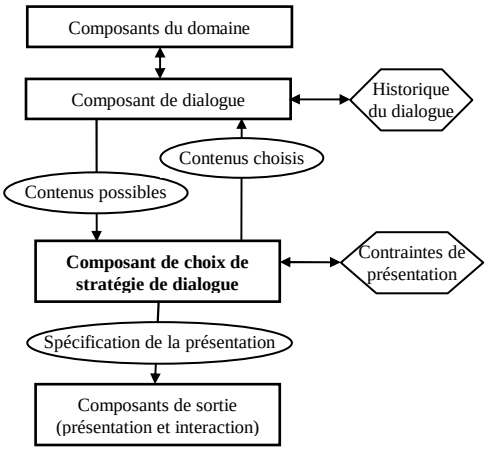

Figure 1. Le composant de choix de stratégie de dialogue dans une architecture basée sur ARCH

Le CCSD doit donc s'intégrer dans des systèmes respectant le méta-modèle ARCH. Pour faire le pont entre le CD et les CPI, il décharge le CD d'une partie de son rôle, et tient compte des informations gérées par les CPI.

Généralement, pour produire la réponse du système, le CD calcule la réaction du système et la transmet aux CPI. Selon notre approche, le CD ne calcule pas la réaction du système dans son ensemble, mais se contente de déterminer toutes les réponses possibles à apporter au système. Idéalement, ces réponses ne se résument pas aux solutions à la requête de l'utilisateur, mais comprennent également les informations supplémentaires possibles, les critères de restriction et les informations permettant un maintien du dialogue. Le CD a donc un rôle de relais à la fois pour les informations des CDo et pour celles relevant de l'historique du dialogue. Dans le cas d'un CD simpliste, le CCSD n'aura aucune décision à prendre et se contentera de transmettre la réponse décidée par le CD aux CPI.

A la différence du CD, le CCSD n'est pas indépendant des modalités, même s'il ne les manipule pas directement. Il comprend un ensemble de règles qui permettent de déterminer la réponse à donner au système, en tenant compte des réponses possibles fournies par le CD, et des contraintes de présentées récupérées par les CPI lors de l'interprétation de l'entrée ou via les terminaux utilisés. Les contraintes de présentation sont mémorisées en dehors du CCSD[1]. Le choix du CCSD résultant de l'analyse des règles est transmis aux CPI sous forme d'une spécification de présentation. Une spécification de présentation est composée d'une ou plusieurs tâches de présentation élémentaires – multimodales ou monomodales – ordonnées. Chaque tâche correspond à une information sémantique unique. L'ordre des tâches peut se traduire différemment lors de la génération de la présentation : si la modalité est perçue auditivement, l'ordre correspond à l'ordre séquentiel d'énonciation ; si elle est perçue visuellement, l'ordre peut correspondre à un ordre d'affichage horizontal, d'onglets, etc.

Parallèlement, le CCSD informe le composant de dialogue des informations sémantiques effectivement présentées, afin de garantir la cohérence de l'historique du dialogue.

La présentation de la réponse à l'utilisateur est du ressort des CPI. Ils sont en mesure de réaliser un certain nombre de tâches de présentation spécifiées lors du développement. Ils sont donc en mesure de concrétiser la présentation spécifiée par le CCSD.

Notre approche nous amène à considérer une multimodalité à deux niveaux. La première, caractérisée par les propriétés CARE, correspond à une coopération de modalités pour une tâche de présentation donnée. La deuxième, d'ordre sémantique, résulte de la réalisation de plusieurs tâches de présentation – tâches qui sont mono ou multimodales. Chaque tâche correspond à une information sémantique et l'ensemble des tâches constitue la réponse du système. Ces deux niveaux de multimodalité sont transparents pour l'utilisateur, mais pas pour le système : alors que la multimodalité classique est traitée au niveau des CPI, la multimodalité sémantique est décidée au niveau du CCSI. Nous avons identifié deux types de multimodalité sémantique pour l'instant : pour un jeu de contraintes de présentations, on parle de tâches ou de spécifications *équivalentes* quand une seule est choisie, et de tâches ou de spécifications *complémentaires* quand les deux sont nécessaires pour construire la réponse du système.

Notre composant permet donc de bien distinguer les étapes suivantes : (1) plusieurs réponses possibles, (2) une réponse choisie en fonction des contraintes de présentation et (3) la génération de la présentation correspondante. Alors que les étapes

[1] Pour l'instant, ces informations sont mémorisées de façon temporaire pour un tour de parole. A terme, les contraintes prises en compte pourraient être élargies, notamment aux préférences générales – explicites ou observées – et aux handicaps de l'utilisateur.

(1) et (2) relèvent du CD dans ARCH, elles sont segmentées dans notre approche. Ainsi, l'étape (2), à la charge du CCSD, correspond très exactement à la couche de contenu dans le modèle de référence pour les IMMPS cité plus haut. Ceci accroît la modularité des systèmes interactifs, et permet de centrer le travail sur le choix de stratégie de dialogue.

Pour valider notre approche, des développements sont en cours de réalisation.

4. VALIDATION LOGICIELLE

Pour nos travaux, nous avons repris un prototype existant, l'annuaire multimodal intelligent d'entreprise @mie. Il permet au personnel d'une entreprise de trouver des informations sur leurs collègues (noms, photos, courriels, numéros de téléphone, bureaux, localisations, équipes), sur les équipes (numéros de fax, noms, acronymes, descriptions) et sur les localisations des sites (villes, pays, plans). Bien que le système puisse être utilisé depuis un ordinateur, nous privilégions son accès depuis un téléphone mobile car les contraintes de présentation qu'il offre sont plus importantes (contraintes du téléphone en lui-même et contraintes dues aux contextes d'utilisation).

Dans le système initial, le choix de la présentation multimodale est figé. L'intégration de notre composant doit permettre de concevoir rapidement des stratégies de dialogue multimodales et coopératives en adéquation avec les contraintes de présentation considérées. Dans un premier temps, nous nous sommes concentrés sur la modalité imposée par l'utilisateur via une exigence sur le sens de perception (*afficher* ou *dire* la réponse), sur la taille de l'écran due à l'utilisation d'un téléphone portable et sur les contraintes inhérentes aux deux modalités utilisées, à savoir la langue naturelle <haut-parleur, langue naturelle> et l'hypertexte <écran mobile, hypertexte+images>. Pour le CD, une partie de ce qui avait été développé pour le prototype a été repris.

Pour exemplifier le déroulement de la génération multimodale, imaginons que l'utilisateur demande au système de lui *dire* le numéro de téléphone de Carole. Le CD informe le CCSD qu'il existe 4 Carole et lui fournit leurs noms, services et numéros de fixes, ainsi que les critères de restriction jugés les plus pertinents, en l'occurrence les noms et services de chaque personne-solution. La présentation devant être orale et le nombre d'informations étant supérieur à trois (les noms des Carole et leurs numéros de fixe), le CCSD utilise la règle selon laquelle le système propose à l'utilisateur de restreindre sa requête. Il informe donc le composant de présentation que les tâches à générer sont les suivantes : (1) informer l'utilisateur du nombre de solutions oralement (2) informer l'utilisateur qu'il peut restreindre selon les critères transmis oralement (3) inviter l'utilisateur à préciser sa requête visuellement (ceci afin de garantir la liberté d'action de l'utilisateur). Ces trois tâches de présentation sont complémentaires d'un point de vue sémantique, mais chacune d'elle est assignée à une modalité de point de vue des propriétés CARE. Les CPI vont dont réaliser chacune des tâches de présentation. Dans notre cas, l'ordre pour les tâches orales est réalisé lors de l'énonciation des informations et il n'est pas pris en compte pour la synchronisation des modalités : les présentations orales et visuelles sont générées parallèlement. Par ailleurs, le CCSD précise au composant de dialogue que l'utilisateur est informé du nombre de solutions et des critères de restriction.

Compte tenu des restrictions de cas pris en compte pour notre validation logicielle, nous avons identifiés quatre conditions de règles (le nombre de solutions, le sens de perception imposé par l'utilisateur, le nombre d'informations présentables pour une modalité, la portée de la requête – i.e. une personne ou une propriété), sept spécifications de présentation possibles et trente deux tâches de présentation.

5. REMERCIEMENTS

Ces travaux sont partiellement financés par le projet européen E-MODE (EUREKA ITEA 04046).

6. CONCLUSION

Au cours de cet article, nous avons traité de la multimodalité en sortie dans le cadre d'un dialogue personne-machine naturel. Cela passe par la production de réponses coopératives tenant compte des contraintes de présentation. Pour produire de telles réponses, nous proposons un composant de choix qui fait le pont entre le composant de dialogue et les composants de présentation et d'interaction classiquement présents dans les IHM.

La validation logicielle de notre approche est en cours, grâce au développement d'une plate-forme intégrant le composant présenté et d'une interface de paramétrage du composant de choix. Par la suite, une collaboration avec un ergonome est prévue afin de tester les stratégies de dialogue implémentées et de travailler sur d'autres règles de choix.

7. REFERENCES

[1] Bordegoni, M., Faconti, G., Feiner, S., Maybury, M. T., Rist, T., Ruggieri, S., Trahanias, P., and Wilson, M. A standard reference model for intelligent multimedia presentation systems. *Computers standards and interfaces, 18,* 6-7 (1997), 477-496.

[2] Nigay, L., and Coutaz, J. A Generic Platform for Addressing the Multimodal Challenge. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI'95)*, ACM Press, 1995, 98-105.

[3] Pfaff G. (Ed.), *User Interface Management Systems.* Springer-Verlag, 1985.

[4] Roth, F., and Hefley, W. E. Intelligent multimedia presentation systems: research and principles. *Intelligent multimedia interfaces*, M. T. Maybury Ed. AAAI Press / The MIT Press, 1993.

[5] Rousseau, C., Bellik, Y., Vernier, F. and Bazalgette, D. A framework for the intelligent multimodal presentation information. *Signal Processing Journal, Special issue on Multimodal Interfaces*, Elsevier, 2006. To appear.

[6] Sadek, D. Design considerations on dialogue systems: from theory to technology - the case of ARTIMIS. In *Proceedings of the ESCA TR Workshop on interactive dialogue for multi-modal systems (IDS'99)*, ISCA Archive, 1999, 173-187.

[7] Siroux, J., Gilloux, M., Guyomard, M., and Sorin, C. Le dialogue homme-machine en langue naturelle : un défi ? *Annales des télécommunications, 44,* 1-2 (1989), 53-76.

[8] The UIMS Tool Developers Workshop. A Metamodel for the Runtime Architecture of an Interactive System. *SIGCHI bulletin, 24*, 1 (1992), 32-37.