A Talking Face for Portuguese

Raquel de Castro Lisboa[1], António J. S. Teixeira, Artur Pimenta Alves[1,2]

Resumo – **A combinação de voz sintetizada com agentes animados, completos ou representados pela cara apenas, tem muitas aplicações, inlusive nas Telecomunicações. Tendo por objectivo contribuir para a existência desta tecnologia para a nossa língua, este artigo apresenta o desenvolvimento dos blocos necessários para uma "cara falante" para o Português. O trabalho incidiu na adaptação de um sistema, o RUTH, o desenvolvimento de uma voz rudimentar para para o Português utilizável no sistema de síntese de voz Festival e o desenvolvimento de uma forma simples de controlar o comportamento da cara pela utilização de anotações XML. Como demonstrador, foi desenvolvida uma aplicação capaz de controlar em tempo real várias "caras falantes", podendo cada uma utilizar uma voz diferente, tendo por entrada texto com anotações XML**

Abstract – **Combination of speech output with animated agents, complete or represented by face alone, has many applications, including in telecommunications. Aiming at contributing to the availability of such technologies for our language, this paper presents the development of the necessary building blocks to have a talking face for Portuguese. Work focused on the adaptation of an available system, named RUTH, the development of a basic Portuguese voice for the Festival speech synthesis system, and development of an easier way of controlling the talking face behaviour by using XML annotations. As a demonstration of the developed talking face, an application capable of controlling in real time several faces, each speaking with a different voice, based on XML annotated text input was implemented.**

Keywords: Talking face, visual speech, computer animation, face animation, audiovisual speech synthesis, Text-to-Speech synthesis.

I. INTRODUÇÃO

Assim como a integração da voz se apresenta muitas vezes como uma mais valia em relação ao texto escrito, a imagem de uma cara animada em 3D surge, também ela, como um complemento importante para a transmissão e compreensão de uma determinada mensagem [3]. É indiscutível que a possibilidade de observar a cara (ou mesmo apenas os lábios) do falante aumenta a inteligibilidade da fala, bastando para isso reportarmo-nos à técnica de leitura labial, usada por pessoas com deficiências auditivas. Não é, portanto, surpreendente que o mesmo se passe com pessoas que não possuam esse tipo de deficiência, quando na presença de ambientes nos quais a comunicação se apresenta de alguma forma degradada (por exemplo, no caso de ambientes ruidosos) [6]. Existe, sem sombra de dúvida, informação preciosa transportada na cara do falante que não deve ser desprezada.

De facto, muitas vezes usamos diversos sinais, verbais e não verbais, para reforçar a ideia chave num discurso. Coisas simples que utilizamos instintivamente, como a entoação, expressões faciais e movimentos da cabeça não são necessariamente redundantes, podendo, muitas vezes, acrescentar algo de novo à interpretação de uma mensagem, transformando-se numa contribuição eficaz para a conversação.

Na área concreta dos agentes de conversação, existem já algumas aplicações direccionadas para o mercado do consumidor, como o fornecimento de serviços de informação (notícias, meteorologia, etc.) em que a própria personagem computorizada surge como complemento do texto e da fala. É ainda de destacar, a título de exemplo, o projecto SYNFACE [7] que tem como objectivo principal facilitar a utilização de um simples telefone a pessoas com deficiências auditivas, conseguido através da conexão de uma cara falante ao telefone.

II. OBJECTIVOS E ABORDAGEM

O objectivo principal deste projecto consistiu no desenvolvimento de uma cara sintética para utilização conjunta com Português falado. Como objectivo complementar, pretendia-se desenvolver uma aplicação de demonstração dos componentes desenvolvidos.

Adicionalmente, no decorrer do projecto, e uma vez que não existe nenhuma voz de Língua portuguesa que podesse ser usada (freeware), começou a fazer sentido tentar construir uma base de desenvolvimento.

O esquema da Figura 1 ilustra, de uma forma geral, os diferentes módulos e programas intervenientes, e sua interacção. O módulo designado por Interface gráfica para além de lidar com a visualização e interacção com o utilizador controla os outros componentes: sistema de síntese Festival e a cara.

[1] FEUP, [2] INESC Porto

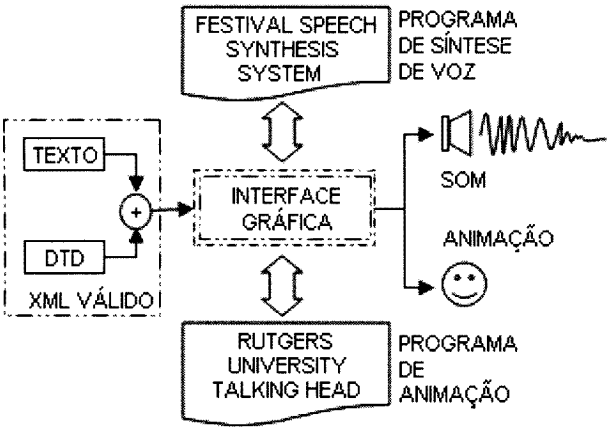


Fig. 1 - Esquema dos módulos, programas e sua interacção

III. SOFTWARE

O primeiro passo na execução deste projecto, foi a pesquisa, análise e teste de diversos programas, a fim de perceber qual o mais adequado à finalidade pretendida. Foram avaliados o *POSER 5*, o *GALATEA*, o *CUANIMATE* [11], o *BALDI* [9,10], LUCIA [5] e o *RUTH* [2].

Considerando factores como a programabilidade, o custo, a plataforma usada e a compatibilidade com o *Festival Speech Synthesis System*, o programa RUTH revelou ser a melhor opção.

Como consequência directa da escolha adequada do programa de animação, a tarefa de desenvolvimento das facilidades de sincronização do sinal de voz com a cara revelou-se desnecessária.

Não pretendendo descrevê-lo exaustivamente, nas subsecções seguintes é feita uma breve caracterização do software utilizado.

A. Rutgers University Talking Head

O RUTH [2] é um sistema de animação facial, em tempo real, para a Língua inglesa que, em sincronização com a fala e movimentos dos lábios, combina entoação, movimentos e expressões faciais. O programa consiste num conjunto de três threads independentes que usam filas para a sua coordenação e comunicação. O diagrama da Figura 2 é ilustrativo dessa arquitectura.

A animação é conseguida fornecendo ao programa texto anotado com informação não só sobre a fala, mas também sobre os movimentos da cara. Tudo isto é possível recorrendo a comandos de alto nível permitidos pelo programa de animação. Estes comandos permitem controlar:

- Na voz
  - a acentuação do pitch (accent)
  - a gama do pitch (register)
  - o tom (tone)

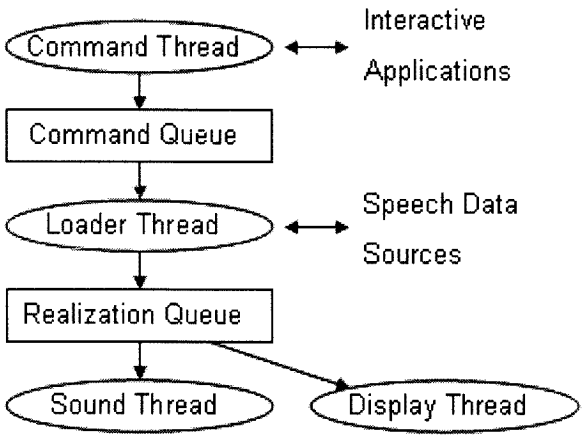


Fig. 2 - Arquitectura do RUTH

- Movimentos
  - o piscar dos olhos (blink)
  - o movimento das sobrancelhas (brow)
  - o movimento da cabeça (jog)
  - o sorriso (smile)

No que diz respeito ao modelo, os seus autores adoptaram a ilustração, por oposição ao foto-realismo, a fim de obterem um resultado atractivo dentro de limites computacionais razoáveis. Além disso, é propositada a ambiguidade em relação ao sexo, raça e idade do modelo.

Ao longo deste ponto, tentou dar-se uma ideia do funcionamento do programa de animação. No entanto, note-se que este programa é composto por mais de 30 ficheiros de código C e outros tantos que mapeiam os visemas [8] e fonemas utilizados para a correcta movimentação da boca, bem como código de programação gráfica que controla a movimentação da face de uma forma geral. Embora não tenham sido todos exaustivamente estudados, foi necessário perceber a sua estrutura e adquirir um grau de confiança sólido para poder manipulá-los de forma correcta, sem prejuízo para as funcionalidades existentes. Como se deve depreender, este é um programa com alguma complexidade e embora haja documentação disponível, esta centra-se essencialmente na sua utilização e concepção, de uma forma geral, não tendo o intuito de o modificar e adaptar, conforme era pretendido neste projecto.

B. Festival Speech Synthesis System

O programa Festival [1] é um sintetizador Text-to-Speech, sendo ele que permite ao RUTH realizar a síntese do texto escrito, conseguido pelo seu funcionamento em modo de servidor. Para os sistemas funcionarem conjuntamente, a comunicação é feita recorrendo a sockets.

Considerando os objectivos inicialmente propostos, apenas seria necessário conhecer este programa ao nível do utilizador, mas a partir do momento que se decidiu construir a nova voz, foi inevitável compreender de uma forma mais aprofundada o seu funcionamento.

Foi realizada a instalação do programa de síntese Festival Speech Synthesis System (versão 1.4.2), em Linux, sendo, para o seu correcto funcionamento, necessário instalar as "speech tools". Foi ainda necessário instalar o OGIresLPC (Residual LPC synthesizer), requerido pela voz brasileira instalada em seguida, tendo-se também procedido à instalação das vozes disponíveis.

IV. ADIÇÃO DE NOVAS FUNCIONALIDADES

Nesta secção serão descritas as funcionalidades acrescentadas às já permitidas pelo programa de animação. Tentando fazer uma abstracção ao código propriamente dito, pretende-se explicar, de uma forma simples, as soluções encontradas para a sua implementação.

Para contemplar todas as alterações que serão referidas, foi necessário alterar o código do programa de animação da cara e construir a interface, de forma conveniente.

No que diz respeito à síntese de voz, as funcionalidades acrescentadas prenderam-se com a instalação e disponibilização de uma voz brasileira, "aga_diphone", e de um sintetizador de voz baseado na concatenação de fonemas, o MBROLA [14], e as respectivas bases de dados disponíveis para as vozes da Língua inglesa, e portuguesa.

O objectivo da adição destas funcionalidades foi, tal como referido, a tentativa de criar um sistema para o português. O método de construção da voz é explicado de uma forma superficial na secção seguinte.

Como consequência directa da introdução de novas Línguas, tornou-se evidente a necessidade de acrescentar os fonemas em falta ao ficheiro de mapeamento existente (para o Inglês), adaptando-o assim às vozes portuguesas usadas. Esta alteração é fundamental, pois os fonemas variam consoante a Língua, e quando aparece um fonema que o RUTH não conhece, não sabe como o há-de traduzir no movimento adequado da boca.

No que diz respeito aos comandos, as funcionalidades acrescentadas relacionam-se com a adição ao código fonte do RUTH, código que permite a comutação para qualquer voz existente no Festival. A implementação desta funcionalidade teve como objectivo, não só possibilitar a comutação interactiva da voz (pelo utilizador), mas também a utilização de mais de uma cara com vozes distintas.

Até este ponto, era possível correr em simultâneo várias caras, mas como a definição da voz era feita no ficheiro de configuração, lido no início da execução do programa Festival, em modo servidor, a voz usada tinha obrigatoriamente de ser a mesma.

Para ultrapassar a limitação referida, foi adicionado ao programa de animação um novo comando, *voice*, através do qual é possível, no decorrer da aplicação, mudar para qualquer voz actualmente instalada no Festival. Para evitar uma sobrecarga desnecessária da interface, e uma vez que se pretende demonstrar essencialmente o uso do Português, apenas é permitida pela interface a escolha entre as vozes da Língua portuguesa (Português Europeu e Português Brasileiro) e inglesa.

Foi criada uma nova opção da linha de comandos, - "position"- que deve ser acompanhada de 2 valores: as coordenadas (x, y) que indicam a posição onde a janela com a cara deve aparecer. Estes valores variam consoante o número da cara, identificado pelo pipe, sendo geridos pela interface. Este comando foi criado para evitar sobreposições das janelas com as caras, evitando, assim ao utilizador o trabalho de ter de as reposicionar no ecrã.

V. CONSTRUÇÃO DA VOZ PARA O PORTUGUÊS

O processo de construção da voz portuguesa passou pela compreensão do modo de construção e funcionamento das vozes sintetizadas que recorrem à técnica de concatenação de difones. Deve ter-se em mente que o MBROLA não é um sintetizador Text-to-Speech, já que não recebe texto simples, mas, em vez disso, aceita informação prosódica e uma lista de fonemas, produzindo amostras de voz. Quando usado juntamente com o Festival, é possível obter uma saída áudio.

A utilização de um programa de conversão de grafemas para fonemas, *String2phon*, disponibilizado pelo grupo de investigação interdisciplinar do IEETA e Centro de Línguas e Culturas da Universidade de Aveiro, traduziu-se num melhoramento substancial da qualidade da voz.

Seguindo o manual do Festival, o método para definir uma nova voz passa pela definição de parâmetros para as várias partes que constituem a voz. Recorreu-se aos exemplos das vozes espanhola e inglesa disponíveis. A sua programação é feita em código Scheme, tendo sido necessário aprender a interpretá-lo para o conseguir utilizar convenientemente.

De seguida, apresenta-se uma breve descrição das partes constituintes da voz, que serviram de guias para a construção de um protótipo de uma voz portuguesa, que usa a base de dados de uma voz feminina (disponível no site do projecto MBROLA [14]).

- **Fonemas** - O bloco básico para a construção de uma nova voz é a definição dos seus fonemas. Esta é uma parte fundamental, já que outros blocos serão construídos à custa deste. É neste bloco que, para cada fonema, se definem os parâmetros característicos que os distinguem. Faz-se a distinção entre vogal e consoante, o tamanho e tonalidade das vogais, o arredondamento dos lábios, o tipo de consoante, e sítio de articulação.
- **Léxico** - Neste bloco são definidas as regras de conversão grafema – fone, conhecidas por regras LTS (Letter To Sound), bem como o dicionário para o Português. Basicamente, consiste em determinar como cada letra ou conjunto de letras se pronuncia, recorrendo ao contexto fonético em que se situa. Para as excepções, palavras difíceis, ou palavras que simplesmente soam mal segundo estas regras, existe um dicionário, onde é definida, manualmente, a sua transcrição. No caso da voz construída, esta tarefa

foi delegada num programa externo, já referido. Optou-se por esta alternativa, já que a construção destas regras não é uma tarefa trivial, ficando fora do âmbito do trabalho apresentado.

- **Fraseamento** - Com este bloco deveria ser possível prever as quebras de frase. Na versão simples adoptada as pausas são determinadas com base apenas na pontuação.
- **Entoação** - Neste bloco deveriam ser definidas as regras de entoação para o Português, pela alteração do pitch do falante. Mais uma vez, a previsão da acentuação das sílabas não é uma tarefa fácil, e o *String2phon* resolve, parcialmente, o problema, de uma forma bastante razoável, para a maioria das palavras.
- **Duração** - Este bloco é também uma parte com substancial importância, já que é aqui que são estabelecidas as durações de cada fonema. Estes valores advêm de médias calculadas a partir de valores medidos. Quanto mais adequados forem estes valores, mais natural resultará a voz.

É perfeitamente plausível assumir que foi criada uma base, sendo de salientar que, embora esteja utilizável, esta voz se encontra num estado primário de desenvolvimento, necessitando ainda de inúmeros ajustes.

Uma vez que a sua criação não era objectivo principal deste projecto, não foi afectado tempo à tarefa de melhoramento da voz, nem tal seria possível.

Os testes efectuados englobam exemplos variados, recorrendo a notícias retiradas de jornais on-line, excertos de textos sobre museus, disponíveis em várias línguas, nos sites oficiais, etc.

VI. ANOTAÇÃO DO TEXTO

Como foi dito anteriormente, o objectivo de possuir texto anotado é o de reforçar as ideias chave, contribuindo de alguma forma para a inteligibilidade da conversação. Analogamente, é vantajoso ter uma forma simples de efectuar a sua anotação, que permita a utilização desta facilidade, sem ter de perceber os detalhes técnicos do RUTH.

Ao longo desta secção, serão explicadas as anotações em causa e o processo para obter um texto correctamente anotado. Para o efeito, aplicou-se uma tecnologia actual – a *eXtensible Markup Language (XML)* [16-19].

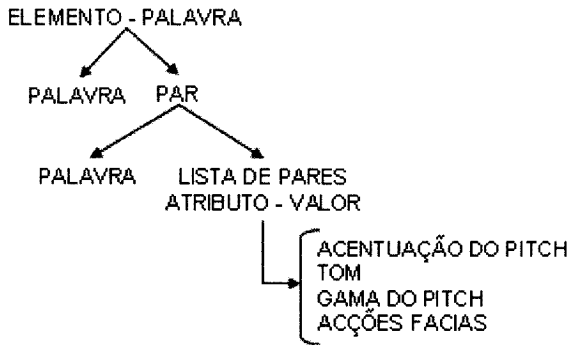
A interface construída evoluiu no sentido da utilização de documentos XML como texto de entrada ao programa de animação. Esta abordagem permite um controlo sobre a introdução dos comandos de alto nível disponibilizados, operando sobre a voz e a cara, com o objectivo de facilitar a sua correcta utilização.

Tendo em vista este objectivo, foi construída uma *Document Type Definition (DTD)* [12] que contém as regras pretendidas para o XML, onde se define o tipo de transformações que a voz e a cara podem sofrer (é de referir que os atributos das marcas não estão tão explícitos como pretendido, limitação resultante da impossibilidade de definir caracteres especiais como elementos enumerados de uma DTD).

Para cada tipo de anotação, existe uma marca de abertura, que possui as propriedades (valor), e uma marca de fecho. No caso das marcas vazias, estas não possuem qualquer atributo. É apresentado de seguida um exemplo ilustrativo de um texto anotado com algumas das marcas XML referidas:

```
<?xml version="1.0" ?>
<!DOCTYPE doc (View Source for full doctype...)>
<doc>
   <SPEAKER nr="1">
      Considerada a mais notável colecção do
      <ACCENT type="L">mundo</ACCENT>
      <BLINK/>
      permite ao visitante compreender não só a
      evolução técnica dos transportes de tracção
      <REGISTER type="HL">
         <BROW type="raise">animal</BROW>
      </REGISTER>
      como acompanhar as mudanças tão bem
      <JOG type="forward">expressas</JOG>
      na ornamentação das viaturas.
   </SPEAKER>
</doc>
```

O parsing das marcas XML é feito pela interface, que as transforma em código Scheme, interpretável pelos programas de animação e síntese. Para melhor se entender o esquema de anotação utilizado, deve considerar-se a seguinte hierarquia:



Analisando o esquema conjuntamente com o exemplo apresentados, verifica-se que as anotações do texto são processadas tendo como base um "elemento-palavra" que pode ser apenas a própria palavra ou uma lista constituída por um par (palavra, (atributo, valor)).

Para as anotações terem o comportamento pretendido, deve ter-se em consideração um conjunto de regras e restrições de uso. Algumas destas limitações são consequência do próprio programa de animação, outras do parsing feito pela interface. Note-se que este parsing não traz apenas limitações, mas também vantagens, uma vez que é mais robusto em diversos aspectos.

Uma das regras utilizadas define que é apenas permitida uma marca de cada tipo, para cada palavra, o que faz todo o sentido. Tome-se como exemplo a marca "ACCENT": não faz sentido marcar a mesma palavra com um pitch alto e baixo.

No que diz respeito às restrições, estas não limitam substancialmente a marcação do texto, mas devem ser tidos em atenção os seguintes aspectos: as marcas não podem englobar mais de uma palavra; não são permitidas marcas em palavras seguidas; as marcas vazias devem ser colocadas depois de alguma palavra e nunca no início de uma frase.

Marcas mal formadas (atributo ou valor inválido) são eliminadas, não permitindo que um pequeno erro impeça todo o texto de ser sintetizado.

Outra regra extremamente importante é a que determina a necessidade da existência das marcas de abertura e fecho "SPEAKER", uma vez que, num universo multi-falante, é esta a marca que define qual a cara a que o texto se destina. Faz sentido que todo o texto fora destas marcas seja ignorado.

Como conselho de utilização, deve recorrer-se frequentemente ao uso de pontuação final (".", "!" ou "?"), uma vez que as regras de fraseamento apenas recorrem à pontuação para produzir pausas no texto. A restante pontuação (",", ";", ":", "(" e ")") é retirada do texto, porque não é aceite pelo RUTH ou, ainda, porque interfere com as marcas, como é o caso dos parênteses.

A síntese é feita de uma forma transparente para o utilizador, sem o conhecimento dos formalismos requeridos anteriormente, quer pelo programa de animação, quer pelo próprio sintetizador, tais como o uso de comandos específicos, parênteses e linhas em branco. No caso de se pretender anotar o texto, é apenas necessário ter em conta as observações feitas anteriormente.

VII. APLICAÇÃO

Embora se pretenda que a aplicação construída seja uma demostração dos componentes desenvolvidos, o principal objectivo da sua elaboração foi o de tornar o programa de animação utilizável por qualquer pessoa, isto é, por utilizadores que não possuam qualquer conhecimento sobre o seu funcionamento. Anteriormente à implementação da interface, a utilização do RUTH era feita extensivamente através da linha de comandos e, de uma forma muito limitada, através do menu que apenas permitia o acesso a demonstrações da sua aplicabilidade.

Esta aplicação não pode ser vista simplesmente como uma interface gráfica, já que não tem como função apenas facilitar a execução de instruções de uma forma intuitiva através de menus e botões, mas também é ela quem controla todos os processos e realiza algumas funções mais específicas, como o parsing das marcas XML.

Tendo em vista os objectivos mencionados, o primeiro passo foi a instalação do Qt C++ toolkit, para Linux. A escolha de desenvolver a interface gráfica, recorrendo a essa biblioteca de classes, deveu-se a esta ser multiplataforma e de fácil uso e integração com OpenGL.

Por uma questão de simplicidade, a abordagem foi a de criar a interface de uma forma independente do programa de animação, recorrendo ao uso de *pipes* para a comunicação entre eles. Para tornar isso possível, foi necessário uma reimplementação da comunicação entre os vários processos, redireccionando a forma como o RUTH recebe a informação: anteriormente pela linha de comandos, agora pelo unnamed pipe.

Ao iniciar a aplicação, é solicitado ao utilizador que escolha o número de falantes pretendidos. Foi estipulado um número máximo de 4 falantes, que pareceu um valor razoável, tendo em vista os objectivos pretendidos.

Existe um menu de opções, apresentado na Figura 3, que permite a qualquer altura, para cada falante, comutar entre as várias vozes existentes. Pelos motivos já mencionados, o sistema apenas está configurado para permitir a comutação entre vozes da Língua Portuguesa (Europeu e Brasileiro) e da Língua Inglesa.

Devido à sua arquitectura, o programa de animação permite determinar facilmente se existe texto a ser sintetizado e, através da troca de mensagens com a interface, é possível evitar por completo a sobreposição das vozes, no caso de existir mais de um falante. Para além disso, esta troca de mensagens serve também para facilitar o acesso a mais informação sobre a eventual ocorrência de erros no decorrer do programa, permitindo fornecer mensagens de erro mais específicas.

Na Figura 4 apresenta-se o ecrã exemplo da interface, acompanhado de duas janelas com as caras. Além de ser mais funcional, uma apresentação gráfica é sempre mais apelativa do que escrever texto numa consola.

Para além das funcionalidades descritas, continua a ser possível guardar e carregar ficheiros com informação

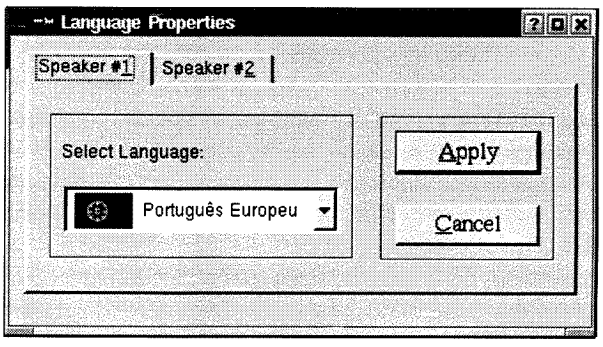


Fig. 3 - Menu de opções

sobre animações previamente realizadas, estando estas funcionalidades acessíveis no menu *Options*.

VIII. CONCLUSÕES

Após um estudo mais ou menos extensivo dos programas intervenientes e tecnologias envolvidas, como resultados relevantes conseguidos salientam-se: o uso de novas vozes pela cara falante; a construção de um protótipo de uma voz de Língua portuguesa; o uso de XML para anotação

do texto de uma forma simples e directa, não só respeitante à voz, mas também à animação da cara em si e devido uso das mesmas; a inserção de novos comandos

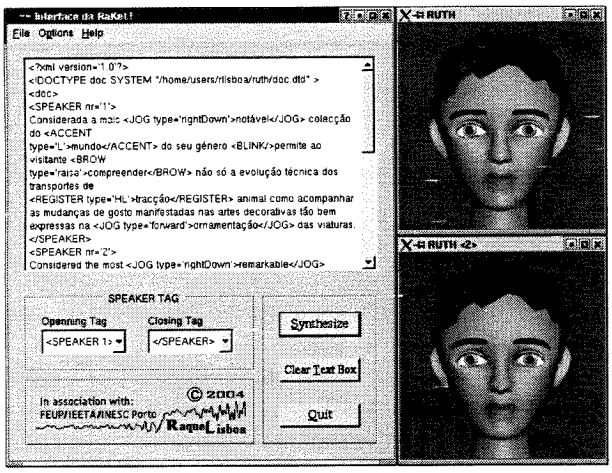


Fig. 4 - Exemplo da interface, com 2 caras

que permitem não só a comutação interactiva da voz pelo utilizador, mas também a utilização de mais de uma cara com vozes distintas.

A evolução natural do projecto levou à construção de uma API simplificada, que permite uma utilização intuitiva por parte do utilizador. Ao manter as funcionalidades existentes e adicionando outras, foi possível aumentar largamente as capacidades e aplicações do programa de animação.

Mais uma vez se reforça a ideia de que os programas envolvidos são complexos e que, por essa razão, o tempo empregue no seu estudo e compreensão não é desprezável.

Como já foi sendo referido ao longo deste artigo, existem diversas áreas onde há ainda lugar para evolução, nomeadamente, no que diz respeito à construção da voz de Língua portuguesa e à anotação automática do texto. Apesar de não estar previsto um envolvimento tão aprofundado na área da síntese de voz, é importante referir que todo o trabalho desenvolvido nessa área se apresenta como um contributo imprescindível para este projecto. No que diz respeito à anotação do texto, seria interessante poder automatizar o processo, quer baseado na pontuação, quer na própria pré-análise do texto a sintetizar. A título de exemplo, pode referir-se a mudança de tom na presença de um ponto de interrogação ou o movimento da cabeça na presença incontestável de afirmações ou negações.

Consideramos que os objectivos inicialmente propostos foram atingidos e, alguns, inclusivamente superados.

Tratou-se de um projecto interessante, educativo e bastante apelativo, numa área em constante movimento e desenvolvimento. A utilização de animação facial em conjunto com voz, em tempo real, desempenhará um papel importante no futuro das comunicações em geral.

Para informações mais detalhadas sobre este Projecto, pode consultar-se a página da internet eleborada para o efeito, disponível em:

http://pwp.netcabo.pt/formatodecores/raquelisboa/projecto.html

AGRADECIMENTOS

Agradecemos todo o apoio prestado durante o desenvolvimento do projecto pelo Eng.º Luís Gustavo Martins (INESC Porto) e pela Professora Lurdes de Castro Moutinho (DLC/Universidade de Aveiro).

REFERÊNCIAS

[1] A. Black, P. Taylor e R. Caley, "The Festival Speech Synthesis System" (http://www.cstr.ed.ac.uk/projects/festival)

[2] D. DeCarlo e M. Stone, "The Rutgers University Talking Head: RUTH" (http://www.cs.rutgers.edu/~village/ruth)

[3] C. Benoît e B. Le Goff, "Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP" , Speech Communication, Vol 26 , n 1-2, Special issue on auditory-visual speech processing, p 117 - 129 , 1998.

[4] J. Hulstijn et al., "Dialogues with a Talking face for Web-based Services and Transactions", In Proceedings Face to Face Workshop, CWI, Amsterdam. Available as CTIT Technical report 99-07.

[5] P. Cosi, A. Fusaro e G. Tisato, "LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model", Proc. Eurospeech 2003.

[6] A. Newell, "Spoken Language and e-inclusion", Proc. Eurospeech 2003.

[7] I. Karlsson, A. Faulkner e G. Salvi, "SYNFACE – a talking face telephone", The Eurospeech Special Event on "Spoken Language Technology in E-inclusion", Proc of EuroSpeech 2003.

[8] J. De Martino e L. Magalhães, "Um Conjunto de Visemas para uma Cabeça Falante do Português do Brasil",Congresso Iberdiscap, Tecnologías de Apoyo para la Discapacidad, Vol. 1, pp.198-203, San José, CR, 2004.

[9] BALDI (http://itakura.kes.vslib.cz/kes/baldie.html)

[10] CSLU toolkit (http://www.cslu.ogi.edu/tts/)

[11] CUANIMATE (http://cslr.colorado.edu/beginweb/cuanimate/cuanimate_paper.html)

[12] DTD Tutorial (http://www.w3schools.com/dtd/default.asp)

[13] INOVANI (http://www.inovani.no/TechSpeech.htm)

[14] MBROLA Project (http://tcts.fpms.ac.be/synthesis/mbrola.html)

[15] TrollTech – Qt C++ toolkit (http://www.trolltech.com)

[16] XED: An XML Document Instance Editor (http://www.cogsci.ed.ac.uk/~ht/xed.html)

[17] XML 1.0 – Extensible Markup Language (W3C) (http://www.w3.org/TR/1998/REC-xml-19980210)

[18] XML Developer Center (http://www.msdn.microsoft.com/xml/)

[19] XML FAQ (http://www.ucc.ie/xml/)