

Projecto NavMI

Navegador Multimodal para Imagens

Vitor M. N. Fernandes
Cadeira de Interfaces Multimodais Inteligentes (IMMI – 2004/2005),
Mestrado em Engenharia Informática e de Computadores (MEIC),
Instituto Superior Técnico (IST)
Av. Rovisco Pais, 1000 Lisboa
vmnf@{yahoo.com, mega.ist.utl.pt}

Resumo

Este documento apresenta o projecto Navegador Multimodal para Imagens (NavMI) realizado na cadeira de Interfaces Multimodais Inteligentes. O projecto NavMI consistiu na elaboração de uma aplicação, que permite a utilização de uma interface de comando, sem a necessidade dos utilizadores estarem fixos num determinado local (PC), utilizando o rato e/ou teclado, para navegarem numa imagem 2D de grande dimensão. Este grau de liberdade foi conseguido através da adição, do teste e da validação da utilização da modalidade voz. O projecto desenvolvido seguiu um processo iterativo que começou pela pesquisa de trabalhos relacionados já existentes, seguido da análise de tarefas, do desenvolvimento de vários protótipos (um não funcional seguido de dois funcionais) e que culminaram na apresentação da versão final. Nesta verificou-se que a modalidade voz é válida para a emissão de comandos e que a taxa de erros era menor do que o limite aceitável em média pelos utilizadores. Os pontos mais fortes são a robustez e flexibilidade na emissão de comandos voz na utilização da interface.

Palavras-chave

Interfaces multimodais, modalidade voz, gramáticas para comandos por voz, robustez e flexibilidade na emissão de comandos por voz, navegação em imagens 2D, visualização de imagens 2D.

1. INTRODUÇÃO

As interfaces WIMP (*Windows, Icons, Menus e Pointers* ou *Windows, Icons, Mice e Pull-down menus*) são o estilo de interacção mais difundido sendo utilizado nos sistemas Windows para PCs, MacOS nos Apple Macintosh e em vários sistemas baseados no X Windows para Unix [Kimani05]. O cenário estudado foi o de adicionar outra modalidade à interface WIMP [Taylor97] e com esta permitir executar todos os comandos que se poderiam realizar de forma convencional com o rato e/ou teclado. A modalidade adicional seleccionada foi a voz ficando o projecto com a possibilidade de serem emitidos comandos por teclado, rato ou voz. O feedback da aplicação será visual na interface gráfica bem como através da utilização de voz sintetizada.

Esta selecção tem por objectivo validar a utilização do ASR (*Automated Speech Recognition*) e o TTS (*Text To Speech*) da Loquendo para uma interface de comando como por exemplo na sala multimédia do campus IST do Taguspark.

Numa primeira fase foi realizada pesquisa sobre as aplicações que realizassem tarefas semelhantes àquelas a que o projecto se propunha. (Ver ponto “Trabalho relacionado”).

A abordagem seguida na realização do projecto foi a iterativa com o ciclo pesquisa, implementação e avaliação como cerne. Estes passos foram realizados para as várias fases: elaboração do protótipo não funcional, 1.º e 2.º protótipo funcional e versão final do projecto. (Informação detalhada na secção “Abordagem” e “Avaliação”).

Os resultados obtidos demonstram a validade da utilização da modalidade voz na interface de comando com o software da Loquendo. (Consultar a secção “Resultados e Discussão”).

2. TRABALHO RELACIONADO

A pesquisa iniciou-se sobre aplicações que permitissem a visualização de imagens para introdução às tarefas a realizar. De seguida apresento trabalhos que utilizam a modalidade voz na interface WIMP. Apresento também dois trabalhos que considerei muito interessantes para desenho / *feedback* da interface por som e comando não discreto da interface de comando por voz (teclado vs. *joystick*).

2.1 Aplicações de visualização de imagens

2.1.1 *Paint Shop Pro*

Aplicação de edição de imagem 2D. Considerando apenas a navegação, é possível deslocar na imagem para a

direita, cima, baixo e esquerda através do clicar e arrastar na imagem, barras de deslocamento e setas do teclado.

2.1.2 Google Earth

Aplicação de visualização de imagem 2D (plana ou assente em esfera) que permite visualizar imagens de satélite em qualquer ponto do planeta. Considerando apenas a navegação, é possível deslocar na imagem para a direita, cima, baixo e esquerda através do clicar e arrastar na imagem, barras de deslocamento e setas do teclado. É também possível visualizar a informação em perspectiva e controlar o ângulo desta.

2.2 Interfaces pós-WIMP

2.2.1 Put-that-there

Parece-me que será impossível realizar um projecto que use uma modalidade não convencional WIMP sem falar do put-that-there [Bolt80].

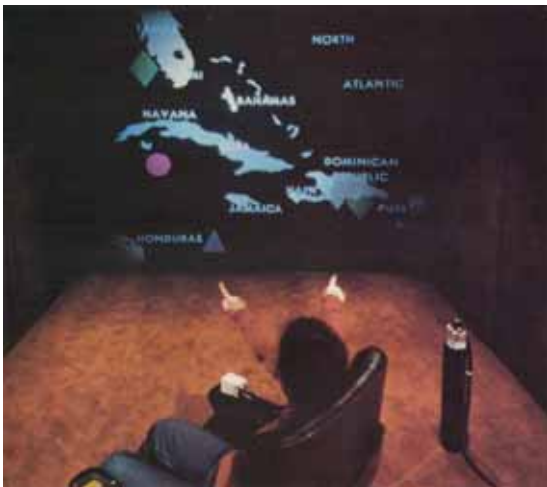


Imagem 1: Put-that-there: media room

Este trabalho é a referência sobre o que é possível realizar sem estar agarrado a um teclado e a um rato. Outro trabalho semelhante foi o Put-that-where? [Billinghamurst98].

2.3 Modalidade voz

2.3.1 Auditory workshop

Será possível desenhar uma interface apenas com som? Este é o desafio do auditory workshop [Kaltenbrunner05] que cobre as áreas de desenho da interface e feedback por som / voz. Sem dúvida uma interface a explorar para pessoas com limitações visuais!

2.3.2 The Vocal Joystick

Trabalho muito interessante realizado na Universidade de Washington em que desenvolveram uma interface de comando por voz para pessoas com dificuldades motoras [Bilmes05]. Os comandos não são discretos (direita, cima) mas são como se fossem realizados com um joystick! Os comandos não são dados por fala (ineficiente para controlo contínuo) mas sim através de vocalizações. (Este tipo de abordagem foi sugerida por Filipe Dias – Grupo IMMI / INESC-ID no inquérito de análise de tarefas!)

3. ABORDAGEM

3.1 Análise de tarefas

O primeiro passo da realização do projecto consistiu na análise de tarefas. Esta análise permite na fase inicial ir de encontro às necessidades dos utilizadores [Sullivan90]. É também com esta análise que se pretende dar resposta a um conjunto de questões como: “Quem vai utilizar o sistema?”, “Quais tarefas executam actualmente?”,... [Baseline00]. Com o objectivo de dar resposta a estas questões foi realizado um inquérito por questionário. Neste estabeleceu-se o perfil de potenciais utilizadores, quais as modalidades com que estes estavam habituados a interagir e como é que realizariam essa interacção.

3.1.1 O inquérito de análise de tarefas

O inquérito por questionário foi organizado em 6 partes distintas: i) dados pessoais – obtenção de dados demográficos, ii) hábitos na utilização do computador, iii) modalidades que usa na interacção com o seu computador, iv) está perante uma aplicação que lhe permite navegar em parte de uma imagem de alta resolução – simulação de utilização de uma potencial aplicação, v) selecção da(s) modalidade(s) mais indicada(s) face ao comando a realizar – analisar a sensibilidade dos utilizadores perante várias modalidades e vi) a interface da aplicação NavMI – indicações para um ponto de partida com vista à elaboração do protótipo de baixa fidelidade. A ferramenta utilizada foi o php Easy Survey Package (phpESP em <http://phpesp.sourceforge.net/>) que permite de forma rápida e simples a elaboração de questionários online. O questionário encerrou com 21 respostas de potenciais utilizadores válidas.

3.1.1.1 Dados pessoais

Os entrevistados foram maioritariamente do género masculino (81,0%), com idades que variaram entre os 20 e os 49 anos (20-24: 23,8%, 25-29: 19,0%, 30-39: 52,4% e 40-49: 5,8%) e a maior parte tem como nível de instrução o ensino superior (61,9%) ou pós-graduação (33,3%).

3.1.1.2 Hábitos na utilização do computador

A maior parte dos utilizadores possuem computador pessoal fixo (95,2%) e/ou portátil (90,5%). Aproximadamente um em cada quatro possui um computador de mão (23,8%). Apenas 1 utilizador (4,8%) referiu utilizar o computador entre 5 a 6 dias por semana tendo os restantes respondido que o utilizavam todos os dias (95,2%).

3.1.1.3 Modalidades que usa na interacção com o seu computador

Todos os utilizadores referiram já ter utilizado o teclado e rato na interacção e destes aproximadamente metade já utilizou também a modalidade voz (52,4%). A confiança na modalidade teclado e rato é muito elevada (6,0 e 5,8 numa escala de 1 a 6) e para a modalidade voz muito baixa (1,7 na mesma escala). Foram ainda referidos outros meios de interacção baseados também no apontar (rato) que foi o *stylus* do PDA (14,3%) e o *joystick* (4,8%) sendo este último considerado apenas para jogos.

3.1.1.4 *Está perante uma aplicação que lhe permite navegar em parte de uma imagem de alta resolução*

Foi pedido aos utilizadores que atribuissem uma escala de grandeza à possibilidade de realização dos seguintes comandos: 1 - navegação relativa (cima, baixo, esquerda ou direita), 2 - navegação para os extremos (tudo para cima, tudo para baixo,...), 3 - zoom de valor em valor fixo seguido (100%, 90%, 80%,...), 4 - zoom para um determinado valor (35%), 5 - apontar para um ponto (x,y), 6 - atribuir um nome a um ponto (x,y = "nome"), 7 - referenciar / centrar num ponto sem nome fora da área visível e 8 - referenciar / centrar num ponto com nome fora da área visível.

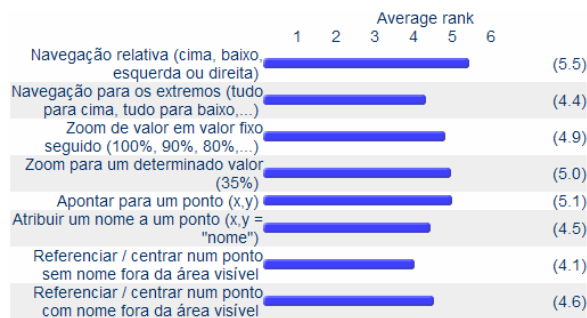


Gráfico 1: Possibilidade da realização de determinados comandos

Numa escala de 1 a 6 verificou-se que os comandos mais passíveis de serem realizados serão a navegação relativa (cima, baixo,...) com 5,5 em 6, a utilização do zoom com 4,9 e 5,0 e o apontar para um ponto x,y.

3.1.1.5 *Seleção da(s) modalidade(s) mais indicada(s) face ao comando a realizar*

Sobre os comandos do ponto anterior (Comando 1 a 8) foi questionada qual a modalidade que utilizaria para a sua execução (de entre teclado, rato e voz) numa escala de 1 a 6.

| | Teclado | Rato | Voz |
|-----------|------------|------------|-----|
| Comando 1 | 4,4 | 5,3 | 3,8 |
| Comando 2 | 5,0 | 4,5 | 4,0 |
| Comando 3 | 4,8 | 5,1 | 3,9 |
| Comando 4 | 5,0 | 3,5 | 3,8 |
| Comando 5 | 3,2 | 5,7 | 3,0 |
| Comando 6 | 5,2 | 2,6 | 4,5 |
| Comando 7 | 4,2 | 4,3 | 3,6 |
| Comando 8 | 4,4 | 4,2 | 4,3 |
| Média | 4,5 | 4,4 | 3,9 |

Tabela 1: Comando vs. modalidade teclado, rato e voz (a negrito a modalidade mais aceite, a normal a intermédia e a cinza a menos aceite)

Verificou-se que as modalidades preferidas são o teclado e o rato em detrimento da voz. Esta preferência está certamente também relacionada com o facto de todos os utilizadores utilizarem o rato e/ou teclado e apenas 52,4% utilizarem a voz

A desconfiança em relação à modalidade voz é bem clara apenas conquistando apenas 3 segundos lugares para os comandos 4 (zoom para um determinado valor), 6 (atri-

buir um nome a um ponto) e 8 (referenciar / centrar num ponto com nome fora da área visível)

Numa questão de resposta aberta foram pedidos outros comandos que em potencial poderiam ser executados. Foram dadas várias sugestões singulares (apenas uma resposta para cada): centrar num ponto, centrar em dois ou mais pontos, rotação, inversão, definir regiões, navegação através de arrastamento, vista em perspectiva (para a frente e para trás) e anotações ou outros objectos sobre a imagem (setas ou X).

Foi também perguntado caso apenas tivesse duas ou uma modalidades das três (teclado, rato e voz) quais a que utilizaria.

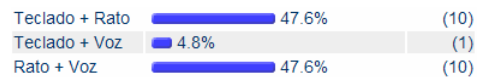


Gráfico 2: Apenas duas modalidades disponíveis

No caso de apenas poder utilizar duas a preferência caiu sobre teclado + rato ou rato + voz ambas com 47,6%. A possibilidade de ficar limitado ao teclado + voz apenas recolheu a preferência de uma pessoa (4,8%).

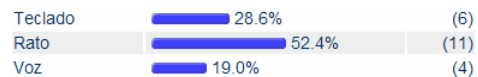


Gráfico 3: Uma única modalidade disponível

Para a situação limite de apenas uma modalidade foi preferido o rato por maioria (52,4%) ficando o teclado com 28,6% e a voz com 19,0%. Da observação destes resultados conclui-se que a modalidade apontar através da utilização do rato é muito importante para todos os utilizadores!

3.1.1.6 *A interface da aplicação NavMI*

De elevada importância é ter uma expectativa do que os utilizadores considerariam uma boa interface para a aplicação em desenvolvimento. Foram com este objectivo colocadas 3 questões. A primeira pretendeu descobrir o que seria mais importante na interface gráfica (menus, botões), a segunda quais as taxas de erro consideradas aceitáveis no desempenho das tarefas e uma terceira de exploração de outras possibilidades de utilização da modalidade apontar sem ser através do rato (esta tinha o objectivo de coordenar a elaboração do projecto com o gesto).

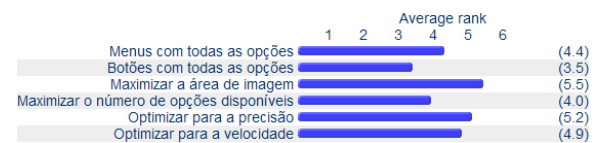


Gráfico 4: O que deve privilegiar a interface gráfica

Os utilizadores responderam que pretendem uma interface com a maximização da área visível de imagem e que seja precisa (reduzido número de erros).

Os utilizadores não são intransigentes em relação à margem de erro da modalidade voz e estão disponíveis para aceitar um nível mais elevado de erro do que com o teclado e/ou rato.

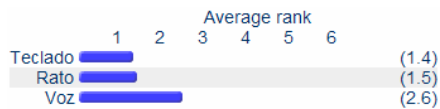


Gráfico 5: Taxa de erros considerada aceitável (1: <5% (muito bom), 2: 5%-9%, 3: 10%-14%, 4: 15%-19%, 5: 20%-24%, 6: > 24% (mau))

A taxa de erro que os utilizadores consideraram aceitável para a modalidade voz foi de 11,4%¹ o que parece um valor aceitável (≈ 1 erro por cada 9 comandos).



Gráfico 6: Aceitaria uma menor qualidade / precisão na interação com a aplicação caso tivesse a flexibilidade de em vez de usar o rato pudesse a) apontar com o braço / dedo ou b) clicar com gestos da mão (1: claro que não a 6: claro que sim)

A aceitação de uma menor qualidade / precisão na modalidade apontar através da substituição do rato pelo braço / dedo ou mão não foi clara. Trocar o rato pelo gesto não parece ser pacífico para os utilizadores questionados.

3.1.2 Resposta às 11 questões da análise de tarefas

3.1.2.1 Quem vai utilizar o sistema?

Os utilizadores do navegador trabalham com desenhos ou imagens de grande dimensão, seja no formato papel, seja no formato digital.

3.1.2.2 Quais tarefas executam actualmente?

Percorrem desenhos ou imagens nas várias direcções através de navegação relativa (cima, baixo, esquerda ou direita) ou navegação para os extremos (tudo para cima,...); ampliam / reduzem a resolução / o tamanho de imagens; apontam para pontos ou áreas e identificam-nos e referenciam pontos em vários locais.

3.1.2.3 Quais as tarefas desejáveis?

Digitalização de todo o documento papel; realização de todas as operações anteriores agora num sistema informático largando o papel; utilização da modalidade voz na realização das diferentes tarefas.

3.1.2.4 Como aprender as tarefas?

Os utilizadores deverão ter os conhecimentos mínimos na utilização do computador, teclado e rato. A aprendizagem da modalidade voz deve ser o mais natural possível no entanto o vocabulário será minimizado de modo a aumentar a precisão e rapidez. O texto correspondente aos comandos voz deve estar disponível para informar o utilizador.

3.1.2.5 Como são desempenhadas as tarefas?

As tarefas são desempenhadas recorrendo a uma aplicação informática que apresenta a imagem a navegar podendo o utilizador 3 modalidades diferentes de comando. Necessitará de ter além de teclado e rato um microfone para a utilização da modalidade voz. Caso

¹ Os utilizadores pontuaram a margem com 2,6 sendo 2 = 5-9% e 3 = 10-14* e considerando os limites inferior 5% (para 2) e superior 14% (para 3) obtemos que o resultado da taxa de erro é $TE_{2,6} = (14\% + 5\%) * (2,6 - 2) = 11,4\%$.

pretenda ter feedback auditivo necessitará também de ter colunas.

3.1.2.6 Qual a relação entre os utilizadores e a informação?

Os dados são imagens de alta resolução (previamente digitalizadas). Não há restrições a nível dos visionamentos ou consultas desde que seja um utilizador com possibilidade de executar a aplicação.

3.1.2.7 Que outros instrumentos possui o utilizador?

O utilizador pode digitalizar e carregar novas imagens no navegador de imagem. O utilizador poderá usar um computador fixo, portátil, PDA ou o sistema presente na Sala Multimédia do Taguspark.

3.1.2.8 Como comunicam os utilizadores?

A aplicação pode ser usada concorrencialmente desde que os vários utilizadores não utilizem a mesma modalidade de comando e ao mesmo tempo. Os utilizadores estarão no mesmo local visionando a mesma imagem, pelo que podem comunicar falando.

3.1.2.9 Qual a frequência de desempenho das tarefas?

As tarefas por ordem de frequência serão: navegação relativa (cima, baixo, esquerda ou direita), apontar para um ponto (x,y), zoom para um determinado valor (35%), zoom de valor em valor fixo seguido (100%, 90%, 80%,...), referenciar / centrar num ponto com nome fora da área visível. A frequência de utilização é elevada sendo realizadas várias tarefas elementares por minuto.

3.1.2.10 Quais as restrições de tempo impostas?

As restrições de tempo colocam-se no redesenho da imagem após um comando e também ao nível do reconhecimento da voz.

3.1.2.11 Que acontece se algo correr mal?

Caso uma modalidade falhe deverá ser possível continuar a trabalhar com as duas restantes. Caso o reconhecedor de voz tenha um baixo nível de performance recorrer à introdução dos comandos via teclado e/ou rato.

3.2 Elaboração do protótipo de baixa fidelidade

O protótipo de baixa fidelidade foi elaborado tendo em conta a recolha de requisitos dos utilizadores no inquérito por questionário elaborado na análise de tarefas.

No protótipo destaca-se a maximização da área de imagem através da sobreposição dos botões principais de comando sobre esta. Os botões disponíveis possibilitam a navegação (esquerda, direita,...) e a alteração do nível de zoom. Todas as três modalidades teclado, rato e voz estão prontas a receber comandos.

Para testar o protótipo de baixa fidelidade foram também criados para este cenários de utilização.

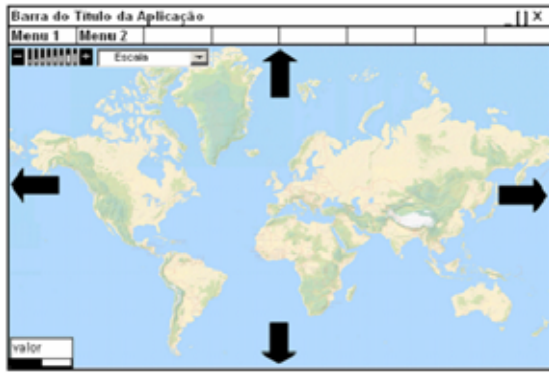


Imagem 2: Protótipo de baixa fidelidade

3.2.1 Cenários de utilização

O cenário de utilização contempla uma sessão de trabalho em que se pretende visualizar diferentes países / cidades da Europa.

3.2.1.1 Tarefa fácil

Visualizar a Europa. A tarefa consiste em na vista todo o mundo “apontar e clicar” sobre a Europa. A Tarefa poderá ser realizada dizendo “Europa” ou clicando sobre a Europa. Espera-se a concretização da tarefa em até 10 segundos e sem erros.



Imagem 3: Imagem após a realização da tarefa fácil

3.2.1.2 Tarefa média

Principais cidades da Península Ibérica. A tarefa consiste em colocar visíveis todas as principais cidades da Península Ibérica. Espera-se a concretização da tarefa em até 20 segundos e sem erros.



Imagem 4: Imagem após a realização da tarefa média

3.2.1.3 Tarefa difícil

Localizar o campus da Alameda do IST. A tarefa consiste em localizar o campus da Alameda do IST em Lisboa. Espera-se a concretização da tarefa em até 2 minutos e com 2 ou menos erros.



Imagem 5: Imagem após a realização da tarefa difícil

3.3 Captura de imagens

A criação das imagens de alta resolução resultou da combinação de várias imagens [Multimap05]. Uma com uma resolução de 3795x2645 (imagem base 1) e outra com o dobro 7590x5290 (imagem base 2). A primeira resultou de 103 e a segunda de 248 imagens individuais no formato GIF. Ambas as imagens foram guardadas no formato BMP a 256 cores com compressão RLE (Run Length Encoding) de modo a reduzir o tamanho dos ficheiros.

Os vários níveis de zoom foram obtidos por *downsampling* das imagens base. As outras 9 imagens, correspondentes a todos os níveis de zoom entre 10% e 90% com saltos de 10 em 10, foram geradas a partir das duas imagens base. O processo, para cada imagem, consistiu em aumentar o número de cores das imagens base para 16M, reduzir a imagem para x% utilizando *bicubic resize* e de seguida foram gravadas em formato BMP com compressão RLE (no processo o número de cores regressa a 256 cores).

(A redução do número de cores a 256 – 1 Byte / pixel - reduz para um terço o tamanho da imagem comparado com a mesma imagem a True Color – 3 Byte / pixel. O formato RLE reduziu também o tamanho da imagem em entre 50 a 70%.)

3.4 Primeiro protótipo funcional

A introdução da modalidade voz foi conseguida através da obtenção de uma licença de teste - 90 dias - através do Professor Joaquim Jorge (IST/INESC-ID) e John Rodrigues (INESC-INOV) para o ASR e TTS da Loquendo [Loquendo05].

O primeiro protótipo funcional foi implementado a partir da aplicação ASR Demo presente no SDK do ASR. A linguagem utilizada no desenvolvimento foi o C++. Neste protótipo já estão disponíveis as três modalidades (teclado, rato e voz). As operações disponíveis contemplam a deslocação relativa da imagem (direita, cima, baixo e esquerda) e a deslocação absoluta da imagem (tudo direita, tudo cima, tudo baixo e tudo esquerda). Todas as

operações estão disponíveis através de comandos através de qualquer modalidade.

A interface gráfica pretendeu já dar resposta aos principais requisitos dos utilizadores como seja a maximização da zona da imagem a navegar e a presença dos botões de comando que podem ser clicados com o dispositivo apontador - o rato - ou acedidos através de atalhos no teclado.



Imagem 6: A interface do 1.º protótipo funcional

Um dos principais objectivos neste primeiro protótipo funcional foi o de validar a modalidade voz e verificar que estava dentro do limiar de erro aceitável pelos utilizadores (11,4%).

Na modalidade voz, a gramática utilizada contempla uma versão muito simplificada dos comandos (apenas três a quatro variantes para cada comando tipo), como por exemplo para “direita”:

```
<direita>=(direita|"para_a_direita"|este)
```

A interface apenas aceitava um comando voz de cada vez através de um botão “push to talk”.

3.5 Segundo protótipo funcional

Após os testes realizados sobre o primeiro protótipo funcional (resultados no ponto “Avaliação”) verificou-se que a gramática utilizada era muito pouco flexível, pois apenas tinha um número reduzido de comandos e para cada comando um número muito baixo de possibilidades. Este facto conduziu a uma interface de comando por voz com um tempo de aprendizagem mais elevado e não natural.

As principais evoluções para este segundo protótipo funcional foram a flexibilização da gramática. Para atingir este objectivo foi realizado um questionário em os utilizadores responderam o que diriam para realizar uma determinada acção (as acções foram apresentadas em imagens de modo a não influenciar os utilizadores nas expressões!).

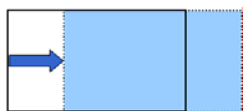


Imagem 7: Qual o comando voz que utilizaria para mudar a área de imagem visível (rectângulo com o limite contínuo) para a nova área de imagem pretendida (rectângulo com o tom azul)?

Com este questionário verificou-se que os utilizadores utilizavam naturalmente prefixos antes do comando como seja a expressão: “deslocar para a direita”. Mas também dizem “para a direita” ou apenas “direita”! Apresento de

seguida a gramática para o comando “direita” que suporta 120 variantes²:

```
<direita>=([<pref_desloca>][<pref_para>]
(direita|éste|leste))
<pref_desloca>=(desloca|deslocar|move|mover|ir|
anda|andar|ver|visualizar)
<pref_para>=(para|para a|para u);
```

Este estudo foi realizado para os comandos de deslocação relativa, absoluta e para o zoom (in / out / valor%) que foi adicionado neste segundo protótipo. Da gramática inicial (1.º protótipo funcional) com menos de 40 variantes chegou-se agora a uma gramática com mais de 2400 possibilidades. A modalidade voz funciona agora em modo contínuo após activação e pára quando o utilizador emite comando “pára reconhecedor”.



Imagem 8: A interface do 2.º protótipo funcional

Além da funcionalidade zoom foi também adicionada um *viewfinder* de modo a permitir ao utilizador saber rapidamente em que local da imagem de grande resolução se encontra. A interface ocupa também, agora, a totalidade do ecrã.



Imagem 9: O *viewfinder*

3.6 Versão final do projecto

Na versão final do projecto a ênfase recaiu na robustez da modalidade voz. O objectivo é reduzir o número de falsos positivos no modo contínuo. O facto do estado do sistema estar oculto (estou a reconhecer, estou a proces-

² O valor resulta de 9 variantes do prefixo desloca ou a sua ausência (10), 3 variantes do prefixo para ou a sua ausência (4) e três variantes de direita (3) em que 10x4x3=120.

Para a avaliação foram utilizadas várias técnicas como os inquéritos por questionário, as entrevistas estruturadas, testes com utilizadores e a realização de vários tipos de tarefas, em cenários de utilização, desde as simples até às mais complexas.

De seguida apresento o protocolo de realização dos vários tipos de testes de avaliação realizados.

4.1 Inquéritos por questionário

Os questionários foram realizados on-line de forma a rapidamente poderem ser publicados e os seus dados recolhidos. A ferramenta utilizada foi o phpESP como referido na secção “Abordagem”. Esta ferramenta permite a colocação on-line dos questionários, o seu pré-teste, realização e visualização de dados. A população alvo pode facilmente ser escolhida através da publicação através de correio electrónico com custos muito reduzidos! A realização de gráficos fica também facilitada pois estes são gerados automaticamente no modo de visualização.

Os inquéritos foram construídos com questões de resposta fechada do tipo sim / não ou escala de valores bem como com resposta aberta. A resposta fechada tem por objectivo validar e/ou medir opções concretas e a resposta aberta permite descobrir novas opções e/ou ideias para concretização.

Os questionários foram utilizados durante o processo da análise de tarefas (protótipo não funcional) e na refinação da gramática em utilização (2.º protótipo funcional). Na fase inicial a sua realização é imprescindível e da 2.ª vez que foi realizado permitiu aferir outras possibilidades não consideradas inicialmente para a gramática voz.

4.2 Entrevistas estruturadas

As entrevistas permitem um contacto directo com os utilizadores e sondar quais as suas opiniões. Têm a virtude de, se soubermos ouvir os utilizadores, descobrir falhas e/ou elementos a adicionar sobre os quais nem pensamos.

Estas necessitam de uma programação mais elevada do que os questionários on-line pois tem de ser realizadas num espaço físico com condições aceitáveis e em que exista possibilidade de utilizar a aplicação.

Ao longo de todo o projecto foram realizadas várias que se distribuíram por salas de aula, gabinetes entre outros espaços. Para a sua execução recomendo ouvir, ouvir, ouvir e claro, como não poderia deixar de ser, escrever / registar o que se ouve (a gravação das entrevistas também pode ser uma mais valia mas é necessário despende um tempo superior).

4.3 Testes com utilizadores

Os testes implicam que a aplicação já funcione. Foram neste projecto realizados sobre os dois protótipos funcionais e permitiram validar de forma clara a modalidade voz.

Para a realização destes testes, com uma modalidade como a voz, é necessário criar um conjunto de condições favoráveis para a sua execução. O ambiente onde decorreram estava longe de ser o ideal, pois tinha ruído de fundo em ambas as ocasiões com outras conversas a

decorrer. A utilização de um auricular, com um microfone de baixa fidelidade e pouco sensível, foi também um factor que pode ter influenciado positivamente os testes. (A utilização de um microfone de elevada qualidade levaria certamente à necessidade de filtrar a banda voz e eventualmente atenuar o ruído existente.)

4.4 Tarefas em cenários de utilização

Os cenários de utilização foram testados duas vezes, uma no início do projecto e outra no final. Menos do que dois testes deste tipo não permitem verificar se o que nos propusemos fazer funciona ou se o que ficou feito funciona ou não também.

Pelo menos para um conjunto mínimo de tarefas deve ser efectuado o teste. De modo a ter um espectro mais largo do tipo de actividades que o utilizador pode realizar foram escolhidas três tarefas, uma fácil, outra de dificuldade média e outra difícil.

Durante a realização dos testes existem duas figuras: a do observador e a do utilizador avaliador. A realização deve seguir um processo com três fases: i) o observador apresenta ao utilizador os objectivos do teste e a aplicação em que este os vai executar (5-10m), ii) o utilizador realiza as três tarefas estando o observador sem ajudar o utilizador mas tomando notas da actividade realizada, tempos de realização e número de erros cometidos e iii) a fase final em que o observador pergunta qual o grau de satisfação do utilizador e toma nota de outras observações que considerar relevante.

5. RESULTADOS E DISCUSSÃO

Os resultados e respectiva discussão são apresentados divididos ao longo das várias fases nas quais se desenvolveu o projecto.

5.1 Protótipo não funcional

A análise de tarefas foi o primeiro passo. Esta serviu para efectuar o levantamento dos principais requisitos e construção do protótipo não funcional. A sua avaliação foi efectuada através de pequenas entrevistas estruturadas com utilizadores em que estes realizaram um conjunto de três tarefas (simples, média e difícil). Antes do início da entrevista foram apresentados vários ecrãs dos estados possíveis da interface para habituação. Os resultados estiveram todos dentro dos limites apresentados na secção “Análise de tarefas”. Para a tarefa simples o tempo médio de realização foi de 6,3 segundos (máx. 10) com 0 erros. Para a tarefa média foi de 12,5 (máx. 20) segundo com 0 erros. Para a tarefa difícil foi de 28,8 (máx. 120) com 0,5 erros por execução (máx. 2).

| | Tempo [s] | Erros [por execução] |
|----------------|-----------|----------------------|
| Tarefa simples | 6,3 | 0,0 |
| Tarefa média | 12,5 | 0,0 |
| Tarefa difícil | 28,8 | 0,5 |

Tabela 2: Protótipo não funcional - Tarefas vs. tempo / número de erros

A clara indicação dos comandos disponíveis bem como a fase de habituação à interface permitiram a obtenção dos bons resultados da tabela anterior. Os tempos bem abaixo dos máximos esperados se devem em parte a apenas exis-

tirem os ecrãs que levavam à concretização das várias tarefas. O baixo número de erros tem a mesma justificação. Participaram nesta avaliação 4 potenciais utilizadores⁴.

5.2 Primeiro protótipo funcional

O primeiro protótipo funcional foi avaliado principalmente sobre a modalidade voz embora também tenham sido recolhidos dados relativos a críticas sobre a interface.

Apresento de seguida as condições de realização do teste: 10 utilizadores, o teste consistiu em reconhecer 22 dos 117 comandos possíveis presentes na gramática direita (3), esquerda (3), cima (4), baixo (4) e zoom (8/103) e foi realizado numa sala de aula com ruído / conversa em fundo. O microfone utilizado foi um microfone de mesa de baixa / reduzida qualidade (um dos mais baratos). O sistema onde o teste esteve a ser conduzido foi um Pentium 4 a 1,6GHz com 256MB de memória. Os comandos a emitir estavam escritos (cada comando foi sempre emitido com o mesmo conteúdo) numa folha entregue aos utilizadores (deste modo estamos num ambiente mais controlado para a realização da avaliação). Para cada comando foi registada a taxa de confiança e o número de erros detectados.

A taxa de confiança mínima foi de 40,5% e a máxima de 89,4% (para reconhecimento correcto). Foram detectados 12 erros em 220 comandos emitidos a que corresponde uma taxa de 5,5% que ficou dentro da expectativa dos utilizadores na análise de tarefas (< 11,4%). O número de erros repetidos num mesmo comando ocorreram no “este” (geográfico) e a maior parte no “zoom” (palavra inglesa). Estes erros surgem da diferença de som entre o este (pronome) vs. este (geográfico) e zoom (português: “zóm”) vs. zoom (inglês: “zume”).

| | |
|--------------------|----------------------|
| este → éste | leste → léste |
| oeste → ôeste, | zoom → zume |
| zoom in → zume ine | zoom out → zume aute |

Tabela 3: Alterações na gramática

Com estes resultados e após as alterações na gramática (imagem anterior) a taxa de erros para os comandos citados será reduzida, tendo-se a expectativa da obtenção do valor inferior a 5% para a taxa de erro⁵.

Além deste teste foi também apresentada a aplicação a outros 3 utilizadores⁶ e em que foram recolhidas várias críticas: a) não ter de pressionar sempre “reconhecer comando por voz” → a aplicação passará a estar em modo de reconhecimento de múltiplos comandos voz (podendo este modo ser desligado e ligado); b) aumentar o número de “sinónimos” para os comandos na gramática

⁴ 4 docentes do ensino secundário.

⁵ Nova taxa de erro será $1 - (220 - 12 + 5) / 220 = 3,2\% < 5\%$. (220 – número total, 12 número de erros inicial, 5 número de erros potencialmente retirados)

⁶ 3 elementos do grupo IMMI: Joaquim Jorge, Frederico Figueiredo e André Martins.

pois os utilizadores descobrem muitas maneiras de fazer a mesma coisa → realizar um segundo inquérito a utilizadores de modo a descobrir o que comando emitem (questões abertas) para fazer as tarefas; c) reduzir o número de falso positivos em modo reconhecimento numa sala onde exista conversação e não apenas comando é importante “explicar” à aplicação quando está perante um comando → foi sugerida a utilização de uma palavra chave para entrar no modo de realizar comando – estilo “Simon says...”; d) ter a possibilidade de encadear comandos com base no contexto → direita, mais, mais... / zoom in, menos, mais.

5.3 Segundo protótipo funcional

Foi realizado um teste semelhante ao descrito no ponto anterior agora com 4 utilizadores que emitiram também um comando de cada tipo (Direita, cima, baixo, esquerda, tudo direita, tudo cima, tudo baixo, tudo esquerda, zoom in, zoom out, zoom 10%, zoom 50% e zoom 100%). Apenas surgiram erros na detecção dos comandos oeste e este. Foram testadas várias possibilidades mas devido às suas semelhanças e após teste a várias escritas fonéticas optou-se por *éste|léste|ôeste*. A média de confiança foi de 75,0% com todos os reconhecimentos acima dos 60%. A este facto não foi estranha a realização de outras correcções fonéticas como “para o leste” estar na gramática como “para u léste”. Foram também realizadas entrevistas com os mesmos utilizadores com vista à obtenção de críticas de forma a melhorar a versão final do projecto. As principais foram: para a modalidade voz: número de falsos positivos / “sem feedback” através do TTS, botão para activar / desactivar o reconhecimento contínuo; na modalidade apontar: não suportar zoom in / zoom out com o rato, não suportar o arrastar (várias problemas com a eficiência); para a modalidade teclado: suportada através do T invertido IJKL em vez das setas de direcção e para a interface gráfica: a colocação dos botões de deslocação, ainda ter uma interface com “debug” e o espaço “reduzido” da imagem a ser visualizada.

5.4 Versão final do projecto

Devido à dimensão do trabalho (interface + modalidade utilizadas) e o pouco tempo disponível foi necessário estabelecer prioridades. Na versão final a prioridade foi tornar o reconhecedor robusto, mostrar o estado do sistema e fornecer ajuda e documentação como referido no ponto abordagem.

Foram entrevistados 4 utilizadores que compararam a interface gráfica do 2.º protótipo funcional com a versão final e estes validaram a 100% a nova colocação dos botões direccionais, zoom, ajuda e sair. Sobre a ajuda estes referiram que esta poderia estar mais completa mas era suficiente. Destacaram o facto da ajuda ser passo-a-passo o que facilita a primeira utilização da aplicação. O semáforo foi também entendido como clara a sua informação.

Foram realizadas duas experiências (uma mais repetição) com dois utilizadores sobre a utilização dos dois submodos no modo de reconhecimento contínuo. Esta teve por

objectivo verificar a fiabilidade dos submodos. Verificou-se que com a experiência nunca ocorreu um falso positivo! Quanto ao protocolo utilizado: pediu-se aos utilizadores que falassem o que entendessem no submodo 1. Verificou-se também que a pontuação de confiança era maior no submodo 1 do que no submodo 2 (mais texto => maior confiança). Os utilizadores têm que dizer antes de emitir um comando “Executar comando. As taxas de confiança variaram entre 81,3% no submodo 1 e 75,0% no submodo 2⁷. Foram também realizadas duas medições das taxas de confiança para os comandos: “Activar modo comando” 83%, “Desactivar modo comando” 76% e “Párar reconhecedor” 68%. O número de erros considerando as expressões na gramática $1/52=1,9\% \ll 5\%$ previstos no ponto anterior, contra $3/52=5,8\%$ considerando as expressões ainda não presentes na gramática.

Sobre a versão final do projecto foi também realizada a avaliação da execução de três tarefas (fácil, média e difícil). As experiências foram realizadas com 4 utilizadores. Apresento de seguida as tarefas, os tempos previstos e o número de erros.

| | Esperado | | Obtido | |
|---------|----------|-------|--------|-------|
| | Tempo | Erros | Tempo | Erros |
| Teclado | 5-10s | 0 | 5,8s | 0 |
| Rato | 5-10s | 0 | 7,5s | 0,125 |
| Voz | 5-20 | 0 | 11,3 | 0 |

Tabela 4: Tarefa 1 (simples): deslocar para a direita

| | Esperado | | Obtido | |
|---------|----------|-------|--------------------|-------|
| | Tempo | Erros | Tempo | Erros |
| Teclado | 10-20s | 0 | 19,5s ⁸ | 0 |
| Rato | 10-20s | 0 | 12,3s | 0 |
| Voz | 10-40 | 1 | 17,5s | 0,125 |

Tabela 5: Tarefa 2 (média): Ir para um canto da imagem

| | Esperado | | Obtido | |
|---------|----------|-------|--------|----------------|
| | Tempo | Erros | Tempo | Erros |
| Teclado | 10-30s | 2 | 18,5s | 0,5 |
| Rato | 10-30s | 2 | 17,3s | 0,25 |
| Voz | 20-60 | 2 | 33,3s | 1 ⁹ |

Tabela 6: Tarefa 3 (complicada): mostrar Austrália com zoom a 80%

A experiência decorreu muito bem com os utilizadores seleccionados e todos os valores ficaram acima das expectativas.

Foram ainda detectados os seguintes problemas em entrevista com os utilizadores que realizaram as tarefas acima: para a modalidade voz: embora quase sem críticas para a gramática ainda fica por pesquisar exaustivamente sinónimos ou outras formas de dizer os comandos e

⁷ 13 comandos x 2 utilizadores = 26 testes por submodo

⁸ Com o teclado, por lapso, não é possível fazer “tudo *”.

⁹ Falar quando o reconhecimento termina e recomeça (50%) e mau reconhecimento (50%)!

incluí-los na gramática; para a modalidade apontar: o não suportar zoom in / zoom out com o rato e o não suportar o arrastar (que ficaram em 2.º plano e para trabalho futuro); para a modalidade teclado: o não suportar os comandos “tudo *” – bug detectado e corrigido na versão final e sobre a interface o *feedback* voz pouco útil e que deve ser melhorado.

6. CONCLUSÕES

Este projecto permitiu concluir que é possível adicionar de forma válida, flexível e robusta a modalidade voz a uma interface WIMP para navegação em imagens de elevada resolução. Os resultados foram muito encorajadores e revelaram que os utilizadores com um tempo de aprendizagem muito baixo a conseguem utilizar.

O lado menos favorável é o facto do tempo de realização das mesmas tarefas ser em média o dobro quando se utiliza a modalidade voz. O mais favorável o facto de poder andar livremente por uma sala a realizar, por exemplo, uma apresentação em que posso falar livremente e quando quero mudar para o acetato seguinte dizer “executa comando seguinte” ou outra variante (com a mesma tecnologia utilizada). Hoje em dia já é possível fazer algo semelhante com um ponteiro electrónico sem fios na mão mas no cenário acima estamos com as mãos totalmente livres!

Outra conclusão importante, é que com uma modalidade como a voz os utilizadores descobrem sempre uma maneira diferente de fazer uma simples tarefa como “direita”. (Esta situação ocorreu durante os testes, mas embora os comandos emitidos não estivessem na gramática devido às suas similaridades com outros existentes resultaram nos comandos correctos.) Tal facto faz com que a implementação da gramática seja ainda um trabalho em conclusão.

Ficou para trabalho futuro e cuja realização seria uma mais valia: a) a adição de feedback mais efectivo por voz utilizando o TTS (no projecto foi apenas utilizador para repetir o comando reconhecido); b) a adição de todos os sinónimos existentes para os comandos reconhecidos (utilizando um dicionário de sinónimos); c) a integração com a sala multimédia do campus do IST do Taguspark.

7. AGRADECIMENTOS

Gostaria de deixar expresso o meu agradecimento a todos os que contribuíram de algum modo na concretização deste trabalho. Agradeço pois os meus alunos e colegas da Escola Secundária de São João da Talha; aos colegas do grupo IMMI e ao Professor Joaquim Jorge pelas suas críticas e ideias construtivas.

Agradeço também às entidades INESC-ID / INOV, na pessoa de John Rodrigues, pela obtenção e renovação da licença do ASR e TTS da Loquendo e claro está à Loquendo ;-).

A todos o meu obrigado!

8. REFERÊNCIAS

8.1 Artigos científicos

- [Baseline00] IE 2013 Baseline. Data for user validation in information engineering. *A Telematics Applications Programme Support Action in Information Engineering - Project IE 2013 - Baseline*, <<http://www.ucc.ie/hfrg/baseline/>>
- [Billinghurst98] Billinghurst, Mark. Put that where? Voice and Gesture at the graphics Interface. *Human Interface Technology Laboratory, University of Washington, 1998*. <http://portal.acm.org/ft_gateway.cfm?id=307730&type=pdf&coll=GUIDE&dl=ACM&CFID=48796827&CFTOKEN=72480303>
- [Bilmes05] Bilmes, Jeff; Li, Xiao; Malkin, Jonathan; Kilanski, Kelley; Wright, Richard; Kirchhoff, Katrin; Subramanya, Amarnag; Harada, Susumu; Landay James A.; Dowden, Patricia; Chizeck, Howard. *Dept. of Electrical Engineering, Dept. of Linguistics, Dept. of Computer Science, Dept. of Speech & Hearing Science, University of Washington, Seattle, WA, Julho de 2005*. <<https://www.ee.washington.edu/techsite/papers/documents/UWEETR-2005-0007.pdf>>
- [Blackwell04] Blackwell, Alan. Human Computer Interaction. *Part II course - University of Cambridge Computer Laboratory, 2004-2005*. <<http://www.cl.cam.ac.uk/Teaching/current/HCI/HCI2004.pdf>>
- [Bolton80] Bolton, Richard A.. Put-that-there. *Architecture Machine Group, Massachussets Institute of Technology, 1980*. <http://www.media.mit.edu/speech/papers/1980/bolt_SIGGRAPH80_put-that-there.pdf>
- [Kaltenbrunner05] Kaltenbrunner, Martin. Auditory Interfaces. <<http://modin.yuri.at/teaching/AuditoryWorkshop/>>

- [Kimani05] Kimani, Stephen. Elements of the WIMP Interface. *Universita' di Roma "La Sapienza", February 2005*. <<http://www.dis.uniroma1.it/~kimani/teach/hci/slides/10Feb2005UIele.ppt>>
- [Sullivan90] Sullivan, Kelly. Using Task Analysis in Documentation Field Research. *Usability Group - Microsoft Corporation, 1990?* <<http://www.microsoft.com/usability/UEPosting/s/p149-sullivan.pdf>>
- [Taylor97] Taylor, Ashley George. WIMP Interfaces. *CS6751 Topic Report, Winter 1997*. <http://www.cc.gatech.edu/classes/cs6751_97_winter/Topics/dialog-wimp/>

8.2 Bibliográficas

- [LoquendoSDK - Documentação]
- Evaluation Kit,
 - Grammar Specifications,
 - Java™ Speech Grammar Format,
 - SpecificationProduct Components and Documentation Plan,
 - Speaker Verification User's Guide,
 - User's Guide,
 - Speech Assistant Toolkit Console Application Users' Guide,
 - Speech Recognition Grammar Specification.

Páginas Web

- [Loquendo05] <<http://www.loquendo.com/>>
- [Multimap05] <<http://www.multimap.com/>>
- [Nielsen90] <http://www.useit.com/papers/heuristic/heuristic_evaluation.html>