EDDY: UM EDITOR GRÁFICO MULTIMODAL COM RECONHECIMENTO DE FALA E GESTOS

Manuel João Fonseca[1,2]
Joaquim Armando Jorge[1,2]

[1]IST, Av. Rovisco Pais, 1000 Lisboa, Portugal
[2]INESC, Rua Alves Redol nº 9, 1000 Lisboa, Portugal

Sumário

A evolução das Interfaces Utilizador no sentido de melhorar a Interacção Pessoa-Computador, conduziu-nos às Interfaces Utilizador Gráficas (GUI) e mais recentemente às Interfaces Multimodais. O primeiro tipo de interfaces, utiliza o teclado e o rato como principais dispositivos de entrada, enquanto o segundo tipo procura explorar toda a gama sensorial do utilizador, incluindo por exemplo a visão, a audição, etc.. As Interfaces Multimodais são mais flexíveis, mas por outro lado exigem do utilizador alguma adaptação. Para a construção destas interfaces é necessário avaliar o grau de adaptação dos utilizadores às técnicas de interacção multimodal e o aumento de produtividade que estas possam introduzir. Para realizar esse estudo desenvolvemos um editor gráfico de formas geométricas que utiliza a fala e os gestos como modalidades de entrada e ícones auditivos, estímulos visuais e animação como modalidades de saída. O estudo consistiu em comparar o editor gráfico desenvolvido com uma aplicação convencional, ao nível da rapidez de execução, agradabilidade, e adaptação dos utilizadores.

Com base nos resultados obtidos, discutimos a possibilidade de utilização deste tipo de interfaces em aplicações interactivas.

0. Introdução

A contínua evolução da Informática, tem levado a um progressivo desenvolvimento das capacidades funcionais e de processamento dos computadores permitindo a sua difusão por um maior número de utilizadores. Paralelamente a este desenvolvimento, as Interfaces Utilizador têm evoluído para aproveitar as maiores capacidades de processamento dos computadores no sentido de integrarem mais capacidades humanas.

As primeiras interfaces baseadas em caracteres foram a solução óbvia num contexto onde a principal preocupação no desenvolvimento de aplicações era o de optimizar tempo de computação e memória. O aparecimento das primeiras interfaces gráficas demonstrou uma maior preocupação com a qualidade das Interacções Pessoa-Computador. No entanto as capacidades necessárias para trabalhar com este tipo de interfaces, tais como: operações sobre teclado e rato, conceitos sobre ícones e janelas, requerem aprendizagem específica para se poder trabalhar com um computador [Ima92].

Segundo [Ima92], o próximo passo na evolução das interfaces poderá passar pela integração,

nas técnicas de interacção, das capacidades pré-adquiridas e usadas diariamente por qualquer pessoa, como por exemplo: a fala, a escrita e os gestos. É neste sentido que aparece um novo conceito de interface, a **Interface Multimodal**. Ao contrário das habituais interfaces gráficas dominadas pelo rato, janelas e teclado, as Interfaces Multimodais incluem um conjunto de modalidades tais como voz, gestos, escrita à mão, seguimento do olhar, etc., para a entrada de dados e informação multimédia para a saída.

A integração de múltiplas modalidades de entrada permite alcançar uma maior expressividade a partir de fontes de informação complementares, e maior fiabilidade devido à utilização de modalidades redundantes. No entanto a simples adição de novas modalidades não garante só por si um aumento da qualidade de uma interface, é necessário que haja uma correcta integração para que a nova modalidade traga algum valor acrescentado à qualidade da interface. Uma integração correcta das várias modalidades passa por dar ao sistema a capacidade de escolher as modalidades de entrada e saída mais apropriadas para cada situação e a capacidade de reconhecer expressões de entrada construídas com informação proveniente de diferentes modalidades.

Começamos por clarificar a noção de Interface Multimodal, apresentando algum trabalho relacionado. Na secção seguinte descrevemos a funcionalidade do editor assim como a sua arquitectura. Finalmente descrevemos o estudo de usabilidade, analisam-se os resultados e apresentam-se as conclusões.

1. Interfaces Multimodais

De acordo com [Gli96][Blat92]e[May93] a definição de modalidade usada na Interacção Pessoa-Computador, advém da definição utilizada em psicologia, onde estes se referem às modalidades sensoriais das pessoas, tais como a visão, a audição, o tacto, o olfacto e o paladar.

As pessoas no seu ambiente diário utilizam vários canais de comunicação (sentidos) para comunicarem entre si. Tendo como referência a comunicação multimodal Pessoa-Pessoa, procurou-se transpor esse tipo de comunicação para as interfaces Pessoa-Computador, de modo a aumentar a largura de banda na comunicação.

De acordo com [Cou91], as Interfaces Multimodais podem ser classificadas em Interfaces Multimodais Exclusivas e Interfaces Multimodais Sinérgicas.

Uma interface multimodal diz-se exclusiva quando existem várias modalidades disponíveis e apenas uma delas é utilizada e diz-se sinérgica quando existem várias modalidades disponíveis e são utilizadas várias dessas modalidades simultaneamente, quer na entrada, quer na saída de informação.

2. Trabalho Relacionado

Esta secção pretende dar a conhecer alguns sistemas realizados que fazem uso duma interface multimodal. Não está no entanto no objectivo desta análise falar dos detalhes técnicos ou mesmo dos pormenores da arquitectura dos sistemas, mas sim mencionar algumas características destes no âmbito das interfaces multimodais.

2.1 Jeanie

O primeiro sistema analisado foi o Jeanie[Vo96], uma agenda electrónica que faz uso de uma interface multimodal. A Jeanie oferece como modalidades: a fala, os gestos através de uma caneta e de um ecrã táctil e oferece ainda a possibilidade de usar palavras escritas.

A interface do sistema pode ser classificada como multimodal sinérgica e como multimodal exclusiva, pois numas situações utiliza várias modalidades e noutras apenas uma.

O utilizador pode fazer uma cruz com a caneta por cima de um apontamento de uma reunião para a anular ou pode-se dizer *"Reschedule this on Friday"* e apontar para uma marcação de reunião para alterar o dia dessa mesma reunião, ou então, pode-se simplesmente desenhar uma seta desde o sítio onde estava a marcação para o novo dia e nova hora na agenda.

2.2 Sync/Draw

O segundo sistema analisado foi o Sync/Draw [Mat97]. O Sync/Draw é um editor gráfico multimodal baseado na interpretação incremental de palavras faladas em japonês e no uso do rato como dispositivo apontador. A linguagem falada é interpretada palavra a palavra. A interface utilizada é do tipo multimodal sinérgica.

Como grandes vantagens deste sistema há a referir a correcção imediata de erros devido a má interpretação, o menor tempo de espera e ainda a vantagem de se poder trabalhar enquanto se vai confirmando o estado da aplicação.

Este projecto serviu para a realização de estudos de usabilidade em sistemas que utilizam ou possam vir a utilizar interfaces multimodais. Assim chegaram à conclusão que ao utilizar o Sync/Draw o tempo de trabalho era reduzido em média cerca de 21.3%, além de reunir as preferências dos utilizadores.

Como resultado do estudo há a realçar a maior eficiência dos sistemas baseados na técnica de interpretação palavra a palavra, sobre os sistemas baseados na técnica de interpretação frase a frase.

2.3 LIMSI-DRAW

O LIMSI-DRAW [Bell95] é um editor gráfico multimodal que foi desenvolvido principalmente numa óptica exploratória.

O editor permite criar e manipular formas geométricas simples (rectângulos, triângulos, círculos) através de um conjunto de comandos multimodais. Para criar um rectângulo, por exemplo, diz-se a palavra **Rectângulo** e depois marcam-se dois pontos no ecrã usando o rato, ou o ecrã táctil.

O editor utiliza como entradas um sistema de reconhecimento de fala, um ecrã táctil e um rato. Suporta multimodalidade sinérgica, podendo receber comandos usando combinação de fala, rato e ecrã táctil.

3. O Editor

A primeira parte do trabalho consistiu no desenvolvimento de uma aplicação, com interface multimodal, que pusesse em prática os conceitos e as características inerentes à multimodalidade. A aplicação a desenvolver seria um editor gráfico de formas geométricas que realizasse algumas das funcionalidades comuns a este tipo de editores. Nomeadamente devia ter capacidades para permitir o desenho de figuras geométricas, a manipulação de atributos das mesmas (cor, tipo de linha) e a execução de certas acções sobre as figuras, como apagar, recortar, preencher, etc..

O objectivo era portanto desenvolver uma aplicação que, recorrendo a reconhecedores de gestos e de fala já existentes, oferecesse um conjunto de técnicas de interacção multimodais, de preferência sinérgicas, com vista a uma posterior avaliação das mesmas em contraponto com outras técnicas de interacção convencionais.

O desenho das técnicas de interacção foi sobretudo orientado para factores de usabilidade da interface, tais como a eficiência, a segurança de utilização e facilidade de aprendizagem. Na definição das interacções a desenvolver seguimos os conceitos inerentes à multimodalidade, onde se destacam:

- Várias modalidades de entrada.
- Várias modalidades de saída.
- Mais que um modo de interacção para cada funcionalidade
- Escolha da modalidade de saída mais apropriada.
- Combinação de várias modalidades de saída.
- Capacidade para interpretar expressões de entrada que integrem várias modalidades.

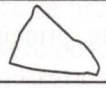
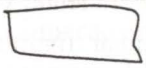
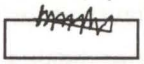
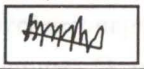
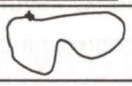
3.1 Funcionalidade

O editor gráfico, tal como foi referido atrás suporta como modalidades de entrada os gestos e a fala. Para colocar estas modalidades acessíveis, utilizamos um reconhecedor de fala[1] e um reconhecedor de gestos [Vil96].

[1] IN^3 Voice Command.

3.1.1 Interacção usando Gestos

Quadro 1 - Alguns dos gestos suportados pelo editor

	Cria uma linha
	Cria um triângulo
	Cria um circulo
	Cria um rectângulo
	Cria um losango
	Cria uma elipse
	Apaga uma figura
	Apaga parte da figura
	Preenche a figura
	Selecciona as figuras que estiverem dentro

No contexto do editor, um gesto é definido como um símbolo desenhado usando uma caneta sobre uma tablete. No **Quadro 1** encontram-se enumerados os diferentes gestos suportados pelo editor. Como se pode ver existem gestos para criação de figuras e gestos para manipular essas mesmas figuras. O editor permite ainda a realização de operações de manipulação directa, como por exemplo mover e redimensionar as figuras, usando a caneta.

3.1.2 Interacção usando Fala

A interacção usando apenas fala, pressupõe que exista pelo menos um objecto seleccionado. Os comandos de fala são fundamentalmente acções sobre as figuras, logo é necessário que o alvo dessas acções esteja identificado.

Quadro 2 – Alguns dos comandos de Fala

Delete	Apaga o objecto seleccionado
Fill <Color>	Preenche uma figura com a cor *Color*
Bigger	Aumenta proporcionalmente a figura
Smaller	Diminui proporcionalmente a figura

Nos comandos constituídos por mais que uma palavra (Ex: Fill Green), não basta dizer as palavras pela ordem correcta, é necessário também dizê-las com um intervalo de tempo

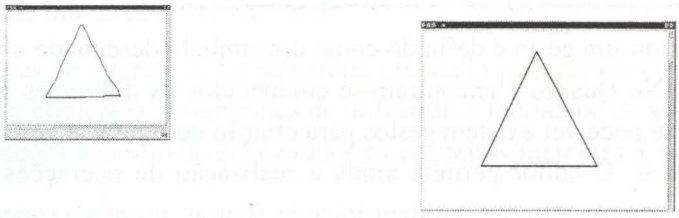
pequeno entre elas, caso contrário serão consideradas duas palavras isoladas.

3.1.3 Vários modos de Interacção para a mesma Funcionalidade

O editor oferece vários modos de interacção para a mesma funcionalidade, de modo a que o utilizador possa escolher a que mais lhe convém. Seguindo este princípio procuramos utilizar técnicas de interacção para executar a mesma acção, que recorressem a modalidades diferentes. O objectivo era permitir ao utilizador escolher a modalidade que mais lhe agradasse, ou que devido às condições da interacção fosse mais eficiente. Como exemplo podemos observar as diferentes técnicas usadas para apagar e desenhar figuras.

Apagar figuras - Para apagar uma figura podemos usar o gesto de apagar descrito anteriormente, ou dar o comando oral **Delete**. Trata-se exactamente da mesma funcionalidade, mas como um dos objectivo é fazer um estudo sobre a usabilidade da multimodalidade, situações como esta onde a complexidade de ambas as técnicas não é muito diferente, são perfeitas para verificar qual a modalidade preferida pelo utilizador.

Desenhar figuras - Para desenhar um triângulo (para as outras figuras o procedimento é similar), existem dois métodos de interacção, o primeiro baseado unicamente em gestos e o segundo recorrendo à fala e à manipulação directa. O primeiro, indicado na **Figura 1**, consiste em desenhar sobre a tablete um rascunho de um triângulo. O segundo requer que seja dita a palavra **Triangle** e que depois se defina com a caneta o centro e a dimensão deste.



(a) (b)

Figura 1 - Desenho de um triângulo usando gestos (a) – Desenho do rascunho (b) – Triângulo reconhecido

Existem no entanto funcionalidades que só podem ser realizadas usando uma técnica de interacção. É o caso do recorte de figuras (só gesto) ou da mudança de cor (só fala).

3.1.4 Comandos de entrada que integram várias modalidades

A combinação de várias modalidades de entrada para realizar acções é um dos pontos essenciais no desenvolvimento de Interfaces Multimodais sinérgicas. Este tipo de interacção permite uma maior flexibilidade no uso do editor, e faz com que a sua utilização seja mais natural e mais agradável. Como exemplos demonstrativos, apresentam-se duas situações onde o comando sinérgico combina gestos e fala.

O editor suporta o desenho de uma figura através do desenho de um rascunho da mesma e simultaneamente permite a alteração da cor da figura usando a fala. Se pretendermos desenhar uma elipse verde teremos que durante o desenho do rascunho da elipse dizer **Green**, e a partir desse momento o rascunho fica a verde, tal como a elipse reconhecida.

Outra situação onde se usam comandos que combinam as duas modalidades, é quando não existem figuras seleccionadas e pretendemos executar uma acção. No contexto de um editor gráfico onde podem existir várias figuras presentes na área de trabalho a melhor maneira de referenciar um objecto é usar o gesto de apontar.

Suponhamos uma situação em que existe um conjunto de figuras desenhadas, onde nenhuma está seleccionada e pretendemos apagar o triângulo. Uma das hipóteses é usar a fala para definir o comando, dizendo **Delete**. Como não existem figuras seleccionadas o sistema requer que seja referido especificamente uma figura. Para informar o utilizador de que está à espera o sistema altera o cursor e a imagem da personagem da animação. Para referenciar a figura a apagar o utilizador terá que usar uma combinação de fala e gesto, isto é, enquanto aponta para o triângulo diz a palavra **This**, resultando na remoção do mesmo. Este método imita situações habituais da vida real onde utilizamos a fala e o gesto de apontar para indicar objectos que se encontram no nosso campo visual. O **Quadro 3** mostra alguns dos comandos multimodais sinérgicos suportados pelo Eddy.

Quadro 3 - Alguns comandos sinérgicos

Fill ▼[2]This Red	Preenche a vermelho o objecto apontado
Delete ▼This	Apaga o objecto apontado

3.1.5 Combinação de modalidades de saída

Qualquer interface que pretenda alcançar bons níveis de usabilidade tem que manter um bom nível de comunicação com o utilizador, de modo a que este esteja constantemente informado sobre o estado da aplicação. A combinação de várias modalidades de saída permite uma maior expressividade na comunicação com o utilizador.

A seguir damos exemplos específicos dos vários tipos de combinações de modalidades de saída disponíveis no editor, tentando sublinhar o que se pretende com as mesmas.

Combinação Cursor/Animação

A utilização do cursor como forma de comunicar o estado da aplicação não é novo, sendo habitual em qualquer aplicação gráfica. O que nós procurámos foi, complementar essa informação usando animações. Nesta perspectiva a combinação cursor/animação é usada

[2] Este simbolo serve para indicar que a palavra que se segue deve ser acompanhada de um gesto simultâneo de apontar.

para aumentar o grau de comunicação disponível.

Para exemplificar suponhamos que se pretende desenhar um rectângulo recorrendo à fala e à manipulação directa, é necessário dizer **Rectangle** e depois marcar dois pontos que definem o rectângulo. Para que o utilizador possa saber que o sistema reconheceu a palavra e que está à espera da marcação de um ponto, o editor usa a animação e o cursor, tal como se pode ver na **Figura 2**.

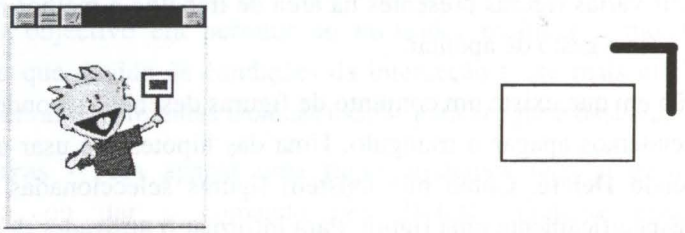


Figura 2 - Combinação do Cursor/Animação quando se reconhece a palavra Rectangle

A combinação destas modalidades é usada em todas as situações similares à descrita. Pretende-se que este tipo de interacção seja o principal auxiliar do utilizador durante uma sessão de trabalho, no sentido de disponibilizar informação sobre o estado actual do editor. A existência desta combinação torna-se relevante devido à existência de temporizadores para limitar o intervalo de tempo entre palavras da mesma expressão. Sempre que o sistema está à espera de uma nova palavra para completar uma expressão indica-o até que o temporizador expire. Quando isso sucede, o cursor e a animação voltam ao estado de espera.

Combinação Som/Efeito Visual

O editor combina duas modalidades para fornecer retorno de acções desencadeadas pelo utilizador. Através de ícones auditivos [Blat90] fornece-se indicação imediata sobre o tipo da figura reconhecida pelo editor e indicações de erro. Vários efeitos visuais, incluindo animações complementam a informação anterior dando indicações específicas sobre o estado do editor em resposta a gestos do utilizador.

Como exemplo podemos observar o que acontece quando se preenche uma figura riscando o interior da mesma, ou proferindo a expressão **Fill <Color>.** Em vez de preencher completa e instantaneamente a figura, acrescentamos um efeito visual que a preenche gradualmente e a aplicação gera um som que se assemelha ao produzido por um recipiente a ser cheio.

Outra situação em que se recorre a este tipo de combinação de modalidades permite fornecer retorno quando se desenham figuras. Associado ao aparecimento da figura desenhada é gerada uma nota musical diferente para cada figura geométrica, de modo a que o utilizador possa associar este "ícone auditivo" à figura reconhecida.

Por experiência verificámos que devido às suas características geométricas certas figuras são confundíveis (Ex. losangos e rectângulos), sendo por vezes difícil de verificar qual a figura reconhecida. Para facilitar a percepção atribuímos notas mais afastadas na escala musical

para figuras com características mais semelhantes.

3.2 Arquitectura do Eddy

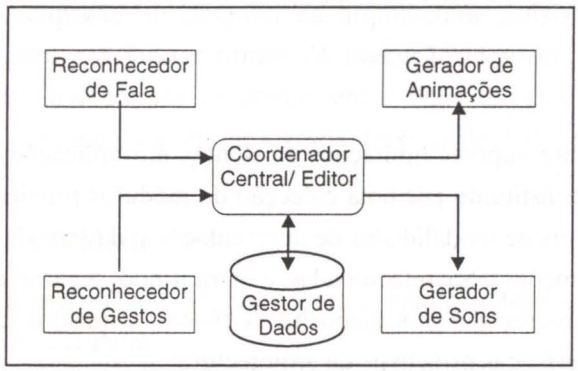


Figura 3 - Arquitectura do Eddy.

O Eddy utiliza uma arquitectura distribuída representada na **Figura 3**. O editor compõe-se de vários componentes autónomos que comunicam entre si através de mensagens simbólicas.

- **Reconhecedor de Fala** – É responsável por reconhecer as palavras ditas pelo utilizador, e por as enviar para o coordenador central. A mensagem gerada pelo reconhecedor é constituída pelo comando reconhecido e por uma marca temporal, que irá ser usada pelo coordenador central na validação dos comandos multimodais sinérgicos.
- **Reconhecedor de Gestos e Figuras** – Reconhece os gestos, ou figuras desenhadas pelo utilizador e envia o resultado para o coordenador central. O reconhecimento de um gesto implica o envio do comando reconhecido e da marca temporal correspondente. No caso das figuras, a informação enviada inclui o tipo de figura identificado pelo reconhecedor, os atributos correspondentes e uma marca temporal.
- **Gerador de Animações** – Este componente ao receber uma mensagem, vai realizar a animação correspondente. As animações podem ser sequências de imagens pré-gravadas, ou animações gráficas geradas pelo editor.
- **Gerador de Sons** – É responsável por produzir os ícones auditivos, e reproduzir sons pré-gravados. Em função da mensagem recebida, o componente escolhe o tipo de som a reproduzir. Estes sons encontram-se gravados em ficheiros com o formato *au*.
- **Gestor de Dados** – Mantém e gere as estruturas de dados utilizadas pelo editor, como por exemplo a lista de figuras existentes, ou a lista de acções efectuadas. É também responsável por garantir a persistência da informação manipulada pelo editor.
- **Coordenador Central/Editor** – Este componente para além de ser responsável pela funcionalidade geral do editor, é também responsável por verificar a validade das mensagens enviadas pelos outros componentes. As mensagens provenientes dos diferentes reconhecedores que são temporalmente próximas e válidas são combinadas de

modo a formarem comandos multimodais sinérgicos, que o coordenador interpreta e executa. O coordenador encarrega-se ainda de trocar informação com o gestor de dados de modo a completar correctamente os comandos dados pelo utilizador, assim como gerar as mensagens necessárias para produzir saídas que utilizem várias modalidades.

3.2.1 Características

A arquitectura do editor permite superar limitações existentes em aplicações monolíticas convencionais. O programa é constituído por uma colecção de módulos funcionais distintos permitindo combinações flexíveis de modalidades de apresentação e de introdução de dados sem necessidade de alterar funções não relacionadas e permitindo separar estruturas de diálogo da configuração do programa e dos dispositivos físicos utilizados. Entre outras, destacamos as seguintes características principais da arquitectura:

- **Autonomia** – Cada um dos componentes existe independentemente dos outros, comunicando entre si por mensagens. O editor compõe-se de uma colecção de processos autónomos que trocam mensagens através de *sockets* utilizando o protocolo *TCP/IP*. O coordenador central cria os processos que concretizam cada componente de acordo com as necessidades do diálogo e dispositivos a utilizar. Desta forma é possível maximizar a independência dos dispositivos e a flexibilidade do diálogo.
- **Concorrência** – Cada um dos componentes autónomos processa concorrentemente diferentes canais de informação, como gestos e fala suportando comandos multimodais sinérgicos de uma forma natural. O tempo adicional gasto na sincronização e escalonamento de processos num ambiente Unix, não afecta de forma perceptível o desempenho dos diálogos, dada a granularidade macroscópica das operações distribuídas.
- **Abertura** – A simplicidade da informação trocada e do protocolo que controla o fluxo de dados, associado à autonomia dos componentes, facilita a substituição de componentes por outros similares, *mesmo durante a execução do programa*, não sendo necessário alterar qualquer um dos outros componentes do sistema desde que a nova unidade cumpra o protocolo preestabelecido. Esta característica é altamente desejável de um ponto de vista da manutenção e configuração do editor.
- **Distribuição** – Entre reconhecedores ou geradores e o coordenador central é apenas trocada informação simples e compacta como palavras identificadas pelo reconhecedor de fala, sons a emitir pelo gerador de audio ou atributos de figuras geométricas reconhecidas. Nada impede que os componentes se executem em máquinas diferentes. No caso do nosso editor, os comandos falados são reconhecidos numa *workstation Sun*, enquanto os restantes componentes do editor existem como processos num computador IBM/PC sob o sistema de exploração Linux.
- **Heterogeneidade** – Uma vez que processos autónomos podem estar distribuídos por diferentes máquinas, cada um dos componentes pode correr sob diferentes sistemas de

exploração em máquinas com distintas capacidades, possibilitando a escolha do computador, sistema de exploração e *software* que melhor se adaptam às necessidades de cada diálogo.

A principal preocupação no desenho da arquitectura consistiu em criar uma estrutura simples e leve para possibilitar ritmos de execução elevados. Esta abordagem altamente modular e aberta, autonomiza os componentes mais directamente ligados às entradas e saídas de informação para maximizar a gestão flexível de diálogos multimodais.

4. Testes de Usabilidade

Para comparar a usabilidade do Eddy com a de um editor com interface convencional, utilizamos o Microsoft Paint.

Pediu-se a um conjunto de utilizadores para executar uma tarefa prática e preencher um questionário. Durante a execução da tarefa foram medidas várias variáveis: tempo de execução, situações inesperadas para o utilizador, número de *undos* efectuados, dúvidas sobre o funcionamento da aplicação, entre outras. O inquérito pretende complementar a informação adquirida na execução da tarefa fornecendo dados sobre a agradabilidade e a facilidade de utilização da aplicação.

Pretendeu-se medir a usabilidade da interface da aplicação desenvolvida, dando particular ênfase aos aspectos específicos da multimodalidade. Nesta perspectiva definimos um conjunto de variáveis que quantificam os aspectos mais relevantes numa Interface Multimodal:

- **Desempenho dos reconhecedores** - Dado que qualquer interface depende da qualidade dos canais de comunicação entre o utilizador e o computador, a medição do desempenho dos reconhecedores reveste-se de particular importância. Neste sentido a medição das taxas de reconhecimento da fala e dos gestos permite extrair dados sobre a utilidade das modalidades que recorrem a estes reconhecedores.
- **Modalidades preferenciais** - Como foi referido anteriormente, o utilizador pode escolher diferentes técnicas de interacção para executar cada uma das funcionalidades do editor. Para determinar a modalidade preferida dos utilizadores para executar uma dada função, mediu-se o número de vezes que cada interacção foi realizada.
- **Reacções subjectivas à multimodalidade** - O uso da fala e dos gestos, de novas técnicas de interacção multimodais, de comandos sinergéticos e a utilização de sons e animações, são elementos novos numa interface, que provocam reacções subjectivas nos utilizadores. Tentámos captar essas reacções observando a atitude dos utilizadores durante a execução da tarefa e através das respostas dos utilizadores ao questionário.

Cada teste individual requeria a realização de uma série de tarefas:

- Desenho de uma figura no Microsoft Paint, e respectiva compilação de dados.

- Visualização de um vídeo demonstrativo sobre o Eddy.
- Treino do reconhecedor de fala no Eddy.
- Medição da taxa de reconhecimento dos gestos no Eddy, através do desenho de dez rascunhos para cada figura disponível.
- Medição da taxa de reconhecimento da fala no Eddy para uma série de vinte palavras.
- Desenho da figura no Eddy, e respectiva compilação de dados.
- Preenchimento de um questionário.

Paralelamente, efectuámos em cada uma das fases uma observação do desempenho e das reacções do participante.

Foram executados 10 testes com a participação de 8 utilizadores. Pretendeu-se variar as características dos vários participantes numa tentativa de obter um domínio mais vasto. Nesta perspectiva, recorremos a utilizadores com diferentes graus de experiência no uso de editores gráficos:

- Dois utilizadores com bastante experiência no uso do Paint;
- Dois utilizadores com um profundo conhecimento da interface do Eddy;
- Quatro utilizadores sem experiência no uso do Paint nem do Eddy, mas que no entanto já tinham utilizado editores gráficos não multimodais.

Com o intuito de medir a evolução registada pelos participantes quando confrontados com duas ou mais utilizações, decidimos executar uma segunda fase de testes com um utilizador experiente no uso Paint e outro não experiente.

5. Resultados

Nesta secção apresentamos uma compilação dos resultados dos testes de usabilidade obtidos para os diferentes módulos que compõem a aplicação, como o reconhecedor de fala e o analisador de gestos, bem como resultados referentes à aplicação em si.

Os valores foram obtidos de forma quantitativa, medindo tempos de execução, ou de forma qualitativa, através das opiniões dos utilizadores em relação à facilidade de utilização do editor, usando uma escala de 1 a 5.

5.1 Reconhecedores de Gestos e Fala

Para avaliar o desempenho dos reconhecedores de gestos e de fala, em situações de execução de tarefas, medimos taxas de reconhecimento antes e durante a execução do desenho. Pretendemos deste modo verificar se as taxas de reconhecimento poderiam ser influenciadas pela integração das modalidades em técnicas de interacção. De seguida apresentamos os valores médios das taxas de reconhecimento de fala e gestos de todos os participantes, para as duas situações acima indicadas.

Como se pode verificar pelos gráficos da Figura 4 e Figura 5 regista-se um menor desempenho no reconhecimento em situações de execuções de tarefas. Esta situação deve-se ao facto de o utilizador ter como referência principal o desenho no seu global e não o gesto ou a palavra que tem que referir.

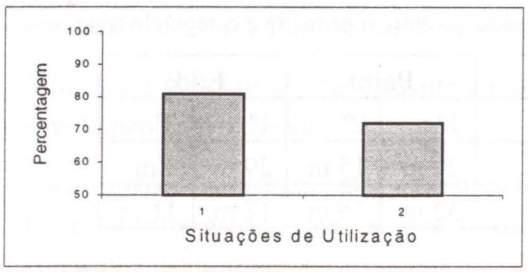


Figura 4 - Taxas de reconhecimento de fala. (1)- Antes da tarefa (2)- Durante a execução a tarefa

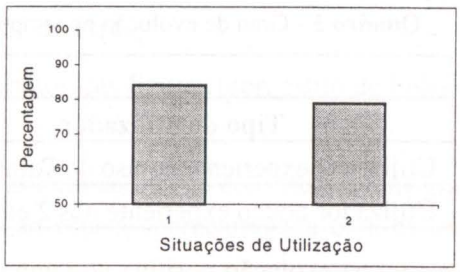


Figura 5 - Taxas de reconhecimento de gestos. (1) - Antes da tarefa (2) - Durante a execução da tarefa

5.2 Produtividade

Para medir a produtividade que se consegue obter em cada uma das interfaces compilamos dados sobre as seguintes variáveis:

- **Taxa de figuras executadas** - todos os utilizadores conseguiram executar a tarefa em ambos os editores
- **Taxa de acções efectuadas** - Em geral o número de acções necessárias para o desenho de uma figura é inferior no Eddy. Esta situação deve-se ao facto de as técnicas de interacção multimodais serem muito mais expressivas do que as utilizadas no Paint. Por exemplo, o desenho de elipses rodadas no Eddy requer uma única acção o que não se passa no Paint.

5.3 Resultados Globais

Outra variável medida durante os testes foi o tempo de execução do desenho, que apresentamos a seguir.

Quadro 4 - Tempo médio de execução

Tipo de Utilizador	Paint	Eddy
Utilizador experiente no uso do Paint	15 min.	23 min.
Utilizador experiente no uso do Eddy	17 min.	6 min.
Utilizador pouco experiente nos dois editores	22 min.	23 min.

O **Quadro 4** referencia os vários tempos de execução do desenho para cada grupo de utilizadores no Eddy e no Paint.

Como seria de esperar cada utilizador tem um melhor desempenho no editor com o qual está mais familiarizado. De notar que o terceiro grupo de utilizadores tem um tempo de execução

superior no Eddy, mas quase idêntico ao do Paint apesar de não estarem familiarizados com o tipo de interface do Eddy. Estes dados revelam uma adaptação inicial relativamente boa às interfaces multimodais.

De seguida apresentamos no **Quadro 5** valores que mostram o grau de evolução registado para os utilizadores que realizaram o segundo teste.

Quadro 5 - Grau de evolução no tempo de execução entre o primeira e o segundo teste

	Paint		**Eddy**	
Tipo de utilizador	1°	2°	1°	2°
Utilizador experiente no uso do Paint	21 m	15 m	29 m	8 m
Utilizador pouco experiente nos 2 editores	32 m	15 m	17 m	11 m

Verifica-se uma evolução positiva no tempo de execução para qualquer um dos editores, o que mostra que ambas as interfaces registam um bom grau de aprendizagem. Devido ao domínio deste teste em particular ser restrito (2 utilizadores), não podemos concluir mais sobre o grau de evolução.

6. Conclusões

As conclusões a extrair dos testes dividem-se em dois tipos: opiniões dos participantes sobre a interface do Eddy, e as observações sobre as reacções dos utilizadores durante a execução da tarefa.

A globalidade dos utilizadores considerou a interface do Eddy agradável e de fácil utilização, após vencida a dificuldade inicial na adaptação ao uso da caneta e da fala para interagir com o editor.

Consideraram uma vantagem bastante grande do Eddy, a rapidez de execução devido à não existência de menus e *toolbars* para definir figuras e modificar atributos.

Entre as desvantagens apontadas à interface multimodal referimos o número limitado de comandos disponíveis, a impossibilidade de recorrer ao desenho livre e a dificuldade de partindo dos comandos conhecidos, conseguir descobrir os restantes. Alguns utilizadores consideraram o uso do retorno sonoro algo maçador e pouco intuitivo, não ajudando na execução das tarefas.

Das observações por nós efectuadas durante o processo de realização dos testes destacamos as seguintes:

- Cada utilizador tende a usar preferencialmente uma modalidade para realizar as várias acções, recorrendo apenas à outra modalidade quando estritamente necessário. Este ponto permite concluir que a existência de várias técnicas de interacção, que recorrem a modalidades diferentes, para executar cada operação é benéfico para a adaptação de cada utilizador.

- Os utilizadores preferem seleccionar uma figura antes de trabalhar sobre ela, ao invés de recorrer ao gesto de apontar para a referenciar.
- A dificuldade inicial de memorizar comandos falados extensos, torna difícil a utilização da fala, sendo comum a troca da ordem das palavras que compõem a expressão. Os utilizadores demonstram uma tendência para se aproximarem do microfone quando pretendem utilizar a fala, distraindo a sua atenção do ecrã.
- Muitas das vezes os utilizadores alteram os atributos das figuras (cor, estilo de linha) em simultâneo com o desenho.
- Para utilizadores inexperientes no uso do Eddy as animações não constituíam uma grande fonte de informação, uma vez que a sua atenção estava sobrecarregada com as outras componentes da aplicação.

7. Discussão

A utilização de fala e gestos permite alcançar um bom desempenho na construção de técnicas de interacção.

Em face do trabalho desenvolvido podemos concluir que o estudo e desenvolvimento das interfaces multimodais é um campo que deve ser mais explorado, devido à excelente aceitação que teve o editor perante os utilizadores. No entanto, este tipo de interface, segundo a nossa opinião, não se coaduna com todo o tipo de aplicações, nomeadamente aplicações que tenham um conjunto de comandos bastante vasto e que sejam difíceis de exprimir de um modo "natural" ou intuitivo.

Em relação às técnicas de interacção que podem vir a ser desenvolvidas, deve-se tirar partido das interacções multimodais sinérgicas e da utilização de várias técnicas de interacção por funcionalidade, já que foram aspectos considerados positivos pelos utilizadores.

Finalmente em relação à arquitectura do editor, convém salientar que esta permite integrar novas modalidades, sem grande dificuldade, devido à sua modularidade. Para juntar uma nova modalidade, basta desenvolver o reconhecedor para ela e adaptar o Coordenador Central da aplicação, de modo a poder processar a informação enviada pelo novo reconhecedor.

Agradecimentos

Queremos agradecer ao Cláudio Salvador e ao Rogério Santos, pelo trabalho desenvolvido na codificação do editor e à JNICT, programa PRAXIS XXI contrato 2/2.1/TIT/1675/95 que financiou parcialmente este trabalho.

Referências

[Bell95] Yacine Bellik, *"Interfaces Multimodales: Concepts, Modèles et Architectures"*, Tese de Doutoramento, 1995

[Bla90] Meera M. Blattner, Denise A. Sumikawa e Robert M. Greenberg, *"Earcons and Icons: Their Structure and Common Design Principles"*, do livro de Ephraim P. Glinert, "Visual Programming Enviroments - Aplications and Issues", IEEE Computer Society Press, 1990.

[Bla92] Meera M. Blattner, Roger B. Danenberg, *"Multimedia Interface Design"*, ACM Press, 1992

[Cou91] Joelle Coutaz e Jean Caelen, *"A Taxonomy for Multimedia and Multimodal User Interfaces"*, Em Proceedings of the ERCIM Workshop of Distributed Systems, 1991

[Fol90] James D. Foley, Andries van Dam, Steven K. Feiner e John F. Hughes, *"Computer Graphics Principles and Practice"*, 2ª edição, Addison-Wesley, 1990.

[Gli96] Ephraim P. Glinert, Meera M. Blattner, *"Multimodal Interaction"*, IEEE Multimedia, Winter 1996, pp. 13-24

[Greg97] J. Gregory Trafton, Kenneth Wauchope e Janet Stroup, *"Errors and usability of natural language in a multimodal system"*, em IJCAI'97.

[Ima92] Takuji Imai, *"The Next-Generation User Interface: From GUI to Multimodal"*, Nikkei Electronics Asia, 1992.

[Mat97] Shigeki Matsubara, Hiroyuki Yamamoto, Nubuo Kawaguchi, Yasuyoshi Inagaki e Katsuhito Toyama, *"An Interactive Multimodal Drawing System based on Incremental Interpretation"*, em IJCAI'97.

[May93] Mark T. Maybury, *"Intelligent Multimedia Interfaces"*, AAAI Press/The MIT Press, 1993

[Usa97] *"Usability engineering and user interface evaluation"*, 1997.

[Vil96] Paulo Vilar e Jorge Barreira, *"Compilador e Analisador de Diagramas"*, Trabalho Final de Curso, 1996.

[Vo93] Minh Tue Vo e Alex Waibel, *"A Multimodal Human-Computer Interface: Combination of Gesture and Speech Recognition"*, Proc. InterCHI'93

[Vo96] Minh Tue Vo e Cindy Wood, *"Building an application framework for speech and pen input integration in Multimodal Learning Interfaces"*, Proceedings of the ICASSP, Atlanta, GA, May 1996