# Estimativa do comportamento vocálico de locutores

**E.L. Fernandes da Silva, H. M. de Oliveira**

Universidade Federal de Pernambuco - UFPE
Centro de Tecnologia e Geociencias, Recife.
E-mail: lizandra_fernandes@hotmail.com hmo@ufpe.br

***Resumo**: Este artigo propõe um método simples para a estimação automática do comportamento espectral de trechos vocálicos de locutores. Uma implementação computacional em Matlab® é apresentada e a validação é conduzida comparando os resultados com uma identificação realizada com intervenção humana empregando o Audacity®. Locutores (masculinos e femininos) foram considerados e os testes foram conduzidos para sete diferentes sons vocálicos da língua portuguesa (a, é, ê, i, ó, ô, u). A abordagem proposta pode ser útil em modelos de trato vocal, na melhoria da qualidade de sintetizadores de voz ou em algoritmos de reconhecimento automático de locutor.*

Os sons relacionados à fala podem ser classificados como vocálicos e não vocálicos [1, 2]. Os vocálicos são obtidos quando o ar que vem dos pulmões passa pela glote e não sofre interrupção parcial ou total por: língua, lábios, dentes, etc., provocando vibrações quase periódicas. Já os sons não-vocálicos são obtidos pela interrupção parcial ou total do ar, durante o percurso dos pulmões até sua saída pelas cavidades oral e nasal [1]. A maior parte dos sons emitidos na língua portuguesa é vocálico e, portanto, de comportamento espectral caracterizado por frequências mais bem definidas (picos/raias espectrais). Este artigo propõe investigar o comportamento espectral de sons vocálicos na língua portuguesa. O procedimento foi implementado em Matlab® e aplicado a seis locutores distintos, sendo os resultados do comportamento espectral apresentados. Compara-se também os resultados com estimativas de *pitch* obtidas para os sinais vocálicos, obtidas por um método clássico [9]. Os dados foram coletados em uma sala fechada e com baixo nível de ruído. A coleta foi realizada com o auxílio de um laptop e microfone acoplado a um fone de ouvido. Foram realizadas sete coletas vocálicas para cada um entre seis locutores diferentes, divididos igualmente entre os sexos masculino e feminino [3, 4]. A idade dos colaboradores variou entre 22 e 35 anos. A taxa de amostragem de frequência foi de 44,1 kHz, a mesma utilizada em sistemas de áudio padrão CD [4]. O programa utilizado para a captura da voz dos participantes foi o Audacity 1.3®. O processo consistiu em realizar coletas de sons das vogais, do alfabeto brasileiro, incluindo as variações regionais para as letras "e" e "o", por intervalos de tempo de aproximadamente 7 segundos sem interrupções. Diferentemente das técnicas usuais para codificadores de voz LPC (*Linear Predictive Coding*), em que se requer a estimativa de *pitch* em uma janela curta (tipicamente 30 ms, trecho quase-estacionário de voz, [5, 6]), o procedimento descrito aqui trabalha com uma repetição bastante longa dos sons vocálicos, de forma a melhor estimar as características do locutor. Modela-se *offline* o comportamento do trato vocal em regiões vocálicas, como um processo de "aprendizado". Pela teoria clássica de Fourier [8], sons periódicos (tais como os sons vocálicos) apresentam essencialmente espectro discreto, caracterizado por picos (raias espectrais). A ilustração do arquivo de áudio para a vogal "a", obtida com o auxílio do mesmo programa pode ser vista na Fig.1. O pré-processamento também foi realizado com o auxilio do Audacity®, para eliminar os trechos de silêncio no início e no final de cada arquivo (empregando algoritmo VAD [7]) e para retirar trechos ruidosos ocasionados, no início e fim, pelo *click* do mouse. A ilustração de um sinal pré-processado, com ausência de silêncio e sem trechos ruidosos no início pode ser vista na Fig. 1. As gravações geraram arquivos na extensão .wav. Este procedimento foi conduzido para cada um dos locutores, para cada uma das vogais, incluindo acentuação (grave/agudo para e, o).

DOI: 10.5540/03.2015.003.01.0469  

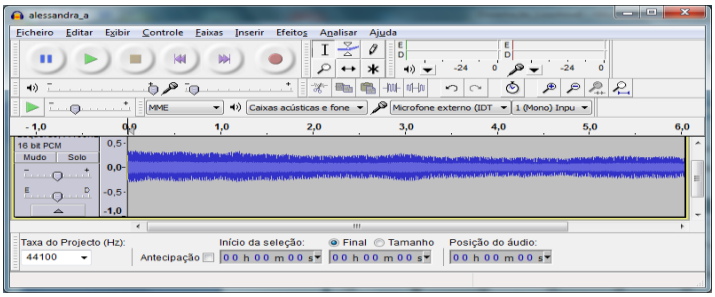


Figura 1: Ilustração do sinal pré-processado correspondente a vogal "a" sendo repetida por 7 segundos pelo locutor Alessandra.

A leitura do arquivo é realizada no início do algoritmo e os sons dos arquivos são amostrados a uma taxa de 44,1 kHz e quantizados em 16 bits. O comprimento da janela utilizado no identificador de *pitch* foi estabelecido com base em observações realizadas no Audacity®. A janela tomada, inicialmente, foi de 128 amostras (comprimento da transformada com a taxa de amostragem fixada), com janelamento de Hamming [2, 6]. Os espectros foram calculados expressando-os em escalas de frequência linear e logarítmica. A escala de frequência linear não apresenta uma visão tão ampla ao longo de seu eixo como a escala de frequência logarítmica, que foi, por esta razão, a escala adotada. Nota-se que para o comprimento $N$=128 não se constata o aparecimento de picos (como seria de se esperar, uma vez que o som registrado é quase periódico). Isto significa que esta resolução espectral adotada ainda é insuficiente. Como forma de melhorar a resolução espectral, optou-se por aumentar o número de amostras por janela. Como pode ser visto na Fig. 2, para a janela de 512 amostras, a distância entre as amostras não sofre variação em relação ao tempo, pois a taxa de amostragem foi mantida constante em 44,1 kamostras/s. No entanto, quando essas mesmas amostras são convertidas para o domínio da frequência, avaliando o conteúdo harmônico nas frequências $n/(NT_s)$, $n$=0,1,...,$N$-1, sendo $N$ o comprimento da DFT. A distância entre as raias espectrais varia com relação ao número de amostras [8] e vale:

$$\Delta f = 1/(NT_s), \tag{1}$$

em que, $\Delta f$ é o valor é a resolução das raias espectrais em Hz/amostras, e $T_s$ é o período de amostragem. A Tabela 1 ilustra a variação do comprimento da transformada discreta de Fourier (DFT) de uma determinada janela com relação à resolução. Como pode ser visto nesta tabela, para o comprimento de janela de 2048 amostras, a resolução é de 21,5 Hz. Essa resolução é satisfatória e por isso é a adotada neste trabalho.

| Comprimento da *DFT* | Resolução (Hz) |
|---|---|
| 128 | 344 |
| 256 | 172 |
| 512 | 86 |
| 1024 | 43 |
| 2048 | 21,5 |

Tabela 1. DFT com Relação à Respectiva Resolução.

A fim de se obter uma melhor visualização dos picos de frequência, variou-se o tamanho da janela utilizada para amostragem. As Figs. 2(a,b,c) ilustram o espectro obtido com tamanho de janelas 512, 1024 e 2048, respectivamente, (todos em escala de frequência logarítmica). Analisando as figuras, observa-se que à medida que o tamanho da janela aumenta, o número de amostras também aumenta, e mais picos de frequência são visualizados. Para tamanho de janelas maiores que 2048, não se notou melhora visual, e este valor foi adotado. A duração do trecho de áudio selecionado (cerca de 7s) fornece cerca de 150 janelas e a média aritmética dos espectros de todas as janelas é tomada como representativa do espectro.

DOI: 10.5540/03.2015.003.01.0469  

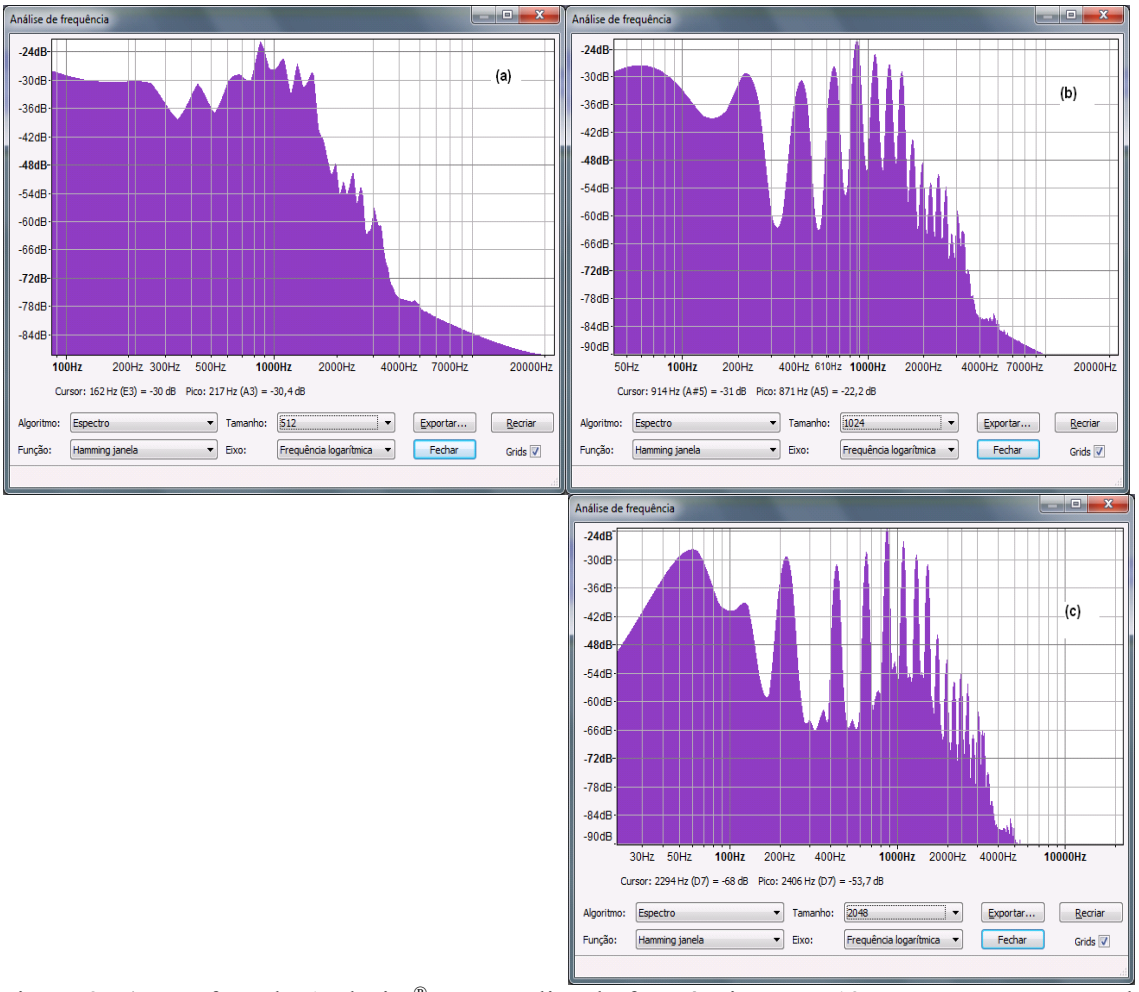


Figura 2: a) Interface do Audacity® para análise de frequência com 512 amostras para a vogal "*a*", sendo repetida por 7 segundos pelo locutor Alessandra. O aumento do comprimento da DFT conduz a explicitar os picos relacionados com os sons harmônicos. b) Interface do Audacity® para análise de frequência com 1024 amostras para a vogal a, sendo repetida por 7 segundos pelo locutor Alessandra. c) Interface do Audacity® para análise de frequência com 2048 amostras para a vogal "*a*", sendo repetida por 7 segundos pelo locutor Alessandra.Verifica-se uma estabilização no formato espectral.

A próxima etapa do identificador espectral é a de tornar a cardinalidade do "conjunto de amostras" um múltiplo de 2048. Para tornar o conjunto de amostras um múltiplo deste comprimento foi realizado um pequeno ajuste, que consiste em dividir o número total de amostras por 2048 e usar o teto deste valor. Em seguida, esse número de janelas é utilizado no cálculo do comprimento do vetor nulo (*zero padding*), que realiza a complementação exata de elementos necessários no vetor de amostras totais, para que esse vetor se torne um múltiplo de 2048. Após o cálculo da FFT, uma normalização da janela final, obtida pelo somatório acumulativo é realizada. Esse cálculo foi necessário para o estabelecimento de critérios de verificação estatísticos, utilizados na localização dos elementos de *pitch*/formantes em etapas posteriores. O cálculo da normalização consiste em dividir o quadrado dos elementos de amplitude espectrais da janela final, obtida depois do cálculo da FFT, pela energia dessa janela. A questão é que há níveis de áudio distintos de gravação e trechos com energia diferente. Após os valores nesta janela serem normalizados realiza-se uma busca, dentro do vetor, pelo máximo valor de amplitude. Esse valor será utilizado em etapas subsequentes.

DOI: 10.5540/03.2015.003.01.0469  

O identificador espectral proposto busca realizar a extração dos elementos do sinal de voz de maneira simplificada. Para isso, utiliza-se um critério de análise, no qual cada amostra obtida até a etapa de normalização deve ser comparada uma a uma. O critério se baseia na comparação entre o valor da amostra atual e o valor da amostra posterior. A estimativa de picos é naturalmente feita no domínio frequencial (isto foi decisivo para adotar taxa 44,1 kHz), obtendo-se a DFT $v_k \leftrightarrow V_n$, examinando-se o comportamento de $|V_n|$, $n$=0,1,...,$N$/2. Se o valor da amostra atual for menor do que o valor da amostra posterior adjacente e se o valor absoluto dessa amostra for maior do que 10% do máximo valor obtido na etapa de normalização, será referenciado o valor 1 a essa amostra, caso contrário, se o valor da amostra analisada for maior do que o valor da amostra posterior e se o valor dessa amostra for superior a 10% do máximo valor obtido na etapa de normalização, será referenciado o valor -1 para essa amostra. Este procedimento é computacionalmente atrativo, pois as amostras espectrais são "reduzidas" (i.e., mapeadas) em ±1, bem mais simples que adotar a magnitude das raias. Ao término desse processo, tem-se um vetor com o mesmo comprimento de 2048 amostras, porém, agora esse vetor só possui dois valores, -1 ou 1. O processo corresponde a uma indicação do sinal (função sgn) da derivada do espectro:

+1 indica áudio em região crescente
-1 indica áudio em região decrescente

Um critério de 10% de tolerância foi adotado para evitar que flutuações, ruídos e/ou erros interfiram no processo. A existência de picos espectrais está associada a pontos de máximo, logo de derivada nula. Se há inversão no sinal da derivada, segue-se que o espectro sai de uma região crescente para decrescente. Finalmente, sobre o vetor obtido, com valores -1 ou 1, são extraídos os picos de frequência. O pico é localizado na região de transição de +1 para -1. Por exemplo, em uma sequencia In expressa por

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $I_n$ | +1 | +1 | +1 | -1 | -1 | -1 | +1 | +1 | ... |
| $n$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | ... |

um pico é localizado na raia 3 (transição +1 $\rightarrow$ -1).

A coleta é realizada apenas para os $N$*=150 primeiros elementos, pois os espectros típicos de voz encontram-se "confinados" nesse intervalo. Se for efetuada a divisão da frequência de amostragem do sinal (44100 Hz) pelo número de amostras da janela (2048), chega-se a um resultado em que todas as amostras encontram-se espaçadas por um intervalo de 21,53 segundos, portanto, são necessárias apenas 150 amostras para cobrir a parte de maior conteúdo de energia do espectro de áudio – desnecessário, pois, avaliar raias correspondentes a frequências mais elevadas. O identificador proposto lida tão somente com sgn(.) da grandeza $\Delta V_n=|V_{n+1}|-|V_n|$, $n$=0,1...,$N$*<<$N$/2. Para finalizar o processo, realiza-se a contagem dos picos de frequência. Essa contagem se dá da seguinte forma: se o elemento correspondente à amostra analisada tiver o valor numérico 1 e se a amostra subsequente possuir o valor -1, cria-se um novo vetor no qual a posição dessa amostra receberá o valor 1, que indica que nessa determinada posição existe um pico; caso contrário, o valor numérico atribuído será 0 e isso indicará a ausência de pico espectral nessa posição. É importante salientar que, ao se realizar a leitura do vetor final, que contém a posição dos picos de frequência, excluíram-se os dois primeiros elementos, já que nessa faixa de frequência tem-se a faixa da rede local de energia elétrica. Apenas picos com frequência acima de 200 Hz foram considerados como relevantes.

Os experimentos foram realizados para 42 arquivos de áudio, pois cada locutor realizou sete locuções, relativas a cada vogal do alfabeto brasileiro mais as duas variações tônicas (ô e ê). Com o auxilio do aplicativo Audacity® foram obtidas também, as raias espectrais relevantes para os mesmos 42 arquivos de áudio citados. O objetivo dessa análise foi o de comparar as posições dos picos de frequência obtidos pelo identificador espectral e pelo Audacity®.

DOI: 10.5540/03.2015.003.01.0469  

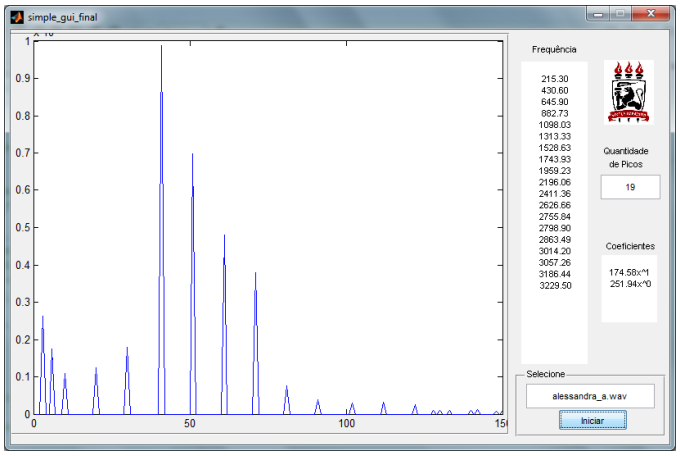


Figura 3: Ilustração da Interface gráfica em Matlab®, do programa identificador de *Pitch* para a vogal "*a*" (Alessandra). Seleciona-se o arquivo extensão .wav (canto direito da tela): 19 picos são mostrados para esta vogal.

Observa-se que as amostras coletadas foram obtidas por método de coleta manual, ou seja, as posições de cada pico de frequência eram observadas na janela da interface do aplicativo, contadas uma a uma. Já no identificador espectral, a contagem é realizada de maneira automática pelo próprio algoritmo.

| Picos | Audacity® (Hz) | Identificador *picos espectrais* (Hz) |
|---|---|---|
| 1º | 59,00 | - |
| 2º | 120,00 | - |
| 3º | 213,00 | 215,30 |
| 4º | 431,00 | 430,60 |
| 5º | 653,00 | 645,90 |
| 6º | 872,00 | 882,73 |
| 7º | 1088,00 | 1098,03 |
| 8º | 1301,00 | 1313,33 |
| 9º | 1518,00 | 1528,63 |
| 10º | 1731,00 | 1743,93 |
| 11º | 1970,00 | 1959,23 |
| 12º | 2156,00 | 2196,06 |
| 13º | 2371,00 | - |
| 14º | - | 2411,36 |
| 15º | 2544,00 | - |
| 16º | - | 2626,66 |
| 17º | - | 2755,84 |
| 18º | - | 2798,90 |
| 19º | 2855,00 | 2863,49 |
| 20º | 3008,00 | 3014,20 |
| 21º | - | 3057,26 |
| 22º | - | 3186,44 |
| 23º | 3280,00 | 3229,50 |

Tabela 2- Picos de frequências, em Hz, para os 19 picos significativos, obtidos pelo Audacity® e pelo identificador para a vogal "*a*" pronunciada por um locutor do sexo feminino.

DOI: 10.5540/03.2015.003.01.0469

A ilustração da interface do identificador espectral com a extração de parâmetros da vogal "*a*" pode ser vista na Fig. 3. Os resultados obtidos pelo identificador de espectro vocálico e pelo Audacity® podem ser vistos na Tabela 2 para o experimento em que a vogal "*a*" é pronunciada por um locutor do sexo feminino. A concordância dos resultados obtidos de modo não automático com o Audacity®, selecionando visualmente os picos, e com o Matlab® desenvolvido, é notável, com $r^2$=0,9997 e EMQ=30 Hz. Para verificar a aderência das raias espectrais identificadas pelo aplicativo e um espectro periódico (fundamental & harmônicos), realizou-se um ajuste linear para cada locutor. Mostra-se (Fig. 4), a título ilustrativo, o ajuste obtido para locutores selecionados. O mesmo procedimento foi repetido para cada uma das "vogais" (a, é, ê, i, ó, ô, u) e para cada um dos locutores testados. A estimativa do passo foi estabelecida usando regressão linear (mínimos quadrados) entre a ordem do pico e sua frequência. Para o exemplo descrito (locutor Alessandra, vogal "*a*"), obteve-se a curva com "coeficiente" inicial 215,3 Hz (apêndice na url http://www2.ee.ufpe.br/codec/vocalico.html) e passo médio em harmônicos 174,6 Hz. Como visto, na determinação de picos, os dois primeiros valores dos elementos de frequência foram retirados da análise do algoritmo de identificador de raias espectrais relevantes, para que perturbações da rede não ocasionassem variações significativas nos valores das amostras coletados.

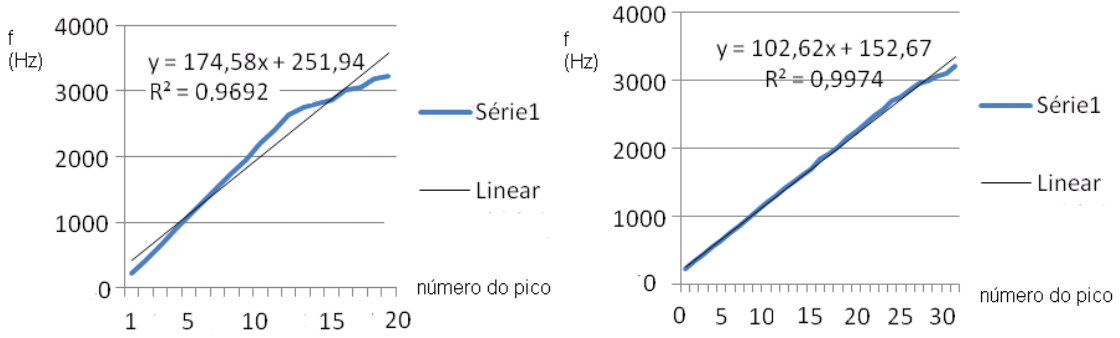


Figura 4: Correlação entre picos identificados pelo aplicativo (ajuste linear) para: a) locutor Alessandra, pronunciando longamente o som vocálico "*a*". b) locutor Ricardo, pronunciando longamente o som vocálico "*a*". Equação de regressão e coeficiente de determinação indicados. Ajustes de regressão linear com *n* pontos, 9<*n*<37, dependendo do som.

Na coleta das amostras do Audacity® foram considerados todos os valores de frequência. A título comparativo, também se empregou um algoritmo de detecção de *pitch* clássico [9] para avalia-lo em cada fonema vocálico repetido, para cada dos locutores. O algoritmo de detecção de *pitch* usado, com base na relação sub-harmônica-harmônica, encontra-se disponibilizado em Matlab na URL http://www.speakingx.com/blog/2008/01/02/pitch-determination. Há uma boa concordância (erro inferior a 5%) entre os valores estimados de *pitch* com o método de estimativa de raias vocálicas deste artigo e as correspondentes estimativas de *pitch* usando um algoritmo construído especificamente para tal função [9]. Isto funciona como um indicador (uma validação parcial) que o método de estimação do comportamento vocálico proposto neste artigo fornece dados coerentes com o esperado. Vale salientar que os resultados contém mais informação que a extração de *pitch* e, por este motivo, a complexidade dos algoritmos não foi comparada. O vetor vocálico assim extraído pode ter múltiplas aplicações: síntese de voz, reconhecimento de locutor, por exemplo. Um "vetor de parâmetros" bastante útil na caracterização do locutor, além do vetor de *pitch* de cada som vocálico, pode ser um vetor com sete elementos contendo a distância inter-raias de cada som vocálico (inclinação da regressão):

$$\rho=((a),(é),(ê),(i),(ó),(ô),(u),(ã)\ ), \qquad (2)$$

em que (.) indica uso da frequência associada à vogal. Por exemplo, os locutores Ricardo e Lizandra têm perfis vocálicos expressos por: $\rho_{Ricardo}$= (102.6 89.0 102.0 82.0 83.3 94.5 87.6) e $\rho_{Lizandra}$=(210.3 197.7 328.0 203.1 224.9 164.4 103.4). Modelo similar foi empregado com sucesso em um recente sistema de reconhecimento de locutor [10]. Uma análise do comportamento espectral, com base no gênero, conduziu a uma constatação experimental

DOI: 10.5540/03.2015.003.01.0469  

curiosa. Ao se obter as curvas de ajuste de regressão para estimar a periodicidade das raias, observou-se (no espaço amostral investigado) que uma correlação maior é obtida para a voz masculina. A aderência das curvas ao modelo linear foi sempre mais fraca para a voz feminina, sugerindo que o grau de periodicidade dos sons vocálicos provenientes de locutores femininos é menor. A despeito da ausência de explicação plausível, o fato é, no mínimo, um achado interessante que provoca pesquisar mais aprofundadamente tais efeitos. O uso do Matlab® e da interface gráfica foi proposto apenas como uma ferramenta de melhor visualização, análise e validação dos resultados. Entretanto, um dos atrativos do método é a sua simplicidade para uso em sistemas embarcados. Os valores de frequência para as amostras coletadas com o auxilio do programa Audacity® são muito próximos dos valores obtidos para o identificador. Resta verificar o comportamento com sinais de voz contendo trechos alternados vocálicos e não vocálicos, como ocorre em fala natural. Constatou-se, em praticamente todos os casos, que há uma flutuação na cadencia das raias, de forma que o espectro apresenta "pequenos desvios" em torno dos "harmônicos teóricos" para um sinal periódico. Uma linha de investigação neste escopo é o uso das séries quase-harmônicas de Fourier, recentemente introduzidas [11], que parece uma ferramenta natural para modelar tal comportamento. Isto pode resultar em sintetizadores com voz mais natural.

**Agradecimento**. Ao prof. Ricardo Campello e a um revisor anônimo cuidadoso.

## Referências

[1] J. Holmes and W. Homes, *Speech Synthesis and Recognition*. Second Edition, Taylor & Francis, 2001.

[2] L. R. Rabiner and R. W. Schafer. Introduction to Digital Speech Processing. Publishers Inc., 2007.

[3] S.V. Vasegnhi. Multimedia Signal Processing - Theory and Applications in Speech, Music and Communications. First Edition, John Wiley, 2007.

[4] F. Rumsey and T MccormicK. Sound and Recording. Fifth Edition, Elsevier, 2006.

[5] W.C. Chu. Speech coding algorithms – Foundation and Evolution of Standardized Coders. John Wiley, 2003.

[6] A. V. Oppenheim; R. W. Schafer. Discrete-Time Signal Processing. Pearson, 2010.

[7] L. Rabiner and B. H. Juang. Fundamentals of Speech Recognition. Prentice-Hall, 1993.

[8] H. M. Oliveira, Análise de Fourier e Wavelets: Sinais Estacionários e não Estacionários. Editora Universitária da UFPE, 2007.

[9] S. Xuejing, Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio, IEEE Acoustics, Speech, and Signal Processing Int. Conf., ICAASP, pp.333-336, (2002).

[10] R.F.B. Sotero Filho, H.M. de Oliveira, Reconhecimento de Locutor baseado em Mascaramento Pleno em Frequência por Oitavas. 7º Congresso de Engenharia de Áudio, São Paulo, SP. Anais do AES Brasil 2009 pp.61-66, (2009).

[11] V. Vermehren V., J. E. Wesen, H.M. de Oliveira, Close Approximations for Daublets and their Spectra, IEEE/SBrT International Telecommunication Symposium, ITS 2010, September 06-09, Manaus, AM , Brazil, (2010).

 