Ambiente interactivo de visualização científica de áudio em tempo-real

Francisco Pinto
INESC Porto / Faculdade de Engenharia da Universidade do Porto
fr.miguel@sapo.pt

Gisela Costa, Hugo Soares, João Soares, Paulo Pinto
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
{ee01211, ee01171, ee01173, paulo.pinto}@fe.up.pt

Aníbal Ferreira
Faculdade de Engenharia da Universidade do Porto / SEEGNAL Research
a.j.ferreira@ieee.org

RESUMO

Apresenta-se um ambiente vocacionado para a representação didáctica e interacção lúdica com características do som, através da integração de técnicas avançadas de processamento de sinal, visualização gráfica 2D e animação 3D por computador, em tempo-real. Descrevem-se os principais módulos gráficos e de processamento de sinal do ambiente, refere-se a sua implementação observando requisitos de tempo-real, e alude-se à sua funcionalidade interactiva. Perspectivam-se também futuros desenvolvimentos e áreas de aplicação.

INTRODUÇÃO

A experiência da audição é intrínseca à existência humana, condicionando o indivíduo, por exemplo, no seu relacionamento inter-pessoal, no seu estilo de vida, e até mesmo nas suas preferências estéticas. Contudo, apesar dos resultados de investigação e dos imensos avanços tecnológicos verificados nas últimas décadas, o mecanismo da audição permanece, ainda hoje, largamente desconhecido, particularmente no que toca à sua impressionante capacidade de interpretar e segregar sons.

Os resultados que aqui se apresentam, traduzem uma preocupação que teve a sua génese num projecto Ciência Viva[1], de ilustrar por via gráfica, a possibilidade de interpretar sons por computador, e de medir com precisão alguns dos seus principais atributos. Aquela preocupação pretende servir dois níveis de interesse: o dos jovens sem conhecimentos específicos que encaram o ambiente desenvolvido como uma plataforma lúdica incentivando a experimentação e a descoberta; e também o interesse de pessoas com conhecimentos sobre o som e audição que encaram o ambiente desenvolvido como uma plataforma de análise científica e investigação de sinais de fala /áudio. Quer uma perspectiva quer a outra dependem fortemente do requisito de funcionamento em tempo-real que foi perseguido (e conseguido) na realização do ambiente interactivo.

Esta comunicação descreve a estrutura e realização do ambiente desenvolvido. Abordam-se os módulos de análise de sinal e os módulos gráficos 2D e 3D do ambiente, destacando, em cada caso, as suas valências interactivas. Conclui-se com uma perspectiva sobre futuros desenvolvimentos e com uma síntese sobre o significado dos resultados alcançados.

MÓDULOS DE PROCESSAMENTO DE SINAL EM TEMPO-REAL

[1] http://www.inescporto.pt/cienciaviva

Cálculo da tonalidade da voz do orador

A tonalidade da voz do orador, ou *pitch*, é estimada a partir de um algoritmo [1] robusto e eficiente, capaz de detectar múltiplas estruturas harmónicas num só segmento do sinal de entrada. É assim possível detectar a presença de vários oradores em simultâneo e estimar o *pitch* de cada um deles. No entanto, devido à configuração escolhida para o algoritmo, a aplicação admite a identificação, no máximo, de dois oradores.

O algoritmo, que actua unicamente no domínio das frequências, requer o cálculo de uma Odd-DFT [2] (ODFT) e a posterior detecção dos picos espectrais que constituem cada estrutura harmónica. Uma vez determinada a frequência precisa de cada pico, a estimativa do valor do *pitch* obtém-se através da minimização da seguinte função de erro

$$e = \sum_{i=1}^{\text{npeaks}} \left(\text{pos}[i] - i \cdot \text{pitch}\right)^2 ,$$

da qual resulta [1]:

$$\text{pitch} = \frac{\sum_{i=1}^{\text{npeaks}} i \cdot \text{pos}[i]}{\sum_{i=1}^{\text{npeaks}} i^2} .$$

Por fim, é atribuída uma categoria ao orador de acordo com a gama de frequências dentro da qual o seu *pitch* se encontra. O orador será identificado como um homem adulto se a tonalidade da sua voz for menor 160Hz, como mulher adulta se o valor se situar entre 160Hz e 260Hz e como criança se o *pitch* for superior a 260Hz. A descrição abreviada da sequência de operações envolvida encontra-se no diagrama da Figura 3 (ramo superior).

Reconhecimento de vogais

Tal como ocorre na detecção do *pitch*, o módulo responsável pelo reconhecimento de vogais opera exclusivamente no domínio das frequências. O processo parte do cálculo da densidade espectral de potência do sinal de entrada *s[n]*, recorrendo novamente à ODFT, e termina na revelação por via gráfica e em tempo-real, da vogal produzida pelo orador. Para o efeito, são utilizados três blocos de processamento (cuja sequência é ilustrada no ramo central do diagrama da Figura 3): detector de envolvente espectral, detector de formantes e decisão final.

A envolvente espectral é obtida por meio de uma análise LPC [3] da densidade espectral de potência *PSD(ω)*. O método consiste no dimensionamento de um filtro só com pólos, capaz de produzir uma forma de onda semelhante a *s[n]* quando excitado com impulsos ou por ruído branco (dependendo respectivamente da natureza do sinal: vozeado ou não vozeado). Os coeficientes deste filtro são obtidos através de um sistema de equações, conhecidas por equações Yule-Walker [3], que podem ser resolvidas analítica ou numericamente. Na nossa aplicação utilizamos um método recursivo, de Levinson-Durbin [3], capaz de chegar à solução do sistema com pouco esforço computacional. O algoritmo (ilustrado no diagrama da Figura 1) requer apenas o vector de autocorrelação *a[n]* do sinal de entrada que, de acordo com a relação Wiener-Khintchine [3], pode ser obtido através da transformada inversa de *PSD(ω)*. Uma vez calculados os coeficientes do filtro, a envolvente espectral de *s[n]* é estimada a partir da magnitude da resposta em frequência do filtro LPC.

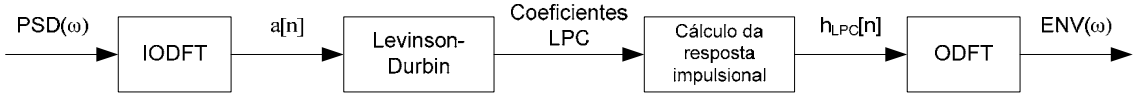

Figura 1: Estimação da envolvente espectral

As frequências das três primeiras formantes (F_1, F_2 e F_3) da estrutura harmónica [3], podem ser identificadas localizando-se os picos mais salientes na envolvente espectral. Se as suas frequências forem suficientemente díspares, cada formante surge como um pico bem definido.

No entanto, quando duas formantes têm frequências muito próximas, os dois picos fundem-se num só. Para contornar este problema, o nosso sistema efectua uma análise de concavidade da envolvente espectral [4], recorrendo à segunda derivada da função, tal como se descreve no diagrama da Figura 2. Desta forma, as formantes surgem como picos mais isolados e facilmente identificáveis.

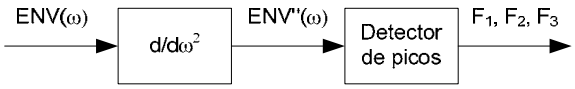

Figura 2: Estimação de formantes

O número de picos a detectar depende do número de formantes que se pretendem obter. Para efeitos de representação, a aplicação estima a frequência das três primeiras formantes. No entanto, apenas as duas com frequência mais baixa são utilizadas para a identificação da vogal. Uma vez conhecidos os valores de F_1 e F_2, estes são representados como um ponto (F_1,F_2) num espaço bidimensional onde cada eixo representa a frequência de uma formante. A localização deste ponto no espaço, dependendo da área de decisão onde se encontra, identifica qual a vogal produzida pelo orador.

MÓDULOS GRÁFICOS 2D

A visualização gráfica e a interactividade são dois aspectos fulcrais no ambiente desenvolvido. No modo de funcionamento 2D, é possível visualizar em tempo-real o resultado das várias fases de processamento, sobre a forma de valores, gráficos, *scatterplots* e até desenhos sugestivos, proporcionando assim uma interface de aparência simpática e de fácil utilização. Estas condições são importantes no sentido de promover a interactividade com o utilizador. A Figura 3 esquematiza a sequência das operações de processamento de sinal e assinala os pontos onde é extraída informação que controla a actualização da interface gráfica, tal como se pormenoriza nas subsecções seguintes.

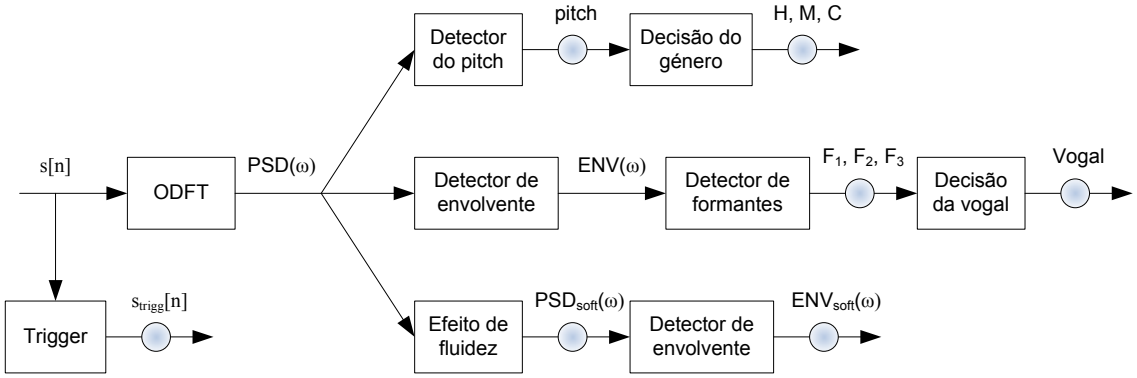

Figura 3: Pontos de visualização (assinalados por ◯)

Representação de sinais nos tempos

Um osciloscópio correctamente calibrado é capaz de representar um sinal periódico completamente estabilizado no tempo, através de técnicas de *triggering*. Pretende-se, da mesma forma, que o sinal de voz captado pela aplicação seja visualizado de forma estável.

Quando se divide um sinal em segmentos, tal como acontece na captação em tempo-real, o tamanho da janela de visualização nunca equivale a um múltiplo exacto do período do sinal. Do ponto de vista do utilizador, este efeito é semelhante a um sinal a deslizar no tempo. Para "travar" o sinal é portanto necessário recorrer a uma técnica de *triggering*. O mecanismo

utilizado no nosso sistema consiste em deslocar no tempo o segmento actual até ao ponto de correlação máxima com o segmento anterior, sendo necessário apenas aumentar a dimensão do *buffer* de entrada em 17% [4]. O resultado é um sinal estacionário que permite ao utilizador observar a natureza periódica desse sinal. A Figura 4 reproduz o painel da interface gráfica onde se representa o sinal áudio no domínio dos tempos.

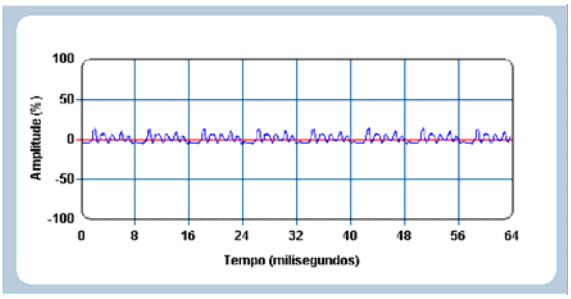

Figura 4: Representação de um sinal no domínio dos tempos

Representação de sinais nas frequências

A representação de um sinal no domínio das frequências é bastante reveladora das suas características e particularidades, especialmente se aquele possuir um carácter harmónico. Por esta razão, o ramo inferior do diagrama da Figura 3 é dedicado exclusivamente à representação gráfica do espectro do sinal de entrada. A janela de visualização nas frequências, ilustrada na Figura 5, exibe quatro aspectos do segmento de sinal: densidade espectral de potência, envolvente espectral, componentes (ou parciais) da estrutura harmónica e localização das formantes.

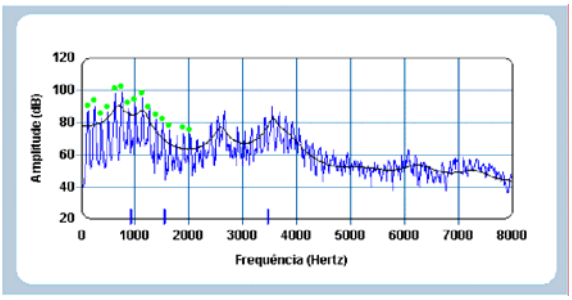

Figura 5: Representação de um sinal no domínio das frequências

O gráfico da densidade espectral de potência visualizado na janela não consiste na função instantânea *PSD(ω)*, mas sim uma versão suavizada da mesma. Esta suavização é feita através de uma filtragem passa-baixo, aplicada ao eixo dos tempos (perpendicular ao plano da figura), dada por

$$PSD_{soft\,\omega}[n] = \frac{1-\alpha}{2} \cdot PSD_{\omega}[n] + \frac{1-\alpha}{2} \cdot PSD_{\omega}[n-1] + \alpha \cdot PSD_{soft\,\omega}[n-1],$$

onde $0 \leq \alpha < 1$. A filtragem produz um efeito de fluidez que preserva contudo a velocidade de resposta a eventos de grande dinâmica sonora. O valor de α deve ser aumentado de acordo com o peso pretendido para a influência da memória, que é responsável pelo retardamento no regresso da função à forma de repouso (em silêncio). A Figura 6 representa a resposta em frequência do filtro quando α = 0.1 .

A envolvente espectral é obtida através do método de reconhecimento de vogais já descrito. Contudo, neste caso, o diagrama da Figura 1 toma por entrada $PSD_{soft}(\omega)$, em vez de $PSD(\omega)$.

Uma vez identificada a estrutura harmónica do segmento de sinal, cada um dos seus parciais é assinalado por uma bola numa posição ligeiramente acima da sua magnitude espectral (Fig. 5).

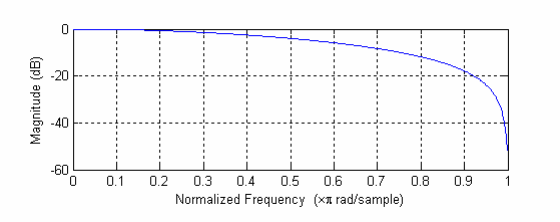

Figura 6: Resposta em frequência do filtro de suavização, com α = 0.1

Por fim, caso o sistema detecte um conjunto de formantes F_1, F_2 e F_3, as suas posições são assinaladas sobre o eixo das frequências. Os seus valores, juntamente com o valor do *pitch*, são também visualizados num painel de valores, semelhante ao da Figura 7.

Como referido anteriormente, os valores da 1ª e 2ª formante constituem um ponto $p \rightarrow (F_1,F_2)$ num espaço bidimensional. O mapa de vogais exibe em tempo-real, através de bolas coloridas, a posição dos sucessivos pontos *p* determinados para cada segmento do sinal de entrada. O orador poderá então tirar conclusões acerca do som produzido, observando a posição dos pontos relativamente às áreas de decisão de cada vogal, como se ilustra na Figura 8.

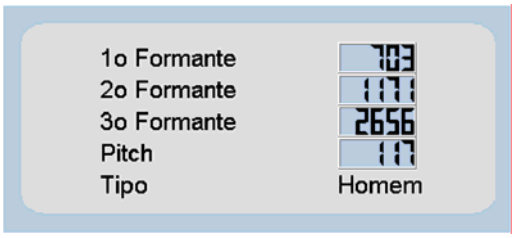

Figura 7: Parâmetros da voz do orador

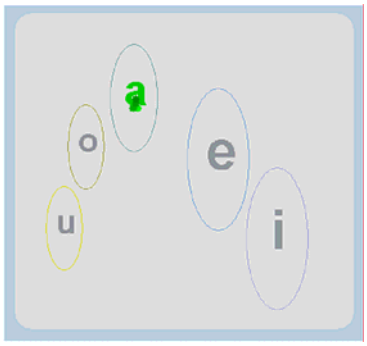

Figura 8: Mapa das vogais

Funcionalidade interactiva

A aplicação disponibiliza vários níveis de interactividade, que permitem despertar o interesse de utilizadores com diferentes níveis de conhecimento. Um utilizador experiente poderá tirar o máximo partido de todos os módulos de representação gráfica descritos nas secções anteriores. No entanto, a natureza ilustrativa do mapa de vogais (Figura 8) proporciona um bom ponto de partida para um utilizador não conhecedor, que pretenda compreender o processo de análise de voz e de reconhecimento de vogais. Numa perspectiva de mais alto nível, o utilizador pode limitar-se a visualizar a vogal identificada. Adicionalmente, a identificação do género do orador é ilustrada por imagens simpáticas (reproduzidas na Figura 9) que têm o objectivo de atenuar sentimentos de carácter ofensivo que possam resultar de uma decisão errada do algoritmo.

Figura 9: Personagens utilizadas para representar as três categorias (© Pixar Animation Studios)
(a) Homem adulto (b) Mulher adulta (c) Criança

Está também prevista a possibilidade de reproduzir sons naturais e sintéticos pré-armazenados, que passam pelo mesmo processo de análise descrito nas secções anteriores.

MÓDULO GRÁFICO 3D

Foi criado um módulo gráfico 3D com o objectivo de convidar à interacção, através da animação de figuras de Lissajous em função do som ambiente. Este módulo é baseado em *Open Graphics Library* (OpenGL) e é activado logo no arranque da aplicação, comutando para o ambiente de interacção mais informativo com as funcionalidades atrás descritas, quando se pressiona um qualquer tecla. Nesta secção detalha-se a implementação e funcionalidade deste módulo interactivo.

OpenGL

O OpenGL é uma interface de *software* para dispositivos de *hardware*. Consiste uma biblioteca gráfica de modelagem e exibição tridimensional, bastante rápida para vários sistemas operativos. Os seus recursos permitem criar objectos gráficos com qualidade e rapidez, além de incluir recursos avançados de animação, tratamento de imagens e texturas. Todas as rotinas do OpenGL são implementadas em C, facilitando assim a sua utilização em qualquer programa escrito em C ou C++. Entre os recursos gráficos disponíveis pelo OpenGL, podem ser destacados os seguintes: Modos de desenho de pontos; Ajuste de largura de linhas; Aplicação de transparência; Activação/desactivação de "*aliasing" (efeito de escada)*; Mapeamento de superfícies com textura; Selecção de janela de desenho; Manipulação de fontes/tipos de iluminação e sombreamento; Transformação de sistemas de coordenadas; Transformações em perspectiva; Combinação de imagens (*blending*). O OpenGL é portanto vocacionado para a animação e reconfiguração paramétrica por computador de objectos tridimensionais, em tempo real, em resultado da análise instantânea de eventos sonoros.

Figuras de Lissajous

As figuras de Lissajous são sintetizadas dinamicamente no ambiente interactivo desenvolvido usando expressões matemáticas e um conjunto de esferas alinhadas segundo a trajectória tridimensional definida por essas expressões [5]. As expressões simplificadas de uma figura de Lissajous a duas dimensões, são as seguintes

$$\begin{cases} x(t) = \sin(t) \\ y(t) = \sin(a.t + b) \end{cases}$$

ilustrando-se na Fig. 10 alguns casos particulares simples.

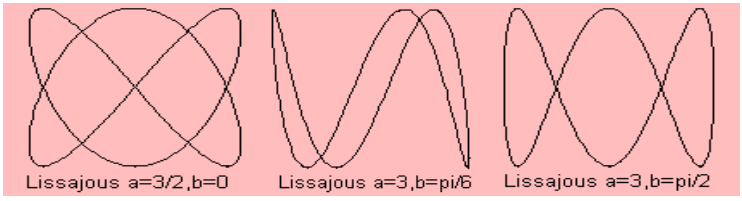

Figura 10: Curvas de Lissajous 2D

Para dar profundidade explícita às figuras de Lissajous, utilizaram-se as seguintes expressões:

$$\begin{cases} x(t) = \sin(\omega.t + \delta) \\ y(t) = \sin(t) \\ z(t) = \cos(\omega.t + \delta) \end{cases},$$

sendo o valor de ω instantaneamente ajustado de acordo com o pitch detectado. O valor de δ é controlado de modo a imprimir uma rotação constante à figura de Lissajous. Ao adicionar o eixo z(t) ao sistema de equações, a função y(t) fica distribuída sobre o plano X,Z de forma circular, proporcionando assim um efeito tridimensional à figura.

A trajectória tridimensional definida por estas equações poderia ser construída com base em objecto cilíndricos ou esferas. A opção por esta última opção deve-se ao menor custo computacional associado, dado que no primeiro caso seria necessário calcular derivadas para orientar os cilindros no sentido do percurso da figura de Lissajous.

A Fig. 11 ilustra um realização tridimensional da figura de Lissajous com ω=5, onde se evidenciam os eixos X, Y e Z e onde o raio das esferas elementares é suficiente para sugerir continuidade de forma ao longo da trajectória. A esfera central é um objecto extra inspirado no símbolo do Visionarium que foi adicionado. A animação resulta assim de uma figura de Lissajous 3D que gira em torno de uma esfera central, e com forma mutante de acordo com o som.

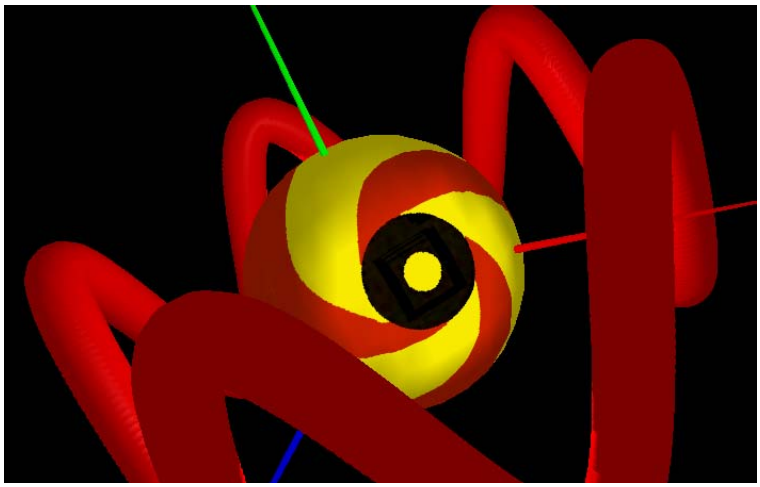

Figura 11: Figura de Lissajous 3D e Esfera Central

A esfera central encontra-se também em constante movimento, rodando em torno do seu eixo, de forma independente do da figura de Lissajous.

Para proporcionar uma maior animação ao módulo foram aplicadas cores, texturas e iluminação ao ambiente. A esfera central cativa a atenção do utilizador através de duas fontes de luz incorporadas e da sua textura helicoidal. A dinâmica da figura de Lissajous foi também acentuada, controlando não só a sua forma e cor, mas controlando também o raio das esferas elementares com se descreve a seguir.

Funcionalidade interactiva

Os atributos da figura de Lissajous variam de acordo com a intensidade sonora e a tonalidade de voz do orador:

- O raio das esferas constituintes da figura de Lissajous é proporcional à intensidade sonora instantânea, ou seja é proporcional à energia do sinal emitido pelo orador ou oriundo de ruído ambiente, gerando assim uma figura com maior ou menor espessura.

- A ordem (i.e. a forma ou complexidade estrutural) da figura de Lissajous e a sua cor, variam de acordo com a tonalidade da voz do orador, segundo uma regra logarítmica.

Existe também a possibilidade de controlar manualmente os parâmetros da animação por computador em tempo real. Para além do raio das esferas e da ordem da figura de Lissajous é possível alterar:

- o raio da esfera central, bem como a sua direcção e velocidade de rotação,
- a variação do número de esferas que constituem a figura de Lissajous, proporcionando uma maior ou menor definição da figura,
- a alteração da velocidade de rotação da figura de Lissajous proporcionando uma maior interacção com o utilizador.

FUTUROS DESENVOLVIMENTOS

O ambiente interactivo aqui descrito foi concebido para utilização no Centro de Ciência do Europarque (Visionarium). As sua valências interactivas sugerem contudo outras áreas de aplicação, nomeadamente em contextos de apoio à pré-aprendizagem da escrita e da fala, ou em contextos de apoio à terapia ou reabilitação da fala. Com este objectivo, está já a ser desenvolvida investigação para melhorar quer a capacidade de reconhecimento de sons [6], quer a robustez funcional do algoritmos utilizados no ambiente interactivo.

CONCLUSÃO

Descreveu-se o projecto, realização e valências interactivas de um ambiente vocacionado para a visualização lúdica e representação científica de alguns atributos importantes dos sinais áudio em geral, e de fala em particular. Este ambiente pretende motivar quer os jovens quer especialistas para a exploração e o conhecimento mais aprofundado das características do som e do funcionamento do sistema auditivo humano. O sucesso já obtido junto do público alvo, motiva também para novos desenvolvimentos privilegiando o som e a fala como o principal *medium* interactivo, nomeadamente em contextos de apoio clínico ou ensino.

A concretização deste ambiente beneficiou do entusiasmo e do trabalho voluntário de várias pessoas, nomeadamente alunos da LEEC (FEUP), a quem se agradece.

REFERÊNCIAS

[1] Aníbal J. S. Ferreira, 1996. Perceptual Coding of Harmonic Signals. 100th Convention of the Audio

Engineering Society, Paper 4177. Copenhagen, Denmark, Maio de 1996.

[2] Aníbal J. S. Ferreira, 1998. Spectral Coding and Post-Processing of High Quality Audio.

Dissertação de Doutoramento, Fac. Eng. da Universidade do Porto, Portugal, Novembro de 1998.

[3] Douglas O'Shaughnessy, 2000. Speech Communications –human and machine.

IEEE Press, 2000.

[4] Vasco Santos, 2002. Analisador de sinais de voz em tempo real. Projecto Final de Curso da Lic. Engª Electrotécnica e Computadores.

[5] Paulo Pinto, 2005. AUDIONARIUM -Plataforma de demonstração científica sobre o som e o sistema auditivo Humana. Projecto Final de Curso da Lic. Engª Electrotécnica e Computadores.

[6] Aníbal J. S. Ferreira, 2005. New Signal Features for Robust Identification of Isolated Vowels. 9th European Conference on Speech Communication and Technology, September 4-8 2005, Lisbon, Portugal.