ANÁLISE E PARAMETRIZAÇÃO DE FONEMAS FRICATIVOS

ANTÔNIO MARCOS DE LIMA ARAÚJO[1], FÁBIO VIOLARO[2]

[1]CEFET/PA & UNICAMP–FEEC–DECOM
[2]UNICAMP–FEEC–DECOM
amarc@decom.fee.unicamp.br, fabio@decom.fee.unicamp.br

ABSTRACT

This paper presents an evaluation of fricative phoneme characteristics for the Portuguese Language, aiming at the development of algorithms to be used in computerized systems to aid the improvement of articulation control of deaf children. Many parameters for the fricative modeling are evaluated: zero-crossing rate, spectral gravity center and spectral peak frequency. We then propose an algorithm based on two new parameters: peak frequency of the smoothed spectrum and spectral spreading coefficient to determine if a given speech frame is fricative and what is its articulation zone. For the algorithm evaluation we used sustained speech signals produced by 8 children with ages ranging from 7 to 10 years. The proposed algorithm allowed the detection of fricatives with error rate below 0.48%, the classification of palatal fricatives with error rate below 14.3%, and the separation between alveolar and labiodental fricatives with error rate below 7.8%.

RESUMO

Este trabalho apresenta uma avaliação das características dos fonemas fricativos da língua portuguesa, visando o desenvolvimento de algoritmos que possibilitem a consecução de sistemas computadorizados para auxílio ao aprimoramento do controle articulatório de crianças com deficiência auditiva. São avaliados alguns dos parâmetros mais utilizados para modelar fricativas, como taxa de cruzamentos por zero, centro de gravidade espectral e freqüência de pico espectral. Ao final, é proposto um algoritmo baseado em dois novos parâmetros, a freqüência de pico do espectro suavizado (FPES) e o coeficiente de espalhamento espectral (CEE), para determinar se o quadro é (ou não) fricativo e qual sua zona de articulação. Na avaliação do algoritmo proposto foram utilizados sinais de voz gerados de forma sustentada por 8 crianças com idades variando entre 7 e 10 anos. O algoritmo proposto permitiu identificar presença de fricativas com taxa de erro inferior a 0,48%, classificar as palatais com taxa de erro inferior a 14,3% e separar as alveolares das labiodentais com taxa de erro inferior a 7,8%, independente da sonoridade.

1. INTRODUÇÃO

Consoantes fricativas são caracterizadas pela formação de uma constrição no trato vocal, produzindo aceleração e turbulência no fluxo de ar proveniente dos pulmões, resultando um ruído de fricção. O ruído primário excita os formantes associados à cavidade à frente da constrição, produzindo o som característico [1]. As consoantes fricativas são contínuas, podendo ser prolongadas, em princípio, tanto quanto permita a expiração.

As constritivas da língua portuguesa são classificadas, quanto à posição de articulação, em labiodentais ou anteriores (surda /f/ e sonora /v/, como em ca**f**é e nu**v**em), alveolares ou médias (surda /s/ e sonora /z/, como em **s**ono e ca**s**a), palatais ou posteriores (surda /ʃ/ e sonora /ʒ/, como em dei**x**ar e ho**j**e). Esses fonemas apresentam características espectrais relativamente fixas, embora sejam observadas variações que dependem dos locutores e/ou do contexto fonético [2]. Na língua portuguesa falada no Brasil, as fricativas correspondem a 12,1% dos sons [3].

As fricativas requerem para audição uma faixa de freqüências acima de 2.000 Hz, faixa esta para as quais crianças com deficiência auditiva usualmente apresentam reduzida sensibilidade auditiva [4]. Por isso as fricativas se situam no grupo de fonemas com aquisição mais difícil por parte dos deficientes auditivos [5].

O objetivo deste trabalho é a obtenção de um modelo paramétrico sobre os sinais de fala que possibilite a discriminação das (e entre) fricativas e que possibilite a sua utilização no desenvolvimento de sistemas computadorizados para auxílio ao aprendizado da fala de deficientes auditivos.

Um projeto neste sentido foi inicializado em 1994 em trabalho conjunto da Universidade Federal do Pará (UFPA) e do Centro Federal Tecnológico do Pará (CEFET-PA) [6]. Este projeto está atualmente em aprimoramento no Laboratório de Processamento Digital de Fala do Departamento de Comunicações da Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas (DECOM/FEEC/ UNICAMP).

2. CARACTERÍSTICAS DAS FRICATIVAS

O espectro do sinal das fricativas posteriores se estende de 3.000 a 6.000 Hz, o espectro das alveolares se estende de 4.500 a 10.000 Hz e o espectro das labiodentais se concentra entre 1.000 Hz e 10.000 Hz. As variantes sonoras apresentam um sinal laríngeo adicional abaixo de 1.000 Hz [7].

As labiodentais apresentam a intensidade mais reduzida de todas as consoantes do português [8]. Essas consoantes são tão débeis que alguns indivíduos praticamente não apresentam nenhuma área de intensidade visível na análise espectrográfica [9]. FANT [1] sugere que talvez devêssemos descrever esses sons como um fraco ruído de faixa larga, sem ressonâncias observáveis.

As Figuras 1, 2 e 3 apresentam espectrogramas de palavras com fricativas surdas anteriores, médias e posteriores, respectivamente.

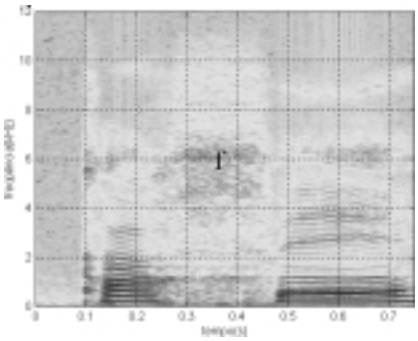

Figura 1. Espectrograma da palavra ca**f**é.

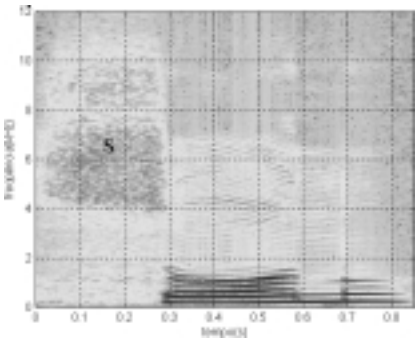

Figura 2. Espectrograma da palavra **s**ono.

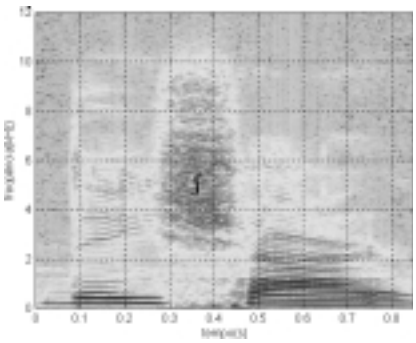

Figura 3. Espectrograma da palavra dei**x**ar.

A Figura 4 apresenta o espectrograma da palavra nuvem que contêm uma fricativa sonora anterior.

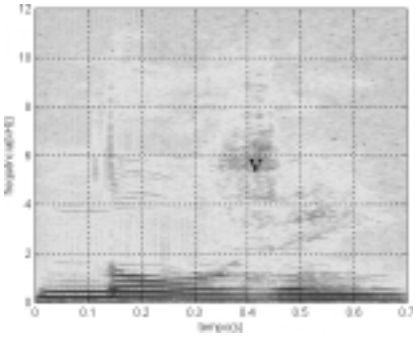

Figura 4. Espectrograma da palavra nu**v**em.

Nas figuras anteriores, pode-se observar que os fonemas fricativos apresentam componentes significativas de elevadas freqüências com duração superior a 50ms. As fricativas sonoras diferem das surdas pela presença (ou ausência) do traço de sonoridade nas freqüências até de 1.000 Hz, conforme se pode verificar pela comparação entre as Figuras 1 e 4.

As fricativas posteriores apresentam a energia por quadro do espectro de freqüência concentrada em uma faixa de freqüências mais baixa que as demais. Os espectros de freqüência das anteriores e médias, apresentam maior concentração de energia em freqüências mais elevadas que as posteriores. O espectro de freqüência das anteriores se espalha por uma faixa maior de freqüências.

A existência de componentes significativas nas altas freqüências, sugere a utilização da taxa de cruzamentos por zero (NCZ) para modelar e localizar quadros contendo fricativas, em especial as surdas [10, 11]. O centro de gravidade espectral (CG), definido como sendo a freqüência abaixo (ou acima) da qual estão concentradas 50% da energia do quadro, também pode ser utilizado para caracterização das fricativas. A freqüência associada ao valor máximo do espectro de potência do quadro (F_{pico}) é também sugerido para diferenciar alveolares das palatais [12].

As Figuras 5, 6, 7 e 8 mostram a trajetória obtida sobre as medidas, tomadas quadro a quadro, da taxa de cruzamentos por zero, do centro de gravidade e da freqüência de pico para as locuções cujos espectros são apresentados nas Figuras de 1 a 4, respectivamente.

A taxa de cruzamentos por zero foi normalizada, sendo apresentada em número de cruzamentos por zero por segundo (NCZ/s). Os sinais foram amostrados a 44.100 Hz e segmentados em quadros de 1024 amostras (23,2ms).

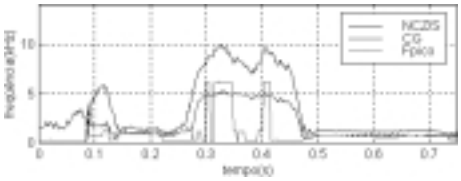

Figura 5. NCZ/s, CG e F_{pico} para a palavra ca**f**é.

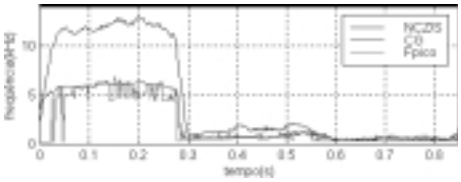

Figura 6. NCZ/s, CG e F_{pico} para a palavra **s**ono.

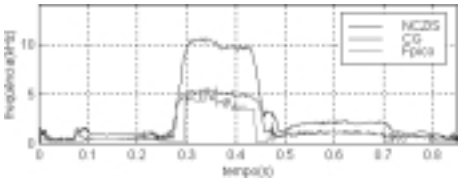

Figura 7. NCZ/s, CG e F_{pico} para a palavra dei**x**ar.

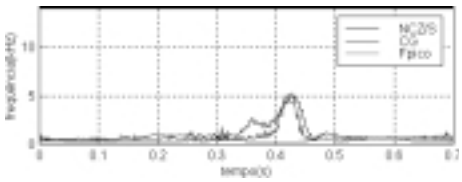

Figura 8. NCZ/s, CG e F_{pico} para a palavra nu**v**em.

Os resultados obtidos indicam que a taxa de cruzamentos por zero e o centro de gravidade apresentam menores variações intra fonema do que a freqüência de pico.

Nas locuções surdas, o NCZ/s mostra relação com a freqüência máxima do espectro com energia significativa. O centro de gravidade e a freqüência de pico mostram mais relação com as freqüência onde concentram-se as componentes mais significativas. Esses resultados são coerentes com os obtidos com sinais tonais puros, quando o NCZ/s deve ser igual a duas vezes a freqüência do tom, e o centro de gravidade e a freqüência de pico devem ser iguais à freqüência do tom.

Os parâmetros avaliados, em especial NCZ/s e CG, apresentam medidas maiores para quadros fricativos alveolares (médias) e menores para quadros fricativos palatais (posteriores).

As características de baixa intensidade e pouca resolução espectral das labiodentais, como caracterizado na literatura, produzem maior variabilidade das medidas sobre essas fricativas, como pode ser observado nas Figuras 5 e 8. Em especial, na labiodental sonora, a trajetória da F_{pico} não a diferencia dos demais fonemas e as trajetórias do NCZ/s e do CG se tornam mais definidas apenas no final do fonema.

Em geral, a sonoridade afeta todas as medidas. Para a mesma posição articulatória, variando-se a fonte (sonora/surda), o centro de gravidade é o parâmetro menos sensível.

3. FREQÜÊNCIA DE PICO DO ESPECTRO SUAVIZADO

As fricativas apresentam características espectrais definidas [1, 2, 7 e 9], portanto a freqüência de pico do espectro e/ou o centro de gravidade podem ajudar a discriminação das (e entre as) fricativas. Deve-se buscar parâmetros que reduzam a influência de picos locais do espectro e/ou da presença (ou ausência) da sonoridade.

Neste trabalho, propõe-se a utilização de uma medida da freqüência de pico do espectro suavizado (FPES), que consiste na medida da freqüência associada ao polo complexo conjugado resultante do modelamento AR de 3ª ordem do sinal.

Para sinais de voz adquiridos com altas freqüências de amostragem (ex. 44.100 Hz) as harmônicas da freqüência fundamental estão relativamente próximas da freqüência zero (abaixo de 1.000 Hz) e podem ser modeladas com a imposição de um polo real, possibilitando ao par de pólos complexos conjugados restante acompanhar a freqüência de pico do espectro suavizado.

As Figuras 9, 10, 11 e 12 apresentam para as palavras anteriores o espectrograma e a trajetória da medida da FPES. Os sinais foram divididos nos mesmos quadros utilizados para a obtenção de NCZ/s, CG e F_{pico}.

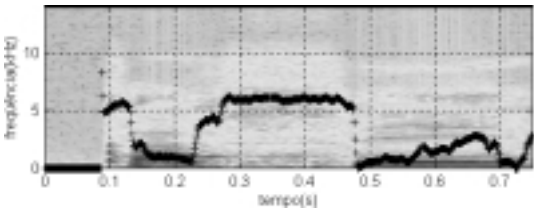

Figura 9. Espectrograma e FPES da palavra ca**f**é.

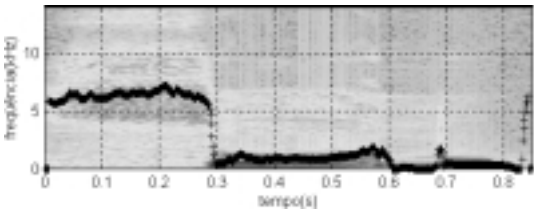

Figura 10. Espectrograma e FPES da palavra **s**ono.

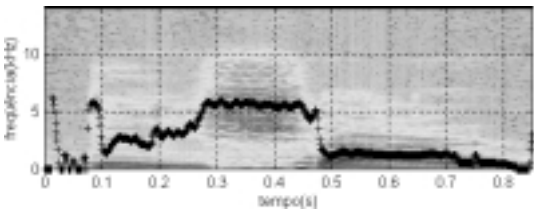

Figura 11. Espectrograma e FPES da palavra dei**x**ar.

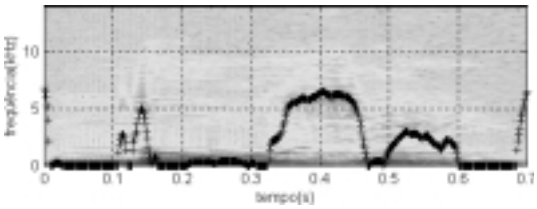

Figura 12. Espectrograma e FPES da palavra nu**v**em.

As figuras mostram que a medida da FPES se mantêm relativamente constante sobre todo o fonema, sendo maior para as labiodentais e alveolares e menor para as palatais. As Figuras 9 e 12, que contêm fonemas labiodentais, mostram que a medida da freqüência de pico do espectro suavizado está mais associada à posição articulatória e é relativamente insensível à sonoridade.

Analisando-se os resultados obtidos para a trajetória da medida da FPES sobre as palavras utilizadas, pode-se verificar que a sonoridade produz pouco efeito sobre a FPES, pois é modelada pelo polo real, permitindo ao par de pólos complexos conjugados restante modelar o filtro a partir da fonte de fricção.

A suavização realizada pelo modelo AR de 3ª ordem foi suficiente para reduzir efeitos de picos espúrios, sem eliminar a informação sobre a região de maior concentração da energia espectral.

A medida da FPES sobre as oclusivas (não consideradas neste trabalho) pode atingir valores elevados nos momentos da oclusão

com curta duração. Este efeito pode ser visto comparando-se as trajetórias da FPES para os fonemas oclusivos /k/ e /d/ nas palavras **c**afé e **d**eixar ilustradas nas Figuras 9 e 11. Nas fricativas, por outro lado, a medida da FPES permanece com valores elevados por mais de 50ms.

A Figura 13 apresenta uma aproximação das funções densidade de probabilidade (fdp) das medidas da FPES de fricativas, computadas a partir dos respectivos histogramas. As medidas foram obtidas da fonação sustentada de 8 crianças com idade entre 7 e 10 anos, 4 do sexo masculino e 4 do sexo feminino, em 40 quadros de 46ms (2048 amostras), totalizando 320 quadros para cada fonema.

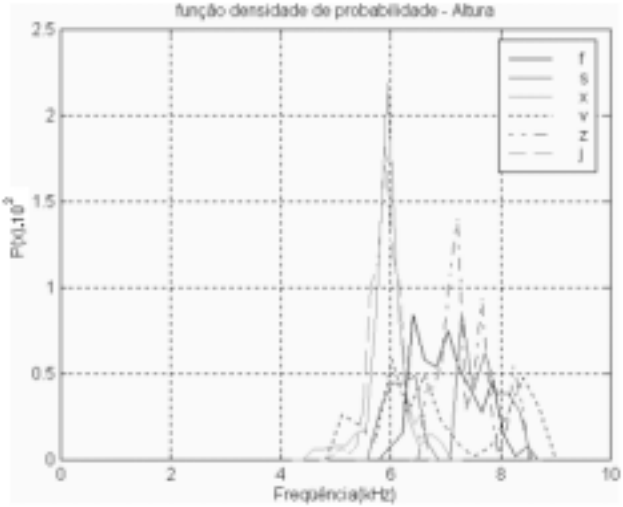

Figura 13. Função densidade de probabilidade das medidas da FPES de fricativas sustentadas.

As medidas da FPES das fricativas estão entre 4.500 e 8.850 Hz, sendo que 99,3% das medidas estão acima de 5.000Hz. As fricativas palatais (mais graves) apresentam menores medidas da FPES (média de 5.925Hz) e as fricativas anteriores e médias apresentam valores médios da FPES mais elevados (6.957 Hz e 7.246 Hz, respectivamente).

Os valores de pico das fdp's, assim como os valores médios da medida da FPES das anteriores são menores que os correspondentes das mediais, resultados concordantes com os apresentados na literatura [7, 8, 9, 12].

A medida da FPES apresentou valores médios de 6.739 Hz para as surdas e 6.788 Hz para as sonoras, variação inferior a 0,73%.

As medidas da FPES dos sons das palatais são menores e com menor desvio padrão que as demais, apresentando-se bastante concentradas pouco abaixo de 6.000 Hz, permitindo assim realizar a sua classificação. A regra resultante para classificação de uma fricativa palatal é a obtenção da FPES entre 5.000 e 6.270Hz. O limiar superior foi obtido por contagem, minimizando a probabilidade de erro. Para medidas da FPES nessa faixa de freqüências as palatais são classificadas com taxa de erro de 14,3%. A decisão sonora/surda deve ser realizada de maneira independente.

Medidas da FPES superiores a 6.270 Hz são características de fonemas fricativos labiodentais ou alveolares.

4. COEFICIENTE DE ESPALHAMENTO ESPECTRAL

A medida da freqüência de pico do espectro suavizado é insuficiente para realizar uma decisão entre fricativas labiodentais e alveolares. A literatura [1, 7, 8, 9, 12] e a observação dos registros espectrográficos de palavras contendo fricativas labiodentais e alveolares mostram que as alveolares apresentam maior concentração do espectro, enquanto as labiodentais apresentam espectros mais difusos. Nesse caso, o parâmetro proposto para a classificação entre labiodentais e alveolares é o coeficiente de espalhamento espectral (CEE), dado por :

$$CEE = \frac{F_{80} - F_{20}}{F_{20}} \quad \textbf{(1)}$$

onde Fx é a freqüência abaixo da qual está concentrada x % da energia do quadro do sinal filtrado acima de 2.000 Hz. Essa filtragem passa altas em 2.000 Hz visa extrair as componentes relativas ao traço de sonoridade. Neste trabalho esta filtragem foi feita no domínio da freqüência, eliminando-se as componentes da DFT abaixo de 2.000 Hz.

A Figura 14 apresenta uma aproximação das funções densidade de probabilidade das medidas de coeficiente de espalhamento espectral para as fricativas médias e anteriores.

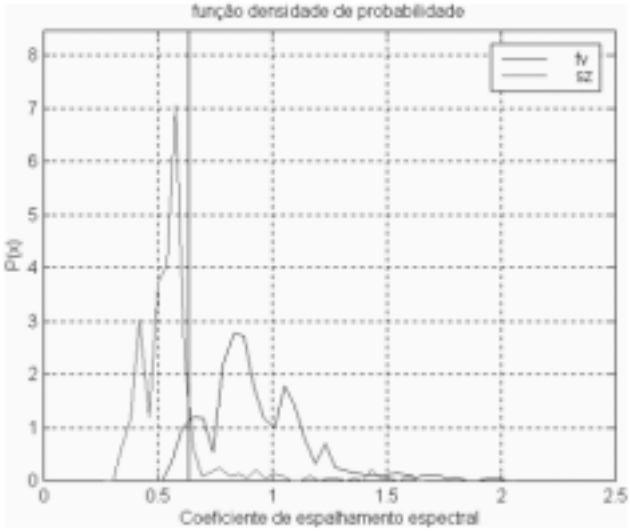

Figura 14. Função densidade de probabilidade das medidas do coeficiente de espalhamento espectral das fricativas médias e anteriores.

A distribuição das fdp's mostra que o espalhamento espectral das médias é menor que o espalhamento espectral das anteriores. A Figura 14 apresenta, ainda, o limiar de decisão obtido minimizando a probabilidade de erro por contagem. O limiar obtido sobre o coeficiente de espalhamento espectral para decisão entre anteriores e médias foi de 0,64, resultando em 7,8% de erros.

5. RESULTADOS

As principais medidas usualmente utilizadas para caracterização de fricativas (NCZ/s e CG) são fortemente afetadas pela presença (ou ausência) de sonoridade. Essas medidas podem ficar mais imunes à sonoridade eliminando-se por filtragem passa altas as componentes abaixo de 2.000 Hz, mas uma filtragem pode fazer com que medidas sobre outros fones, como as vogais /i, ê/, se confundam com as fricativas. Além disso, as características de baixa intensidade e pouca resolução espectral das labiodentais produzem muita variabilidade das medidas do número de cruzamentos por zero, da freqüência do pico do espectro e do centro de gravidade sobre essas fricativas, em especial no caso das sonoras.

Para sinais adquiridos com freqüências de amostragem altas, a utilização de um modelamento AR de 3ª ordem permitiu que um polo real modelasse a sonoridade e possibilitou que os dois pólos complexos conjugados restantes modelassem a freqüência de pico do espectro suavizado (FPES). O modelo foi suficiente para reduzir efeitos de picos espúrios, mantendo as informações sobre a região de maior concentração do espectro de energia do quadro considerado.

A avaliação de sinais fricativos com crianças entre 7 e 10 anos permitiu verificar que todas as fricativas apresentam freqüência de pico do espectro suavizado superior a 5.000 Hz. Para crianças na faixa etária indicada e para fonação sustentada, a altura média obtida foi de 6.957 Hz para labiodentais, 7.254 Hz para alveolares e de 5.609 Hz para palatais.

A medida da freqüência de pico do espectro suavizado média é de 6.486 Hz sobre quadros surdos e de 6.517 Hz sobre os quadros sonoros, que corresponde a uma variação inferior o 0,73%, indicando este parâmetro é relativamente insensível à sonoridade, como proposto.

Medidas da freqüência de pico do espectro suavizado acima de 5.000 Hz são característicos de fricativas.

A classificação como palatal pode ser realizada com taxa de erros de 14,3% quando a medida da freqüência de pico do espectro suavizado está entre 5.000 Hz e 6.270 Hz.

Medidas da freqüência de pico do espectro suavizado superiores a 6.270 Hz são características de fonemas fricativos labiodentais ou alveolares.

6. CONCLUSÕES

A freqüência de pico do espectro suavizado é bastante independente da sonoridade, permite caracterizar uma dada produção sonora como sendo (ou não) uma fricativa e ainda possibilita caracterizar as palatais.

O coeficiente de espalhamento espectral é também bastante independente da sonoridade. O seu valor quando obtido sobre quadros de sinais labiodentais é maior que o obtido sobre quadros de sinais alveolares, permitindo realizar a classificação entre estas duas zonas de articulação. Medidas do coeficiente de espalhamento espectral acima de 0,64 são característicos de fricativas labiodentais enquanto medidas inferiores a 0,64 são característicos de alveolares, resultando esse critério em 7,8% de erros. A decisão sonora/surda deve ser realizada à parte.

Os limiares foram obtidos para crianças entre 7 e 10 anos em fonação sustentada, sendo portanto válidos somente para esta faixa etária e na condição especificada. Para outras idades e para fala fluente os limiares deverão ser alterados.

Estes parâmetros foram incorporados a jogos computacionais para aprimoramento da locução de sons fricativos em crianças com deficiência auditiva.

7. REFERÊNCIAS

[1] Fant G. *Acoustic theory of speech productions*. The Haguc, Mouton, 1960.

[2] Klatt D.H. "Review of text-to-speech conversion for English". *Journal Acoustical Society of America*, 82(3):737-793, 1987.

[3] Alcaim A., Solewicz J.A e Moraes J.A. "Freqüência de ocorrência dos fones e listas de frases foneticamente balanceadas no português falado no Rio de Janeiro". *Revista da Sociedade Brasileira de Telecomunicações*, 7(1):23-42, 1992.

[4] Ling D. "Amplification to speech". In Calvert D. R. and Silverman S. R. *Speech and deafness*, 2nd ed., Alexander Graham Bell Association for the Deaf, Washington, 1978.

[5] Calvert D. R. and Silverman S. R. *Speech and deafness*. 2nd ed., Alexander Graham Bell Association for the Deaf, Washington, 1978.

[6] Castro, D.B., Anjos, J.C.A., Klautau, A. e Araújo, A.M.L. "Sistema áudio-gráfico visual computadorizado para auxílio no treinamento de deficientes auditivos". *Congresso Nacional de Matemática Aplicada e Computacional*, Curitiba, Setembro de 1995, pag. 725-729.

[7] Launay C. and Borel-Maisonny S. *Distúrbios da linguagem da fala e da voz na infância*. Roca, São Paulo, 1989.

[8] Russo I. e Behlau M. *Percepção da fala: análise acústica do português brasileiro*. Lovise, São Paulo, 1993.

[9] Santos M.T. "Uma análise espectrográfica dos sons fricativos surdos e sonoros do português brasileiro". *Monografia de Especialização*, Escola Paulista de Medicina, São Paulo, 1987

[10] Rabiner L.R. and Schafer R.W. *Digital Processing of Speech Signals*. Prentice-Hall, New Jersey, 1978.

[11] Vieira, M.N. "Módulo frontal para um sistema de reconhecimento automático de voz". *Dissertação de Mestrado*, UNICAMP, Campinas, 1989.

[12] Kent R.D. and Read C. *The acoustic analysis of speech*. Singular Publishing, San Diego, 1992.

AGRADECIMENTOS

Este trabalho foi parcialmente financiado pelo CNPq através da bolsa de Doutorado, processo 144383/1996-9.