# Representação Visual do Modelo Articulatório para o Estudo da Produção da Fala

António Branco, Ana Maria Tomé, FranciscoVaz

***Resumo*** **- Neste artigo descreve-se a implementação de um sintetizador articulatório de voz que tem como objectivo ser usado como ferramenta de estudo da produção da fala. Um som pode ser gerado a partir de configurações do tracto vocal editadas pelo utilizador. O simulador também permite estudar o problema inverso, isto é, determinação de uma configuração do tracto vocal correspondente a um sinal de voz.**

***Abstract*** **- In this paper we present the description of an articulatory speech synthesiser which is to be used as a tool for the study of speech production. The sound might be produced as the result of vocal tract configurations edited by the user. It also allows the study of the inverse problem- the estimation of a vocal tract configuration given a speech sound.**

## I. INTRODUÇÃO

A síntese articulatória de voz modela o aparelho vocal com parâmetros fisiológicos que variam lentamente com o tempo(ex. corpo da língua, lábios, velo, etc.) e promete naturalidade em aplicações de texto-para-voz. O tracto vocal pode ser representado com modelos articulatórios de distância sagital [1,5] ou modelos de área [6].

Para a determinação da resposta do sistema a partir do modelo articulatório existem métodos que resolvem o problema no tempo[9], os híbridos no tempo e frequência[3,7] ou do tipo *filtros de onda digitais*[8]. Sondhi [7] implementou o método híbrido no tempo e na frequência em que primeiro é calculada a função de transferência e a resposta impulsional é calculada usando a Transformada Inversa de Fourier e o valor da saída obtido através de convolução (da excitação glotal com a resposta impulsional). Neste artigo descreve-se a implementação de um sintetizador articulatório de distância sagital que combina o método da frequência de Sondhi com um algoritmo que calcula geradores de formantes discretos, ligados em paralelo, a partir da função de transferência [3].

A determinação da forma do tracto vocal a partir do sinal de voz é um problema que perdura, devido à não unicidade do mapeamento acústico-articulatório Schroeder[11], Wakita[16], Sondhi[14] and Atal[1]. Schroeter *et al.*[12] usaram um *codebook* combinado com programação dinâmica. Rahim *et al.*[10] utilizaram vários perceptrões multicamada, cada um para uma zona específica o espaço articulatório. Neste trabalho foi utilizado um conjunto de redes neuronais de Kohonen [3,4,5,6] para criar um *codebook* e fazer o mapeamento inverso acústico-articulatório.

## II. MODELO ARTICULATÓRIO

É usado um modelo articulatório de distância sagital baseado no modelo de Mermelstein[1] com algumas modificações introduzidas por investigadores do *Mind-Machine Research Center, University of Florida*[2].

Os pontos representados na Fig. 1 são os articuladores principais. O ponto V é o Velo, o ponto C representa o centro da língua, o ponto T representa a ponta da língua, o ponto J representa o maxilar, o ponto P serve para determinar o contorno superior do osso hióide, LIPP indica o alongamento dos lábios, LIPO indica a abertura dos lábios. A parte inferior da faringe é descrita por WH, HK1 e G1K.

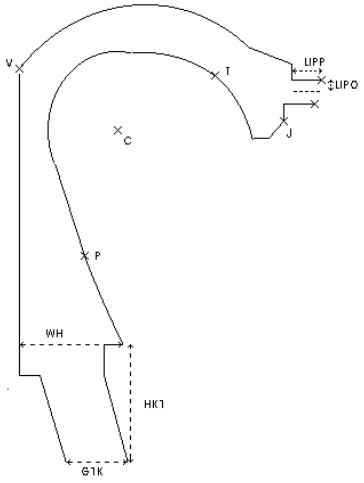


Fig. 1 - Modelo articulatório de distância sagital.

Uma vez especificadas as posições dos articuladores, as dimensões sagitais são calculadas, sobrepondo uma grelha não uniforme sobre o traçado do tracto vocal (ver Fig. 2). A grelha divide o tracto vocal em seis regiões diferentes com um total de 60 secções. Todas as linhas da grelha são quase perpendiculares ao traçado do tracto vocal e são calculadas tendo em conta que nas zonas de fronteira entre regiões têm que ter inclinações semelhantes.

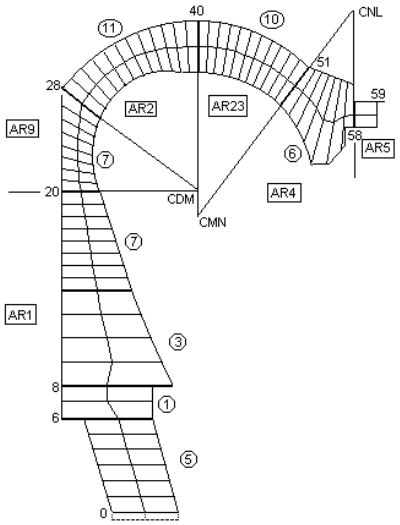


Fig. 2 - Grelha não uniforme usada para a determinação das distâncias sagitais.

As dimensões sagitais são o comprimento do segmento da linha da grelha interior ao tracto vocal.

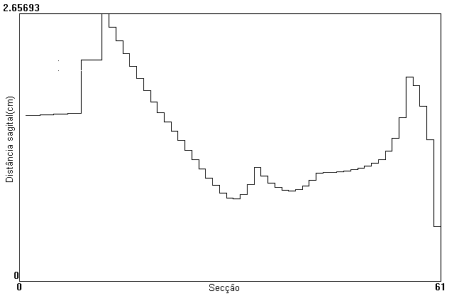


Fig. 3 - Exemplo de distâncias sagitais ao longo do tracto vocal.

As dimensões sagitais(ver Fig. 3), são convertidas para áreas seccionais usando diferentes formulas, dependendo da região do tracto vocal tal como descrito em Mermelstein[1]. Por exemplo na região faringea(AR1), tal como na região dos lábios(AR5), as áreas seccionais são aproximadas pela forma de área da elipse, com um eixo dado pela distância sagital e outro calculado de acordo com as características da região[1].

## III. Função de transferência

Considerando que o tracto vocal tem perdas, assumindo uma propagação de ondas planas ao longo deste, reduzindo o tracto vocal a um conjunto de K secções (cada uma de forma cilíndrica (Ver Fig. 4)), sendo a última secção terminada pela impedância de radiação calcula-se a função de transferência global do tracto vocal.

A função de transferência global H(w) é obtida calculando os ganhos de cada secção, desde a secção de radiação até à glótis (1),

$$H(w) = \frac{U_o}{U_g} = \prod_{i=1}^{K} \frac{U_{i-1}}{U_i} \qquad (1)$$

onde $U_o$ é o fluxo na radiação e $U_g$ é o fluxo glotal.

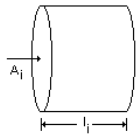


Fig. 4 - Secção *i* com área *Ai* e comprimento *li*.

Em cada secção, a função de transferência é o quociente entre o fluxo de entrada ( $U_{i-1}$ ) e o de saída ( $U_i$ ) (3),

$$\frac{U_{i-1}}{U_i} = \frac{1}{\cosh\Gamma_i + \frac{Z_B}{Z_i}.\sinh\Gamma_i} \qquad (3)$$

Onde $Z_i$ e $\Gamma_i$ são a impedância característica e a constante de transferência da secção, respectivamente. O cálculo de $Z_i$ e $\Gamma_i$ depende da área *Ai* e do comprimento *li* de cada secção [7] e, pressupõe-se perdas uniformes na superfície por viscosidade e condução de calor. $Z_B$ é a impedância de carga de cada secção calculada dos lábios para a glote e actualizada para trás por(4).

$$Z_B = \frac{P_i}{U_i} = Z_i.\frac{\cosh\Gamma_i - \frac{U_{i-1}}{U_i}}{\sinh\Gamma_i} \qquad (4)$$

Para calcular a impedância de radiação (impedância de carga da última secção) foi utilizado o modelo SKF[10].

## IV. Geradores de formantes

Os geradores de formantes foram implementados através de filtros em paralelo derivados a partir da expansão em fracções parciais da função de transferência. Tendo sido utilizado para o cálculo das singularidades da função de transferência um método de convergência mais rápida[3]. Este método procura as raízes do denominador e do numerador, em intervalos de 75Hz, da função de transferência sem perdas. Em seguida, as singularidades são interpoladas, utilizando o método de Newton, procurando em intervalos de 1Hz mas agora considerando a função de transferência com perdas. Este método converge rapidamente porque no caso sem perdas os cálculos usam apenas números reais e amostram a função de transferência com intervalos de 75Hz.

*B. Expansão em fracções parciais*

A função de transferência $H(s)$ pode ser expressa por um soma de fracções parciais, sabendo as raízes do denominador (pólos) e os correspondentes resíduos [3].

$$H(s) = \sum_{n=1}^{\infty} H_n(s)$$

onde,

$$H_n(s) = \frac{A_n}{s - s_n} + \frac{A_n^*}{s - s_n^*}$$

ou,

$$H_n(s) = \frac{2.\alpha_n.(s - \sigma_n) - 2.\beta_n.\omega_n}{s^2 - 2.\sigma_n.s + \omega_{on}^2} \qquad (5)$$

onde $^*$ representa o conjugado complexo, $s_n = \sigma_n + j\omega_n$ é o pólo e $A_n = \alpha_n + j\beta_n$ é o resíduo complexo em $s_n$.

O resíduo é primeiro calculado tendo apenas em consideração a parte imaginária do pólo, isto é, para $s_n = j\omega_n$ (6)

$$A_n = \frac{H_z(j\omega_n)}{\Delta H_p(j\omega_n)} \qquad (6)$$

Em seguida é feita uma correcção do valor do resíduo considerando agora a parte real do pólo. Assim cada singularidade ($s_k = \sigma_k + j\omega_k$) da função de transferência contribui para o factor de correcção do seguinte modo

$$C_{k,n} = \frac{s_n - s_k}{j\omega_n - s_k} = \frac{(\sigma_n - \sigma_k) + j(\omega_n - \omega_k)}{-\sigma_k + j(\omega_n - \omega_k)}$$

O resíduo correcto $\overline{A}_n$ depende das contribuições dos pólos e zeros,

$$\overline{A}_n = A_n \frac{C_{j,n}}{C_{i,n}} \qquad (7)$$

onde $C_{j,n}$ representa as contribuições dos pólos e $C_{i,n}$ as contribuições dos zeros.

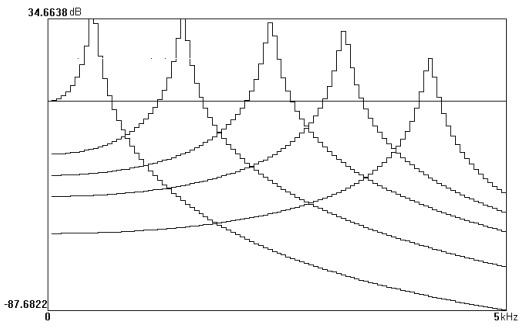


Fig.5 - Expansão em fracções parciais.

*B. Filtros digitais*

O equivalente discreto de cada $H_n(s)$ (5) é calculado considerando cada gerador de formante como a cascata de um filtro de primeira ordem de zeros e de segunda ordem de pólos.

O filtro discreto dos pólos é determinado com o método da resposta impulsional invariante. O filtro dos zeros é decomposto usando diferenças para trás e usando uma derivada discreta.

Os coeficientes do filtro são dados por,

$$b_1 = 2.E.\cos(\omega_n.T)$$

$$b_2 = E.E$$

$$a_1 = \frac{\alpha_n}{\alpha_n - (\sigma_n.\alpha_n + \beta_n.\omega_n).T}$$

$$g = \frac{2.\left[\alpha_n.(1/T - \sigma_n) - \beta_n.\omega_n\right].\left|1 - b_1/z_n + b_2/z_n^2\right|}{2.\sigma_n.\omega_n}$$

Onde T é o intervalo de amostragem

$$E = e^{\sigma_n.T}$$

$$z = e^{s_n.T}$$

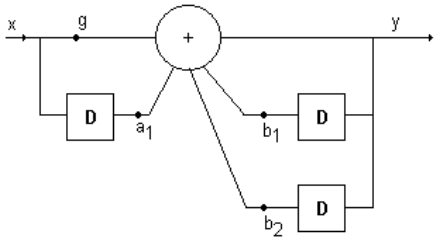


Fig. 6- Filtro digital de um formante.

Os filtros (Fig.6) podem ser implementados por uma equação de diferença,

$$y_n(i) = g.x(i) - a_1.x(i-1) + b_1.y_n(i-1) - b_2.y_n(i-2)$$

A saída global é dada por,

$$y(i) = \sum_{n=1}^{N} y_n(i)$$

## V. BREVE DESCRIÇÃO DO SINTETIZADOR

SINART é um simulador do modelo do tracto vocal implementado em C++. O simulador foi desenvolvido utilizando metodologias e metáforas de uma aplicação WINDOWS. Para além disso tendo como principal

objectivo que o simulador constituísse uma ferramenta de estudo tentou-se que toda a informação relativa ao modelo ficasse disponível ao utilizador. Assim, praticamente todas as etapas do processamento no modelo articulatório podem ser visualizadas: o traçado do tracto vocal, os pontos auxiliares para determinar o referido traçado, a forma dos lábios, as funções de distância e área sagital, a função de transferência com perdas e sem perdas, as singularidades da função de transferência, a função de transferência obtida a partir da expansão em fracções parciais, a excitação glotal e o som sintetizado. O simulador permite ainda obter uma representação gráfica das várias configurações articulatórias bem como a possibilidade de criar novas configurações alterando o valor dos parâmetros articulatórios armazenados em ficheiro. No entanto, sendo processo de representação interactivo também se pode proceder à alteração da forma do tracto vocal, seleccionando um articulador e deslocando-o, repercutindo-se a modificação para todas as janelas de visualização (Fig.4). Assim, analisando a informação fornecida pelo simulador consegue-se compreender a relação entre os parâmetros articulatórios e a correspondente descrição acústica (como por exemplo os valores dos formantes).

Para implementar o simulador foi usado o conceito de documento. Cada documento é responsável por coordenar os valores e visualizações de uma sequência de configurações articulatórias. O Documento cria, destroi e manda mensagens para as instâncias de visualização. Por exemplo se o utilizador altera os dados do documento as instâncias de visualização recebem uma mensagem para alterar os seus dados. Os documentos estão associados a um ficheiro permitindo guardar todos os valores correspondentes a uma dada sequência de configurações.

O documento contém uma sequência de configurações articulatórias, um conjunto de filtros de segunda ordem de pólos e de primeira de zeros em paralelo. Cada configuração articulatória contém os parâmetros articulatórios. Para cada configuração articulatória são ainda guardados os parâmetros referentes ao modelo de excitação glotal.

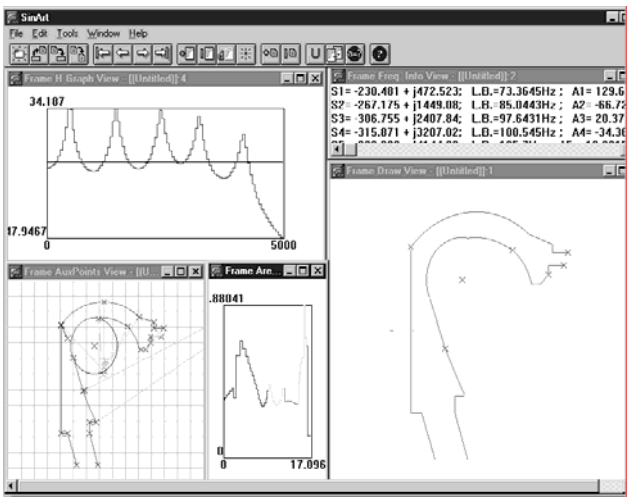


Fig. 5 - Imagem do sintetizador articulatório SINART.

## IV. RESULTADOS DE SIMULAÇÃO

### A - Síntese de Vogais Orais

Nos modelos articulatórios as cordas vocais, nos sons vozeados, são modeladas pela fonte de excitação glotal. Neste caso utilizou-se o polinómio de Rosenberg[4] que, embora não tenha interacção com o tracto vocal, é de complexidade computacional reduzida.

Usando descrições articulatórias e estudos acústicos das vogais Portuguesas [11] editamos as configurações articulatórias para cada vogal. A Fig.7 mostra o gráfico dos formantes F1/F2 que corresponde às configurações articulatórias das vogais orais sintetizadas. A localização no gráfico F1/F2 das vogais sintetizadas descreve um triângulo acústico semelhante aos obtidos em [11].

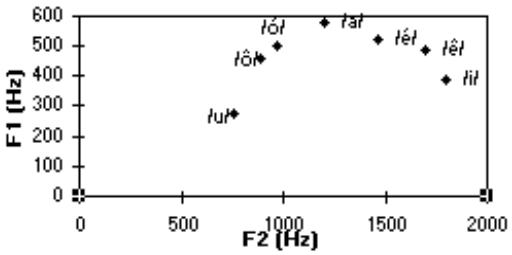


Fig. 7 - Gráfico dos formantes F1/F2 das vogais orais sintetizadas.

Na Fig. 5 estão representadas a configuração do tracto vocal e módulo da função de transferência para a síntese da vogal oral /ô/.

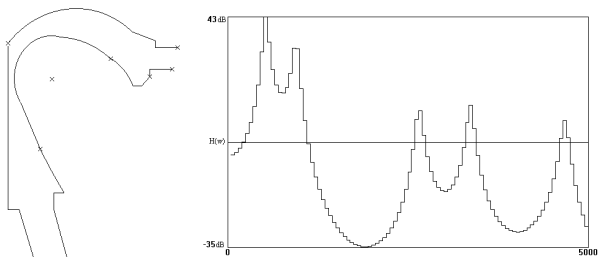


Fig. 6 - Exemplo da configuração do tracto vocal e função de transferência H(w) para a vogal oral /ô/. F1= 502 Hz, F2=932 Hz, F3= 2632 Hz.

### B - Determinação da forma do tracto vocal

Para fazer o mapeamento inverso neste trabalho foi utilizado um conjunto de redes neuronais de Kohonen [3,4,5,6]. Na primeira camada a rede mapeia parâmetros Cepstrais pesados derivados de LPC[9]. Cada neurónio na primeira camada contém uma sub-rede associada numa

segunda camada que mapeia o espaço articulatório e Cepstral.

A rede principal é uma rede de Kohonen com dimensão lateral de 6x6 com 18 parâmetros Cepstrais de entrada. Cada centroide Cepstral contém um sub-rede 2x2, com parâmetros Articulatórios e Cepstrais como entrada (ver Fig.7).

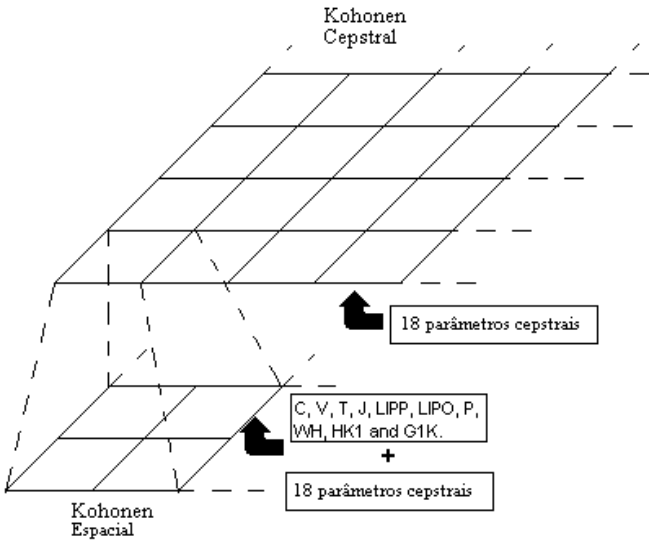


Fig.7 - Representação do conjunto de redes de Kohonen.

Os parâmetros articulatórios usados com entrada são:
- C, V, T, J, LIPP, LIPO, P, WH, HK1 and G1K.

A regra de treino das redes neuronais efectuado com base no trabalho de Kohonen[3, 4,5,6]. No entanto, tendo em atenção a topologia sugerida (ver Fig.7) o treino é efectuado com três etapas sucessivas. Em primeiro lugar é treinada a rede cepstral, em seguida o conjunto de treino é dividido em grupos. Cada grupo contém os padrões para os quais um neurónio cepstral foi vencedor. Assim, a rede espacial ligada ao neurónio cepstral é treinada com o respectivo grupo de padrões.

O mapeamento acústico-articulatório é feito determinando primeiro o vencedor da rede Cepstral . Por fim é calculado vencedor espacial da sub-rede espacial correspondente ao vencedor Cepstral.

E o conjunto de treino acordo com os exemplos descritos na literatura [17] e com interpolações entre eles obtivemos *Codebooks* de configurações das vogais portuguesas. Foram feitas algumas experiências de mapeamento inverso de vogais orais de um orador de 24 anos. Comparando os valores obtidos com os estudos anteriores o gráfico F1/F2 obteve uma configuração do triângulo acústica semelhante [26] [27].

A topologia das redes neuronais ajuda estudar o problema da não unicidade do mapeamento inverso porque junta na mesma sub-rede espacial configurações com as mesmas caracteríticas cepstrais mas pode conter diferenças espaciais. Inspecionando as configurações das sub-redes pode-se encontrar diferenças articulatórias entre as configurações assegurando semelhança cepstral (Fig.8).

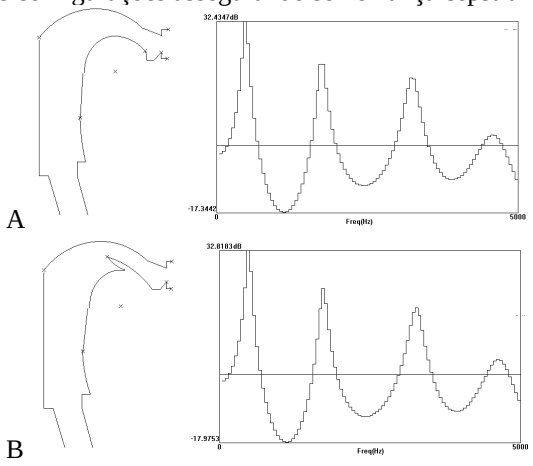


Fig.8 - Exemplo de configurações articulatórias diferentes mas com funções de transferência semelhantes.

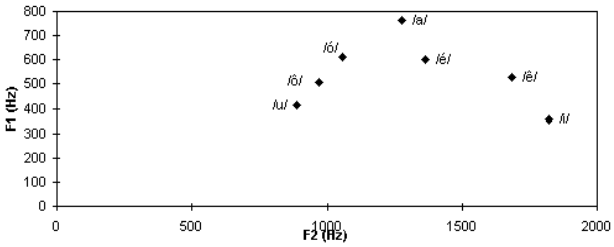


Fig. 9- Gráfico da localização F1/F2 das vogais orais.

C- Ajudas Visuais Para Terapia da Fala

A informação fornecida pelo método inverso permite obter um conjunto de dados articulatórios que não estão usualmente disponíveis nos métodos convencionais de terapia da fala. A construção de ajudas visuais para terapia da fala é usualmente realizada em forma de jogos que permitem ao indivíduo com problemas de fala praticar em casa sem a ajuda do terapeuta.

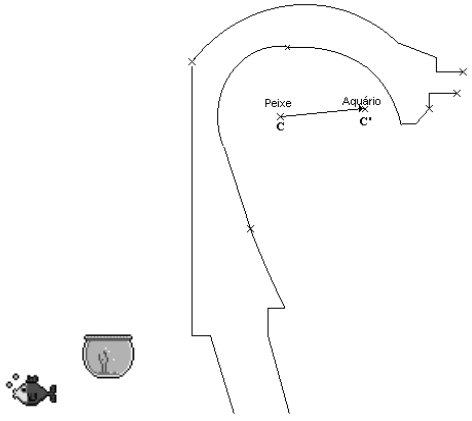


Fig 10- Ajuda visual em forma de jogo.

Fez-se uma implementação de um jogo protótipo em que o objectivo é colocar os articuladores numa dada posição (ver Fig.10).

O objectivo deste jogo é fazer com que o peixe entre dentro do aquário. Neste caso a posição do aquário é fixa e a posição do símbolo do peixe depende da posição do centro da língua C. Atinge-se o objectivo quando o peixe está dentro do aquário.

## VII. CONCLUSÕES

Neste artigo descreveu-se a implementação de um sintetizador articulatório utilizado para o estudo da produção da fala. Foi usado um método rápido e robusto para a determinação dos coeficientes dos filtros geradores de formantes.

O modelo pode ser estendido para produzir sons não vozeados, permitindo produzir os vários tipos de consoantes. Para trabalho futuro seria importante continuar o estudo de métodos inversos, de modo a obter mapeamentos mais robustos e rápidos. Será importante para a construção de ajudas visuais, no futuro, entrar em cooperação com terapeutas e médicos ligados ao ramo, para criar um sistema adaptado às necessidades reais da terapia da fala.

## AGRADECIMENTOS

Um dos autores (A.Branco) agradece ao programa PRAXIS XXI pelo suporte financeiro durante a realização deste trabalho. Esta investigação foi efectuada como satisfacção parcial dos requisitos do programa de Mestrado em Electrónica e Telecomunicações pela Universidade de Aveiro [27].

## REFERÊNCIAS

[1] Mermelstein, P. "Articulatory model for the study of speech production," J.Acoust. Soc. Am. Vol. 53, pp. 829-841, 1973.

[2] Prado, P. ,"A Target-based articulatory synthesizer," Phd thesis, University of Florida, 1991.

[3] Lin, Qiguang "A Fast Algorithm for Computing the Vocal-Tract Impulse Response from the Transfer Function," IEEE Transactions on Speech and Audio Processing, Vol.3, nº6, November 1995.

[4] Cummings ,"Glotal Models for Digital Speech Processing: A historical survey and new results," Digital Signal Processing, A Review Journal - Academic Press, Vol.5 , 1995.

[5] Coker, C. "A model of articulatory dynamics and control," Proc. IEEE 64, spp.452-460, 1976.

[6] Stevens, K., House, A. , "Development of a Quantitative Description of Vowel Articulation," J.Acoust. Soc. Am. Vol 27(3), pp.401-409, 1955.

[7] Sondhi, M., Schroeter, J., "A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer," IEEE Transactions on Acoustics, and Signal Processing, Vol. ASSP-35, nº7, pp.955-967, 1987.

[8] Rubin, P., Baer, T., Mermelstein, P. , "An articulatory synthesizer for perceptual research," J.Acoust. Soc. Am. Vol 70(2), pp. 321-328, 1981.

[9] Ishizaka, K., Flanagan, J. , "Synthesis of voiced sounds from two-mass model of the vocal cords," Bell Syst. Tech. J., vol. 51, nº 6, pp. 1233-1268, 1972.

[10] Badin, P., Fant, G. ,"Notes on vocal tract computation," STL-QPSR, nº2-3, pp.53-107, 1984.

[11] Martins,M. , "Análise acústica das vogais orais tónicas em Português" Instituto de Fonética da Faculdade de Letras de Lisboa, Novembro de 1971.

[12] Lin, Q. , "Vocal-Tract Computation: how to make it more Robust and Faster"STL-QPSR 4/1992.

[13] Atal, B., Chang, J., Mathews, M., Tukey, J.;"Inversion of articulatory-to-acoustic transformation in the vocal tract by computer sorting technique," J.Acoust. Soc. Am. Vol 63(5), pp.1535-1555, 1978.

[14] Juang, B., Rabiner, L., Wilpon, J, (1987) "On the use of band-pass liftering in speech recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-35, pp. 947-954.

[15] Kohonen, T. (1982), "Self Organized Formation of Topologically Correct Feature Maps," Biological Cybernetics, nº 43, pg. 55-69.

[16] Kohonen, T. (1982b), "Clustering Taxonomy, and Topological Maps of Patterns," Proceedings of the 6th international conference on pattern recognition, October.

[17] Kohonen, T. (1987), "Adaptive, associative, and self-organizing functions in neural computing," Applied Optics, Vol. 26, nº 23, 1 December.

[18] Kohonen, T. (1988), "The Neural phonetic typewriter," Computer, March.

[19] Rahim, M., Goodyear, C., et all (1993), "On the use of neural networks in articulatory speech synthesis,"J.Acoust. Soc. Am. Vol. 93(2), pp. 1109-1121.

[20] Schroeder, M.(1967), "Determination of the Geometry of the Human Vocal Tract by Acoustic Measurements," J.Acoust. Soc. Am. Vol 41(4), pp. 1002-1010.

[21] Schroeter,J., Sondhi, M. (1994),"Techniques for Estimating Vocal-Tract Shapes from the Speech Signal,"IEEE Transactions on Speech and Audio Processing, Vol.2,nº1,PART II, January 1994.

[22] Sondhi, M. (1974) "Model for wave propagation in a lossy vocal tract," J. Acoust. Soc. Am., Vol. 55(5), pp. 1070-1075.

[23] Sondhi, M. (1979), "Estimation of Vocal-Tract Areas: The Need for Acoustical Measurements," IEEE Transactions on Acoustics, and Signal Processing, Vol. ASSP-27, nº3, pp.268-273.

[24] Wakita, H. (1979), "Estimation of Vocal-Tract Shapes from Acoustical Analysis of the Speech Wave: The State of the Art," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-27, nº1, pp. 281-285.

[25] Borden, G., Harris,K., Raphael, L. (1980) "Speech Science Primer," William & Wilkins.

[26] A.Branco, A.Tomé, A.Teixeira, F.Vaz (1997), "A Method to Extract Articulatory Parameters from the Speech Signal using Neural Networks", DSP97, July 1997 / Greece.

[27] A.Branco (1997), "Representação Visual do Modelo Articulatório para io Estudo da Produção da Fala", Tese de Mestrado em Electrónica e Telecomunicações pela Universidade de Aveiro.