Criação de modelos estocásticos para a síntese de vogais considerando os pulsos glotais de Rosenberg e de Liljencrants-Fant com parâmetros unificados

Diego Santana Marques Bahiano e Edson Cataldo

***Resumo*— Na produção da voz, a variação dos comprimentos dos ciclos glotais em relação a um valor médio causada pelo movimento (quase) periódico das cordas vocais gera o fenômeno aleatório conhecido por *jitter*. Seu estudo é importante devido a aplicações como a identificação de patologias relacionadas à voz, melhoria da naturalidade da voz sintetizada e a calibração de algoritmos de processamento de sinais para identificação de ciclos glotais. O objetivo deste trabalho é propor modelos estocásticos para a produção da voz, com geração do *jitter*, considerando a variação instantânea do intervalo de tempo glotal como um processo estocástico. Sinais de vozes são sintetizados com baixos níveis de *jitter*, vozes normais, e com níveis de *jitter* mais altos, gerando vozes roucas.**

***Palavras-Chave*—*jitter*, processo estocástico, sinal glotal, processamento de sinal de voz.**

***Abstract*— In voice production, the variation of the glottal cycles lengths in relation to a mean value caused by the (quasi) periodic movement of the vocal folds generates the random phenomenon known as jitter. Its study is important due to applications as the identification of voice-related pathologies, the improvement of the synthesized voice naturalness and the generation of voice signals to calibrate signal processing algorithms and to help in detecting glottal cycles, as well. The aim of this work is to build stochastic models for voice production, with generation of jitter, considering the instantaneous variation of glottal time interval as a stochastic process. Voice signals are synthesized with low levels of jitter, normal voices, and with higher levels of jitter, generating hoarse voices.**

***Keywords*— jitter, stochastic process, glottal signal, voice signal processing.**

I. Introdução

O ser humano é capaz de expressar seus sentimentos, pensamentos e vontades através da voz, sendo esta um meio de comunicação de grande importância para a vida em sociedade. O fluxo de ar proveniente dos pulmões, após passar pelas cordas vocais, torna-se um sinal de pressão acústica (quase) periódico que irá se propagar pelo trato vocal, sendo filtrado e amplificado, até ser irradiado pela boca, gerando a voz.

A especificidade dos órgãos envolvidos na produção da voz torna o sistema fonador único e injetivo para cada indivíduo. É possível identificar neste contexto alguns parâmetros que estão, inclusive, associados a patologias nas cordas vocais, o que é o caso do parâmetro *jitter*, que diz respeito às pequenas flutuações dos intervalos glotais dos pulsos glotais gerados. Medidas de *jitter* superiores a 1% causam rouquidão na voz, podendo até ser indício da existência de patologias nas cordas vocais [1].

Modelos de *jitter* também podem auxiliar na melhora da naturalidade de vozes sintetizadas, simular vozes roucas e/ou com características de patologias e, ainda, ajudar a calibrar sistemas de processamento de sinais de voz que necessitem identificar ciclos glotais.

Schoentgen apresentou em [2] modelos de *jitter* que simulavam os distúrbios nas frequências instantâneas de vibração das paredes glotais causando perturbações nos comprimentos dos ciclos glotais. Em [3], é apresentado um algoritmo que transforma a voz normal em uma voz rouca. Já em [1], [4], a partir de modelos mecânicos massa-mola-amortecedor para o movimento das cordas vocais, e considerando a rigidez da mola como um processo estocástico, foi obtido *jitter* em sinais de voz sintetizados, sendo possível controlar o nível do *jitter* com mudança de certos parâmetros.

Este trabalho implementa, a partir dos modelos determinísticos de pulso glotal de Rosenberg e de Liljencrants-Fant com parâmetros unificados, modelos estocásticos para a variação instantânea do intervalo de tempo glotal, visando produzir sinais correspondentes a vogais sustentadas sintetizadas semelhantes às naturais produzidas por um indivíduo, com e sem rouquidão.

II. Modelagem matemática determinística da produção da fala: o modelo fonte-filtro

Essa teoria acústica apresentada por Gunnar Fant [5], em 1970, consiste em modelar o mecanismo de produção da fala como a convolução entre uma fonte de excitação (o pulso glotal) e um sistema de filtros digitais lineares [6], representando o trato vocal e a irradiação pelos lábios/narinas, conectados em série. No domínio do tempo, o sinal de voz produzido pode ser representado pela convolução $s[n] = g[n] * t_v[n] * r[n]$, onde $s[n]$ é o sinal da fala, $g[n]$ é o sinal da fonte de excitação (sinal glotal), $t_v[n]$ é a resposta ao impulso do trato vocal e $r[n]$ é a resposta ao impulso da boca (irradiação).

A. Fonte de excitação

Na literatura, são encontrados vários modelos matemáticos determinísticos para o sinal glotal (ou fluxo glotal), como o modelo de Rosenberg [7], o de Fant [8], o de Liljencrants-Fant (LF) [9] e o de Klatt [10]. Em [11], foi proposta uma

Diego Santana Marques Bahiano e Edson Cataldo, Escola de Engenharia, Departamento de Engenharia de Telecomunicações, Programa de Pós-graduação em Engenharia Elétrica e de Telecomunicações Universidade Federal Fluminense (UFF), Niterói-RJ, Brasil. E-mails: diegobahiano@live.com e ecataldo@id.uff.br.

estrutura unificada para estudar as propriedades nos domínios de tempo e frequência dos modelos de fluxo glotal, mostrando que os mesmos podem ser representados por cinco parâmetros: Amplitude de vozeamento (A_v); Velocidade de fechamento (E_e); Quociente de abertura (O_q); Coeficiente de assimetria (α_m); e Constante de tempo de fase de retorno (T_a).

Nesse trabalho são utilizados dois modelos de pulsos glotais, o de Rosenberg e o de LF, o primeiro por ser bem mais simples e o outro por ser mais próximo da realidade, embora mais complexo. O modelo de Rosenberg é descrito pela Equação (1) :

$$U_g(t) = \begin{cases} \frac{A_v}{2}[1-\cos(\pi t/T_p)] & ,\ 0 \le t \le T_p \\ A_v \cos(\pi(t-T_p)/(2T_n)) & ,\ T_p < t \le T_p + T_n \\ 0 & ,\ T_p + T_n < t \le T_0\,. \end{cases} \tag{1}$$

Este modelo trigonométrico é dado por uma função definida por várias sentenças. A primeira sentença descreve o comportamento do fluxo glotal durante a fase de abertura das cordas vocais e a segunda sentença durante a fase de fechamento, onde T_0 é o período fundamental, $T_p = \alpha_1 T_0$ é o tempo de abertura das cordas vocais e $T_n = \alpha_2 T_0$ é o tempo de fechamento, sendo α_1 e α_2 parâmetros relacionados às fases de abertura e fechamento, respectivamente. Em relação aos parâmetros unificados, há uma relação direta estabelecida para este modelo e, para isso, basta tomar $T_p = \alpha_m O_q T_0$; e $T_n = T_e - T_p = (1-\alpha_m) O_q T_0$. Sendo assim, conclui-se que $O_q = \alpha_1 + \alpha_2$ e $\alpha_m = \alpha_1/(\alpha_1+\alpha_2)$.

Já o modelo de LF, com parâmetros unificados [11], é representado pela Equação (2).

$$U_g(t) = \begin{cases} U_{g1}(t) & ,\ 0 \le t < O_q T_0 \\ U_{g2}(t) & ,\ O_q T_0 \le t \le T_0\,. \end{cases} \tag{2}$$

$$U_{g1}(t) = -\frac{E_e e^{-\varepsilon_1 O_q T_0}}{sen\left(\frac{\pi}{\alpha_m}\right)\left(\varepsilon_1^2 + \left(\frac{\pi}{\alpha_m O_q T_0}\right)^2\right)}\left(\frac{\pi}{\alpha_m O_q T_0} + \varepsilon_1 e^{\varepsilon_1 t} sen\left(\frac{\pi}{\alpha_m O_q T_0} t\right) - \frac{\pi}{\alpha_m O_q T_0} e^{\varepsilon_1 t} cos\left(\frac{\pi}{\alpha_m O_q T_0} t\right)\right). \tag{3}$$

$$U_{g2}(t) = -E_e\left(\frac{1}{\varepsilon_2 Q_a (1-O_q) T_0} - 1\right)\left(T_0 - t + \frac{1-e^{\varepsilon_2 (T_0 - t)}}{\varepsilon_2}\right). \tag{4}$$

Este modelo é composto por uma parte senoidal modulada por uma exponencial crescente (entre 0 e $O_q T_0$), seguida de uma fase de retorno exponencial decrescente (entre $O_q T_0$ e T_0). A expressão temporal da equação acima corresponde ao modelo no caso de um fechamento não abrupto ($Q_a > 0$), cujo parâmetro Q_a tem relação com a duração efetiva da fase de retorno através de $T_a = Q_a(1-O_q)T_0$. O parâmetro E_e representa a máxima amplitude de excitação da derivada do fluxo glótico. As constantes ε_1 e ε_2 são obtidas através das condições de continuidade do fluxo glótico e de sua derivada no instante de excitação máxima.

B. Trato vocal

No caso em estudo, tem-se o trato vocal como um filtro ressonante que faz com que o fluxo de ar que passa pelas cordas vocais chegue aos lábios e produza a voz humana. Durante a emissão sonora, a energia dos vários harmônicos da fonte glotal não é transmitida igualmente, uma vez que as frequências baixas, que incluem aí também os formantes mais baixos, concentram maior energia fornecida pela fonte glotal. É importante dizer que, embora a largura de faixa do formante não seja necessariamente um fator crítico na percepção de vogais, há possivelmente uma largura de faixa ótima que facilita a discriminação e identificação de vogais [12]. Portanto, há que se considerar, além dos formantes, suas larguras de faixa.

C. Radiação labial

A radiação labial pode ser aproximada através de um filtro FIR (resposta ao impulso finita) passa-alta de primeira ordem, tendo como função de transferência $R(z) = 1 - \alpha z^{-1}$, para $\alpha \in [0.95, 0.99]$. Seu efeito é ampliar as componentes de alta frequência do sinal referente à cavidade bucal.

III. O FENÔMENO DO *jitter*

O *jitter* é um fenômeno presente na voz humana que provoca pequenas perturbações aleatórias na frequência fundamental. É uma característica acústica do sinal de voz que é largamente utilizada para detecção de patologias no aparelho fonador humano. Para determinação do *jitter*, são utilizados diferentes tipos de medidas. Neste trabalho, a análise ficou restrita ao *jitter* local, definido pela Equação (5). Nesta, tem-se T_i como o intervalo de tempo do pulso i, $i \in \{1, 2, 3, ..., N\}$, N como o número de pulsos executados e Jit_{loc} como a medida de dispersão do *jitter* utilizada. Para valores acima de 1,040%, as vozes começam a ficar mais ásperas e roucas, podendo ser um indício da existência de patologia no sistema fonador do indivíduo.

$$Jit_{loc} = 100\left(\frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i}\right). \tag{5}$$

IV. MODELOS ESTOCÁSTICOS PARA O SINAL GLOTAL

Para cada um dos modelos determinísticos de sinal glotal utilizados neste trabalho, o intervalo de tempo glotal será considerado uma variável aleatória e tomaremos a variação no tempo desse intervalo de tempo glotal como um processo estocástico. Para cada modelo, Rosenberg e LF, duas densidades espectrais são associadas ao processo estocástico proposto. Dessa forma, são criados, no total, quatro modelos estocásticos para o pulso glotal. Seja t_g o intervalo de tempo glotal; isto é, o intervalo de tempo associado a um pulso glotal, no caso determinístico. No caso real, t_g terá comprimento variável. Isto é, para cada ciclo glotal, t_g terá um valor diferente, sendo considerado uma variável aleatória, que será denotada por T_g. No caso determinístico, ao ser discretizado o sinal, o intervalo de tempo glotal pode ser dividido em N intervalos de comprimento δt; isto é, $t_g = N\delta t = \sum_{i=1}^{N} \delta t$, com δt e N escolhidos. No caso aleatório, tem-se $T_g = \sum_{i=1}^{N} \Delta t(t_i)$, sendo

$\Delta t(t_i)$ amostras de uma realização do processo estocástico $\Delta t(t)$ que é o grande objetivo desse trabalho e descrito a seguir.

Considere T_{g_i} cada instante to intervalo de tempo glotal, com $i = 1, \ldots, N$. Assim, tem-se que $T_{g_{i+1}} = T_{g_i} + \Delta t(t_i)$, com $i = 1, ..., N$. E, $\Delta t(t_i)$, $i = 1, \ldots, N$ são amostras de uma realização do processo estocástico $\Delta t = \{\Delta t(t), t \in \mathbb{R}\}$, construído baseado nas ideias do modelo proposto em [1], com as seguintes características:

1) Para todo t, $0 < \Delta t_0 \leq \Delta t(t)$, onde Δt_0 é uma constante positiva independente de t.
2) $\Delta t(t)$ É um processo estacionário, não gaussiano, uma vez que só pode ter valores positivos.
3) $E\{(\Delta t(t))^2\} < +\infty$ para todo t (processo estocástico de segunda ordem), tal que $E\{\Delta t(t)\} = \underline{\Delta t(t)} > \Delta t_0 > 0$, sendo considerado contínuo no sentido quadrático médio para garantir a existência da densidade espectral de potência associada.

É introduzido um processo estocástico real gaussiano de segunda ordem $Y = \{Y(t), t \in \mathbb{R}\}$ centrado e contínuo sob o ponto de vista quadrático médio, estacionário e ergódico, fisicamente realizável. A representação do processo estocástico $\Delta t(t)$ pode ser escrita como:

$$\Delta t(t) = \Delta t_0 + (\underline{\Delta t} - \Delta t_0)(\underline{y} + Y(t))^2, \quad \forall t \in \mathbb{R}, \tag{6}$$

onde $\underline{y}$ é um parâmetro tal que $E\{(\underline{y} + Y(t))^2\} = 1$ e $E\{(\underline{y} + Y(t))^4\} < +\infty$, uma vez que $E\{\Delta t(t)\} = \underline{\Delta t}$ e $E\{(\Delta t(t))^2\} < +\infty$. O processo estocástico gaussiano $Y(t)$ é construído como um filtro linear, $Y = h * N_\infty$, dado pela convolução do ruído branco gaussiano centrado N_∞, cuja função densidade espectral de potência é constante e igual a $S_{N_\infty}(\omega) = 1/(2\pi)$ para todo real ω, pelo filtro $h = \mathfrak{F}^{-1}\{H\}$, que é a transformada inversa de Fourier da função resposta em frequência $H(\omega)$ [13] .

Duas densidades espectrais de potência (PSD) associadas ao processo estocástico Y serão consideradas neste trabalho:

$$S_Y(\omega) = \frac{1}{2\pi} \frac{a^2}{\omega^2 + b^2} \quad , a > 0, b > 0; \quad e \tag{7}$$

$$S_Y(\omega) = \frac{1}{2\pi} \frac{a^2}{(b^2 - \omega^2)^2 + 4\xi^2 b^2 \omega^2} \quad , a > 0, b > 0, \xi > 0. \tag{8}$$

Pode ser deduzido que o processo estocástico $\{\dot{Y}(t), t \in \mathbb{R}\}$, derivada do processo estocástico $\{Y(t), t \in \mathbb{R}\}$, é um processo estocástico de segunda ordem, porque $\int_{\mathbb{R}} \omega^2 S_Y(\omega) d\omega < +\infty$. Para o primeiro caso; isto é, densidade espectral definida pela Equação (7), o processo $Y(t)$ pode ser obtido como solução da equação diferencial estocástica de Itô, para $t >> t_0$, sendo t_0 um número real positivo:

$$dY(t) = -bY(t)dt + adW(t), \quad t > 0. \tag{9}$$

Com a condição inicial $Y(0) = 0$ a.s., onde W é processo de Wiener indexado por $[0, +\infty[$, pode ser provado [13] que a Equação (9) tem uma única solução.

Da Equação (7) temos que:

$$E\{(\underline{y} + Y(t))^2\} = 1 \\ \iff \underline{y}^2 + \int_{-\infty}^{+\infty} \frac{a^2}{2\pi(\omega^2 + b^2)} d\omega = 1 \iff \quad \underline{y}^2 = 1 - \frac{a^2}{2b}. \tag{10}$$

Consequentemente, $0 < a < \sqrt{2b}$ e $b > 0$.

A fim de garantir a convergência da solução da Equação Diferencial de Itô, é importante que seja analisado o comportamento do valor esperado e do momento de segunda ordem para realizações do processo estocástico Δt para fins de verificação de ergodicidade do processo estocástico e convergência da solução da equação de Itô.

Para o segundo caso; isto é, para a densidade espectral definida pela Equação (8), considera-se $\{\mathbb{Y}(t) = (Y(t), \dot{Y}(t), t \geq 0)\}$ como processo estocástico com valores em $\mathbb{R}^2$. Para valores de $t >> t_0$, sendo t_0 um número real positivo, O processo $Y(t)$ é definido a partir da solução da equação diferencial estocástica de Itô:

$$d\mathbb{Y}(t) = -[\alpha]\mathbb{Y}(t)dt + [\beta]d\mathbb{W}(t), \quad t > 0, \tag{11}$$

com a condição inicial $\mathbb{Y}(0) = (0, 0)$ a.s., onde $\mathbb{W}(t), t \geq 0$ é processo estocástico normalizado de Wiener indexado por $[0, +\infty[$, $[\alpha]$ é uma matriz real de dimensão (2×2) e $[\beta]$ é um vetor real:

$$[\alpha] = \begin{bmatrix} 0 & 1 \\ -b^2 & -2\xi b \end{bmatrix} \quad , \quad [\beta] = \begin{bmatrix} 0 \\ a \end{bmatrix} \quad . \tag{12}$$

Pode ser provado [13] que a Equação (11) tem uma única solução $\{\mathbb{Y}(t), t \geq 0\}$.

Da Equação (8) temos que:

$$E\{(\underline{y} + Y(t))^2\} = 1 \iff \\ \underline{y}^2 + \int_{-\infty}^{+\infty} S_Y(\omega) d\omega = 1 \quad \iff \quad \underline{y}^2 = 1 - \frac{a^2}{4\xi b^3} \quad . \tag{13}$$

Consequentemente, $0 < a^2 < 4\xi b^3$, $b > 0$ e $\xi > 0$.

As soluções das Equações (9) e (11) foram obtidas através de método numérico, utilizando-se o método semi-implícito de Euler Maryuama, considerando Y após um instante grande o suficiente que garanta a convergência da solução.

Para o primeiro modelo de densidade espectral, é possível variar apenas dois parâmetros, um deles controla o nível de *jitter* e o outro está relacionado à variação de frequência e, mais precisamente, à convergência da solução da equação de Itô. Para o segundo modelo de densidade espectral, surge mais um parâmetro, que será importante para controlar o chamado colorido frequencial do sinal de voz gerado.

V. SIMULAÇÕES

Os parâmetros utilizados nos modelos são os seguintes:

1) Para a fonte glotal: Os pulsos glotais gerados com o modelo de Rosenberg apresentam $f_0 = 98Hz$, $Av = 7$, $\alpha_1 = 0.50$ e $\alpha_2 = 0.30$. Já pelo modelo de LF, foram utilizadas as seguintes constantes: $f_0 = 98Hz$, $\varepsilon_1 = 221$, $\varepsilon_2 = 1566$, $O_q = 0.80$, $\alpha_m = 0.75$ e $Q_a = 0.30$. Foi escolhida uma frequência de amostragem para o sistema de $f_m = 20000Hz$, a constante $\Delta t_0 = 1/40000$ e estipulado um tempo $T_f = 3\,s$ de sustentação da vogal.
2) Para o trato vocal: Os formantes, assim como as suas respectivas larguras de faixa, estão definidas, em Hertz, na Tabela I. Os três primeiros formantes (F_1, F_2 e F_3), ao contrário dos formantes superiores (F_4 e F_5), têm menor dependência com o locutor e prestam-se, principalmente, para diferenciar as vogais. Os formantes,

juntamente com as bandas de passagem, foram selecionados da literatura [14].

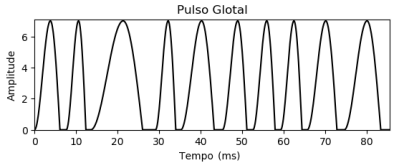
TABELA I: Formantes e bandas de passagem, em Hertz.

Fonemas	F_1	F_2	F_3	F_4	F_5
\a\	900	1300	2000	2200	2500
\e\	450	1700	2000	2200	2310
\i\	300	1900	2100	2200	2490
\o\	500	800	2150	2200	2490
\u\	360	700	2170	2200	2330
Banda de Passagem	41	52	70	32	100

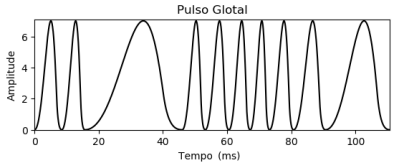
3) Para a radiação labial: Foi escolhido $\alpha = 0.95$, resultando na função de transferência $R(z) = 1 - 0.95z^{-1}$.

Nas simulações desenvolvidas neste trabalho, os parâmetros supracitados serão fixos e, portanto, será analisada a influência das variáveis associadas às duas diferentes densidades espectrais de potência utilizadas para os processos estocásticos estabelecidos. Em primeiro lugar, foram analisados os modelos de pulsos glotais de Rosenberg e LF com $S_Y(\omega)$ a dois parâmetros. Foi verificado que para um número de iterações, de cada realização do processo estocástico $\Delta t(t)$, superior a 400, o valor de $E\{\Delta t(n)\}$ converge.

Na Figura 1, estão apresentados dois exemplos de sinais glotais, obtidos com as simulações, para valores de *jitter* local 50,36% e 48,83%, para o modelo de Rosenberg e para o modelo de LF, respectivamente. Esses valores de *jitter* são muito maiores do que, em geral, são obtidos nas vozes de humanos. Foram escolhidos, porém, para que ficasse clara a geração de *jitter* com os modelos propostos.



(a) Sinal glotal com modelo Rosenberg com *jitter*.



(b) Sinal glotal com modelo LF com *jitter*.

Fig. 1: Sinal glotal com *jitter* ($a = 14.1$ e $b = 100.0$).

Para o objetivo deste trabalho serão escolhidos valores de a e b que gerem valores de *jitter* mais próximos da realidade, tanto para vozes normais como para vozes roucas, podendo caracterizar ou não patologias. O comprimento de cada pulso sofre uma expansão ou contração, devido à modelagem do *jitter*, sendo $a = 0$ o caso determinístico. Os algoritmos foram desenvolvidos em Python. A Tabela II apresenta os valores obtidos de *jitter* de acordo com variações de a e b. Como os valores de *jitter* obtidos com o modelo de Rosenberg e de LF são bem próximos, decidiu-se apresentar apenas os gerados com o modelo de LF. A maior diferença entre os pulsos gerados por um ou outro desses dois casos (Rosenberg e LF) está na inteligibilidade do som gerado. Para valores de *jitter*, as diferenças estão nas duas possibilidades de densidade espectral.

TABELA II: Valores de *jitter* locais simulados, para o modelo de LF, com densidade espectral a dois parâmetros.

Casos	Jit_{loc}	Casos	Jit_{loc}
a=10 e b=100	46.85%	a=0 e b=10000	0.0%
a=10 e b=1000	10.04%	a=2 e b=10000	0.222%
a=10 e b=10000	1.016%	a=5 e b=10000	0.600%
a=10 e b=100000	0.096%	a=10 e b=10000	1.016%
a=10 e b=1000000	0.011%	a=100 e b=10000	8.483%

Uma outra forma interessante de verificar a presença de *jitter* consiste na construção da função densidade de probabilidade (PDF) da variável aleatória, denotada por F_{0g}, dada pelo inverso variável aleatória associada ao intervalo de tempo glotal; isto é, $F_{0g} = 1/T_g$. A Figura 2 mostra o gráfico da PDF de F_{0g} para diferentes valores de a e b .

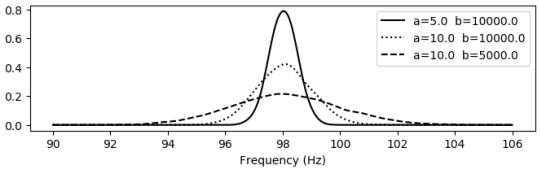


Fig. 2: PDF de F_{0g} para o modelo LF considerando a PSD ($S_Y(\omega)$) com dois parâmetros.

Em seguida, foram analisados os modelos de pulsos glotais de Rosenberg e LF com $S_Y(\omega)$ a três parâmetros. Define-se um novo parâmetro c_b, tal que $b = 2\pi f_0 c_b$, sendo f_0 a frequência fundamental desejada para o sinal simulado. Esse novo parâmetro é importante para mostrar a sensibilidade da variação de frequência do sinal gerado em relação a esse parâmetro. A Tabela III mostra valores de *jitter* obtidos a partir de combinações dos valores de a, c_b e ξ, para o modelo de LF e densidade espectral de potência a três parâmetros, sendo $a = 0$ o caso determinístico.

TABELA III: Valores de *jitter* locais, para o modelo de LF, com densidade espectral de potência a três parâmetros.

Casos	a	c_b	ξ	*jitter* Local
I	0.0	1.0	0.1	0.0%
II	1.0	1.0	0.1	0.109%
III	3.0	1.0	0.1	1.022%
IV	5.0	1.0	0.1	3.126%
V	10.0	1.0	0.1	9.722%
VI	50.0	1.0	0.1	87.33%
VII	3.0	1.5	0.1	0.169%
VIII	3.0	2.0	0.1	0.067%
IX	3.0	5.0	0.1	0.001%
X	3.0	1.0	0.3	1.006%
XI	3.0	1.0	0.5	0.952%

Em relação à variável c_b, percebe-se que ela tem relação direta com a frequência em que a densidade espectral de

potência apresenta valor máximo. Outrossim, percebe-se um decréscimo do efeito *jitter* com o seu aumento, como pode ser observado na Tabela III, nos casos II, VII, VIII e IX. Nos casos II, X e XI, o aumento de ξ proporcionou uma redução no efeito *jitter*. Embora todos os parâmetros modifiquem os valores de *jitter*, cada um tem um objetivo no modelo. Os valores de a regulam o nível de *jitter* de forma mais sensível, c_b está diretamente ligado a pequenas variações da frequência fundamental e, principalmente, à convergência da equação de Itô e ξ está diretamente ligado ao colorido frequencial do sinal.

A Figura 3 mostra a PDF de F_{0g} para diferentes valores de a, c_b e ξ. Percebe-se que a diminuição do valor de a gerou uma redução na dispersão dos valores das frequências fundamentais dos pulsos gerados. O aumento do parâmetro c_b acarreta um deslocamento do gráfico de $S_Y(\omega)$ para a direita, aumentando o valor do harmônico mais importante do sinal. A variação de ξ muda a distribuição de frequências do sinal ou, de maneira mais informal, no seu colorido frequencial.

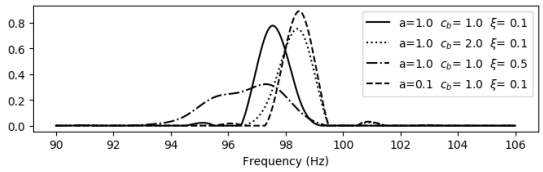


Fig. 3: PDF de F_{0g} para o modelo LF considerando a PSD ($S_Y(\omega)$) com três parâmetros.

Os arquivos de áudio das vogais sintetizadas com diferentes níveis de *jitter* podem ser ouvidos em:

1) Modelo de Rosenberg e densidade espectral a 2 parâmetros: https://www.dropbox.com/sh/x1ff47phep4s944/AAD8Z3dne2RKJ2pxcO5_DR5Ya?dl=0 ;
2) Modelo de Rosenberg e densidade espectral a 3 parâmetros: https://www.dropbox.com/sh/y75w052vx9x90rc/AABGc3_i8ovZPTvICvETZDlsa?dl=0 ;
3) Modelo de LF e densidade espectral a 2 parâmetros: https://www.dropbox.com/sh/tdlorg477pivnd4/AAB6cBfKfYlBeVpM-ZEeiBsra?dl=0 ; e
4) Modelo de LF e densidade espectral a 3 parâmetros: https://www.dropbox.com/sh/c0vx0erwqhguldl/AAC9U33XL0QxNAgeR0lNBz0fa?dl=0 .

Foi aplicada uma técnica de modulação da amplitude do sinal de saída, o que forneceu aos áudios uma melhora na característica atenuadora no início e no final da pronúncia de cada vogal. Os arquivos de áudio das vogais, sintetizados com diferentes frequências fundamentais, podem ser ouvidos em: https://www.dropbox.com/sh/dcwynhpxcvptakj/AAAbsF0Ny274V3YahenvpwUYa?dl=0 .

VI. Conclusões

A modelagem estocástica do sinal glotal mostrou-se eficiente para gerar o *jitter* nos sinais de voz. Foi possível obter sinais com diferentes níveis de *jitter* e, principalmente, níveis que podem indicar rouquidão e/ou patologias relacionadas às cordas vocais. O *jitter* está presente em todas as vozes humanas e acrescenta um caráter mais natural à voz sintetizada.

É possível perceber que o som gerado pelo modelo de Rosenberg é mais metalizado e com maior energia acústica quando comparado ao de LF, o qual é mais suave, mais aflautado, com diferentes propriedades, o que permite inferir que a forma de onda do pulso glotal está interligada não somente à energia como também à qualidade do som gerado. Conforme era esperado, identificou-se que na voz gerada com *jitter*, à medida que o Jit_{loc} superava 1% nas vogais sintetizadas, o som ganhava uma característica adicional que o tornava áspero e rouco, podendo ser sinal de alguma patologia.

Para o modelo de densidade a dois parâmetros, verificou-se que o aumento do valor do parâmetro a acarreta no aumento do nível de *jitter*. Uma vez escolhido o valor de b, o parâmetro a atua como um ajuste fino, estabelecendo valores de *jitter* de acordo com as pequenas variações desejadas.

A comparação entre as funções de densidade de probabilidade da frequência fundamental permite perceber que o *jitter* foi gerado, como inicialmente proposto. Em particular, para as densidades espectrais a três variáveis, há um número maior de possibilidades de geração de *jitter* e modificações das frequências do sinal gerado. O passo adiante é o de identificar parâmetros dos modelos resolvendo problemas estocásticos inversos a partir de sinais de vozes reais.

Agradecimentos

Os autores agradecem ao CNPq pelo apoio financeiro.

Referências

[1] E. Cataldo and C. Soize, "Voice signals produced with jitter through a stochastic one-mass mechanical model", *Journal of voice*, Vol. 31, pp. 111.e9-111.e18, 2017.

[2] J. Schoentgen, "Stochastic models of jitter", *The Journal of the Acoustical Society of America*, Vol. 109, n. 4, p. 1631-1650, 2001.

[3] D. Ruinskiy and Y. Lavner, "Stochastic models of pitch jitter and amplitude shimmer for voice modification", in *IEEE 25th Convention of Electrical and Electronics Engineers in Israel*, pp. 489-493, 2008.

[4] E. Cataldo and C. Soize, "Stochastic mechanical model of vocal folds for producing jitter and for identifying pathologies through real voices", *Journal of Biomechanics*, Vol. 74, pp. 126-133, 2018.

[5] G. Fant, "Acoustic theory of speech production", 2nd edition, Mouton, The Hague, 1970. 328 p.

[6] L. R. Rabiner and R. W. Shafer, "Digital Processing of Speech Signals", Prentice-Hall, 1978. 962 p.

[7] G. Degottex, "Glottal Source and Vocal-Tract Separation - Estimation of glottal parameters, voice transformation and synthesis using a glottal model", Tese de doutorado, Université Paris, 2010.

[8] G. Fant, "Vocal-Source Analysis - A Progress Report", *STL-QPSR*, nos. 3-4, pp. 31-53, 1979.

[9] J. Liljencrants, G. Fant and Q. Lin, "A Four Parameter Model of Glottal Flow", *STL-QPSR*, n. 4, pp. 1-13, 1985.

[10] L. C. Klatt, "Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers", *Journal of the Acoustical Society of America*, n. 87, pp. 820-857, 1990.

[11] N. H. Bernardoni, C. d'Alessandro e B. Doval, "Glottal flow models: waveforms, spectra and physical measurements", *Forum Acusticum*, Sevilla, Spain, p. 1, 2002.

[12] R. D. Kent e C. Read, "Análise Acústica da fala", Tradução de Alexsandro Rodrigues Meireles, 1a. edição, São Paulo, Cortez, 2015. 503 p.

[13] C. Soize, "The Fokker-Planck Equation for Stochastic Dynamical Systems and its Explicit Steady State Solutions", *World Scientific*, Singapore, 1994.

[14] L. E. B. Sandoval e E. Cataldo, "Comparação de modelos de sinal glotal na síntese de vogais, nos casos de vogal sustentada e de voz cantada, considerando sons na língua espanhola", Dissertação de Mestrado, Universidade Federal Fluminense, Niterói, Rio de Janeiro, 2018.