JOÃO PEDRO HALLACK SANSÃO

MEDIDA DA RELAÇÃO HARMÔNICO/RUÍDO EM VOZES DISFÔNICAS PELO PROCESSAMENTO DIGITAL DE IMAGENS ESPECTROGRÁFICAS

Belo Horizonte
09 de junho de 2009

Universidade Federal de Minas Gerais
Escola de Engenharia
Programa de Pós-Graduação em Engenharia Elétrica

MEDIDA DA RELAÇÃO HARMÔNICO/RUÍDO EM VOZES DISFÔNICAS PELO PROCESSAMENTO DIGITAL DE IMAGENS ESPECTROGRÁFICAS

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Engenharia Elétrica.
Área de Concentração: Engenharia de Computação e Telecomunicações
Linha de Pesquisa: Sistemas de Computação

JOÃO PEDRO HALLACK SANSÃO

Belo Horizonte
09 de junho de 2009

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Medida da relação harmônico/ruído em vozes disfônicas pelo processamento digital de imagens espectrográficas

JOÃO PEDRO HALLACK SANSÃO

Dissertação defendida e aprovada pela banca examinadora constituída por:

Prof. Dr. Maurílio Nunes Vieira – Orientador
Departamento de Física / Instituto de Ciências Exatas
Universidade Federal de Minas Gerais

Prof. Dr. Hani Camille Yehia – Co-orientador
Departamento de Engenharia Eletrônica / Escola de Engenharia
Universidade Federal de Minas Gerais

Prof. Dra. Ana Cristina Côrtes Gama
Departamento de Fonoaudiologia / Faculdade de Medicina
Universidade Federal de Minas Gerais

Dr. Cristiano Rodrigues de Carvalho
Google, Brasil

Belo Horizonte, 09 de junho de 2009

Resumo

Este trabalho apresenta a S^2NR, *Spectrographic Signal-to-Noise Ratio*, uma medida da relação sinal/ruído obtida através do processamento da imagem do espectrograma de uma vogal. O algoritmo utilizado baseia-se em ferramentas de identificação de impressões digitais, as quais apresentam traçados com linhas paralelas que se assemelham aos espectrogramas de vogais. Para validação do algoritmo, estabeleceu-se uma plataforma de testes que permite a síntese de diferentes vogais, com controle de freqüência fundamental, ruído branco aditivo e perturbações ciclo-a-ciclo na amplitude (*shimmer*) e no período fonatório (*jitter*). Para fins de comparação, geraram-se vogais com níveis conhecidos da relação sinal/ruído. Em seguida, para cada caso mediu-se a relação sinal/ruído utilizando a S^2NR e um algoritmo baseado na demarcação da periodicidade da vogal. A S^2NR mostrou-se, na maioria das situações com voz sintética, mais robusta a perturbações de *jitter* e de *shimmer* e com menor sensibilidade à vogal. Foram testadas freqüências fundamentais masculinas e femininas com tratos vocais para as vogais /a/, /i/ e /u/. O teste inicialmente foi feito variando, de forma independente, o nível de *jitter* e de *shimmer* desde a condição de inexistência até valores extremos (0% a 3% para *jitter* e 0% a 30% para *shimmer*). Sob *jitter*, com $f_o = 120\,Hz$, os valores de desvio máximo em relação à referência foram de $2,1\,dB$, $11,5\,dB$ e $2,9\,dB$ para as vogais /a/, /i/ e /u/, respectivamente. Já sob *shimmer*, estes valores foram de $2,5\,dB$, $4,4\,dB$ e $3,6\,dB$. Em seguida, aplicaram-se as perturbações simultaneamente, não ocorrendo perdas de desempenho diferentes das observadas com perturbações individuais. Finalmente, o algoritmo S^2NR foi testado com vozes reais disfônicas predominantemente soprosas, resultando numa relação consistente com a classificação perceptiva de soprosidade. Em adição a estes testes, mostrou-se a utilização do algoritmo S^2NR em fala encadeada.

Abstract

This work presents the S^2NR, Spectrographic Signal-to-Noise Ratio, a signal-to-noise ratio measurement obtained from the processing of vowel spectrograms by using adaptations of fingerprint image enhancement algorithms. In order to validate the S^2NR method, a test bench was set to generate synthetic vowels with controlled values of fundamental frequency, amplitude, additive white noise, and cycle-to-cycle perturbations in the waveform amplitude (shimmer) and phonatory period (jitter). For comparison purposes, vowels were synthesized with known signal-to-noise ratio values. Next, the signal-to-noise ratio was measured with the S^2NR algorithm and a method based on time domain periodicity analysis. In most of the synthetic voices, the S^2NR exhibited a behavior more robust to jitter and shimmer perturbations than the time based algorithm, having also a reduced sensitivity to the vowel type. Both male and female fundamental frequencies were tested with /a/, /i/, and /u/ vocal tract shapes. Initially, jitter and shimmer were assessed independently, the simulated perturbation values varying from inexistent to extreme conditions in the human voice (0% to 3% for jitter, and 0% to 30% for shimmer). With jitter and $f_o = 120\,Hz$, the measured S^2NR estimates deviated from the reference values by $2.1\,dB$, $11.5\,dB$, and $2.9\,dB$ for /a/, /i/ and /u/ respectively. With shimmer, these differences were $2.5\,dB$, $4.4\,dB$, and $3.6\,dB$. Subsequently both perturbations were varied simultaneously within the same ranges, no performance degradation occurring other than those observed with separated perturbations. Finally, the S^2NR algorithm was tested with real, dysphonic, and predominantly breathy voices. Results showed a consistent relation between S^2NR values and perceptual ratings of breathiness. Additionally, the potential application of the S^2NR algorithm in running speech was explored.

Agradecimentos

Aos meus pais, Raphael e Suraia, e à minha irmã Júlia pelo apoio, amor e paciência dados durante a realização deste trabalho.

Ao professor Maurílio Nunes Vieira que me guiou durante toda minha formação acadêmica e tanto me ensinou da arte da engenharia.

Ao professor Hani Camille Yehia pelo inestimável apoio e por ter operado milagres que permitiram a conclusão deste trabalho.

Aos colegas do CEFALA que me motivaram a concluir este trabalho.

À professora Ana Cristina Côrtes Gama e ao Dr. Cristiano Rodrigues de Carvalho que deram valiosas contribuições para a melhoria deste texto.

Aos pacientes da *Royal Infirmary of Edinburgh* que cederam amostras de voz para a realização deste trabalho, e às fonoaudiólogas Ana Paula e Mariana que ajudaram na classificação perceptiva destas vozes.

Ao Prof. Guilherme Augusto Silva Pereira que ministrou a disciplina que me inspirou a criar o algoritmo apresentado neste trabalho.

Ao Dr. Peter Kovesi e a todos aqueles que compartilham seus códigos fonte para o bem do desenvolvimento científico.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

A análise da qualidade de voz é um campo vasto, que abrange áreas como a ciência e tecnologia da Fala, telecomunicações, a fonética e fonoaudiologia. A medida quantitativa da qualidade da voz faz que seja necessário criar instrumentos que traduzam características perceptivas em valores numéricos de atributos como rouquidão, aspereza e soprosidade (Kreiman e Gerratt, 1998).

A fonação é a produção sonora através da vibração das pregas vocais. Voz disfônica é aquela que apresenta anormalidades durante a fonação (Titze, 1995). Estas são entendidas como mudanças em parâmetros como a variação de freqüência fundamental, amplitude e outros atributos que fazem a voz de um determinado falante diferir daquela considerada normal para o próprio falante, ou para o grupo de mesma idade, sexo, dialeto e grupo cultural (Beech et al., 1993). Em geral, este tipo de anomalia decorre de alterações que afetem tanto a estrutura anatômica das pregas vocais devido a lesões (nódulos, pólipos, etc.) quanto seu comportamento fonatório devido à tensão anormal nos músculos da região da laringe.

Uma forma de se estudar a qualidade da voz é através de escalas perceptivas. Estas escalas surgem da rotulação (atribuição de adjetivos) de impressões observadas pelos ouvintes de uma determinada voz. Para evitar confusão entre os termos e manter uma padronização entre os especialistas, há propostas de normalização que na prática são aceitas pela a comunidade (Beech et al., 1993), destacando-se o *Voice Profile Analysis Scheme*, Esquema de Análise do Perfil de Voz (Laver et al., 1981) e a escala GRBAS (Hirano, 1981), explicados abaixo.

O VPAS é baseado fortemente nas possibilidades fisiológicas do trato vocal representadas por ajustes laríngeos e supralaríngeos (velofaríngeos, orais, nasais e labiais). As vozes são comparadas com uma referência (ajuste neutro) onde o modo de vibração das pregas é regular e sem escape de ar audível ou outros ruídos. No que tange os ajustes laríngeos possíveis, destacam-se os tipos de vibração:

- Modal ou falsete;

- Com ou sem crepitação (*creaky/vocal fry*);
- Com ou sem escape de ar (*whisper/breathiness*);
- Com ou sem irregularidade vibratória (*harshness*);
- Com ou sem desvio de tensão laríngea (hiper/hipotensão).

O VPAS está detalhado em Laver et al. (1981). O formulário do perfil de voz é completado em cerca de 20 minutos por profissionais treinados e a amostra de voz é analisada em duas passagens. Inicialmente, na primeira passagem, as características são classificadas a *grosso modo* como neutra ou não-neutra. A categoria não-neutra é também classificada entre normal e anormal. Na segunda reprodução, os ajustes recebem notas em uma escala de 6 pontos, divididas em duas partes, 1-3 e 4-6, ajuste normal e anormal, respectivamente.

Já a escala GRBAS (Hirano, 1981) é baseada em adjetivos (dimensões) decorrentes da impressão psico-acústica e resulta da análise direta de amostras de pacientes com voz disfônica. Esta escala tem origem em uma técnica estatística chamada "Técnica Semântica Diferencial" (Snider e Osgood, 1969). Para o uso desta técnica, cria-se inicialmente um espaço perceptivo multidimensional, onde cada dimensão é definida por dois adjetivos "bipolares" com significados opostos, estabelecendo os sentidos de cada eixo. Após a definição do espaço multidimensional, vários julgadores classificam as amostras em uma escala abrangendo os eixos bipolares. Finalmente, os dados são submetidos à "Análise Fatorial" determinando o número mínimo de dimensões (quase) indepententes (com baixa correlação), que irão abranger a maior parte da variância nos julgamentos.

Nos primórdios da escale GRBAS, partiu-se de 17 pares de rótulos adjancentes (Hirano, 1981) usando uma escala de 4 pontos (0 - normal, 1 - ligeiro, 2 - significativo, 3 - extremo). Após a análise fatorial, as 17 dimensões foram reduzidas a 4:

- ***R****oughness* - rouquidão, associada a pulsos glóticos irregulares;
- ***B****reathiness* - soprosidade, relacionado ao ruído turbulento gerado na glote ou devido ao escape excessivo devido a uma fenda;
- ***A****sthenia* - astenia, relacionada à percepção de fraqueza ou falta de potência na voz;
- ***N****ormality* - normalidade, associada ao grau geral de disfônia.

Partindo destas dimensões (Hirano, 1981), renomea-se a *Normality* como *Grade* e adiciona-se uma nova dimensão, *Strain*, relacionada ao esforço excessivo na fonação. Na prática (Vieira, 1997), a dimensão ***R*** correspondem a uma mistura da aspereza e do sussuro do VPAS. ***B*** contém o sussurro do VPAS, ***S*** corresponde a tensão. Já ***G*** está correlacionada ao grau geral de distúrbio da voz.

Atributos como os citados acima são subjetivos e apesar de serem empregados clinicamente na avaliação de problemas na voz, ainda há dificuldade do estabelecimento de escalas perceptivas consistentes entre ouvintes (Kreiman et al., 1993; Kreiman e Gerratt, 1998).

As escalas objetivas vêm complementar a avaliação perceptiva da qualidade da voz e possuem a vantagem de dependerem de análises acústicas, e não da avaliação de um ouvinte.

A dificuldade com medidas objetivas tem sido a robustez na análise (Kreiman et al., 1993; Kreiman e Gerratt, 1998). Em geral, as medidas automáticas têm sua confiabilidade e precisão reduzida justamente quando as perturbações a serem medidas aumentam.

Um índice objetivo usual na avaliação da qualidade de voz é a relação sinal ruído (SNR), ou mais frequentemente, a relação harmônico ruído (HNR) (Yumoto et al., 1984). Este tipo de medida idealmente quantifica tudo que não é harmônico da fonte de excitação glótica como ruído. Numericamente, a determinação do valor do "sinal" não é trivial e os algoritmos tendem a ser influenciados por perturbações, afetando seu cálculo (Cox et al., 1989; Vieira, 1997).

O objetivo desta dissertação é estabelecer um algoritmo para o cálculo da relação harmônico-ruído que tenha maior robustez a perturbações de freqüência de excitação e amplitude do sinal.

Especialistas freqüentemente servem-se do espectrograma da voz para avaliação perceptivo-visual, tais como ruído, periodicidade, amplitude (Yanagihara, 1967). A Figura 1.1 mostra um espectrograma de uma voz disfônica, com presença de ruído e perturbações na freqüência fundamental.

Este tipo de imagem seria portanto um plataforma adequada para extração de dados, utilizando técnicas de processamento de imagens e não o tradicional processamento unidimensional, somente no tempo ou na freqüência. Aliando as características tempo-freqüência desta imagem, a similidaridade do relevo de uma impressão digital aos picos e vales da resposta harmônica da voz, aplicaram-se algoritmos de processamento de imagem, como os usados no realce de impressões digitais para identificação das regiões harmônicas do sinal de voz na imagem do espectrograma. Desta imagem, estabelece-se a relação sinal ruído espectrográfica ou S^2NR (*Spectrographic Signal Noise Ratio*), como será detalhado ao longo desta dissertação.

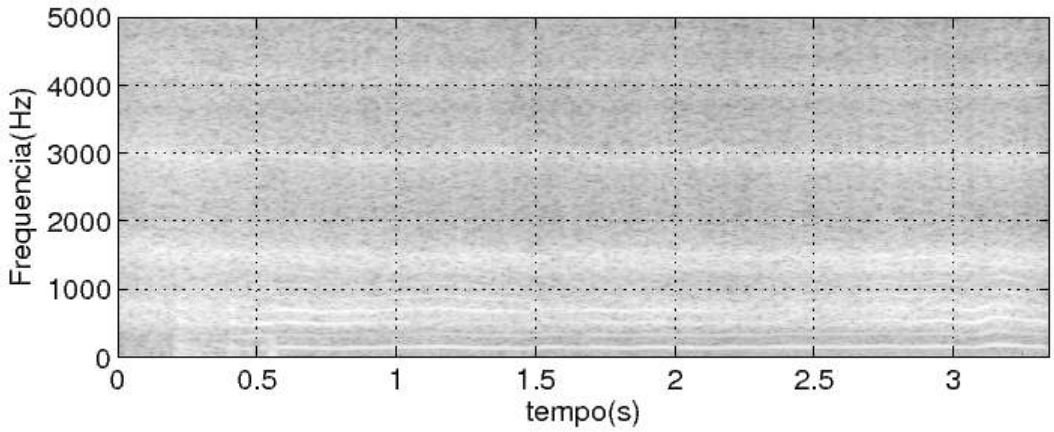

Figura 1.1: Espectrograma de Voz disfônica, na escala de tons de cinza, a parte clara indica maior intensidade. Assim permite-se uma melhor visualização de detalhes, incluindo fragmentos da estrutura harmônica, no caso de haver maior ruído em altas freqüências.

1.2 Estrutura da dissertação

Esta dissertação foi organizada em 4 capítulos. O capítulo 2 apresenta modelos de análise e síntese da voz onde é feita uma revisão do modelo de simulação das pregas vocais a duas massas e mostra-se sua variação assimétrica a quatro massas. Em seguida, investiga-se o fluxo glótico e o filtro do trato vocal através da filtragem inversa do sinal de voz e do fluxo aéreo. A partir disto, apresenta-se uma metodologia para geração de um banco de sinais de teste de voz com freqüência fundamental, perturbação em freqüência e amplitude, e ruído controlados.

No capítulo 3 é feita uma revisão das medidas de relação sinal ruído da voz mais comuns. Em seguida, descreve-se matematicamente o algoritmo de cálculo da S^2NR e sua implementação computacional.

No capítulo 4 são apresentados testes do algoritmo S^2NR sob variadas condições de relação sinal ruído, freqüência fundamental (masculino e feminino) variando-se individualmente *jitter* e *shimmer* e, em seguida, variando-se ambas perturbações simultaneamente. Compara-se o algoritmo S^2NR com o $SNR(t)$, baseado na demarcação de períodos glóticos. Aplica-se o algoritmo S^2NR à voz real, com variado grau de soprosidade individual, e compara-se o valor apurado a uma classificação subjetiva. Finalmente, mostra que o algoritmo funciona em fala encadeada.

Capítulo 2

Modelos de análise e síntese de sinais de voz

A fala é o principal método de comunicação do ser humano (Flanagan, 1972). Ela é o produto da ação voluntária de um complexo aparato, o aparelho fonatório, e é um campo extremamente ativo de pesquisa.

Para a boa compreensão do funcionamento deste sistema, buscou-se a construção de modelos que simulam o funcionamento das estruturas responsáveis pela produção da fala. A abordagem fonte-filtro (Fant, 1960), por exemplo, define a fonte como um fluxo de ar gerado pela vibração das pregas vocais, enquanto o filtro é composto pelas cavidades supraglotais: oral, faríngea e nasal. A configuração deste filtro, ou trato vocal, apresenta uma determinada função de transferência.

Neste capítulo faz-se uma revisão de alguns modelos usados no estudo da fonte e do filtro vocal. As seções são ilustradas com resultados de simulações computacionais realizadas pelo autor.

2.1 Modelos de Fonte

2.1.1 Modelo a duas massas

Um modelo freqüentemente utilizado para a fonte é o modelo a duas massas, descrito em Ishizaka e Flanagan (1972). Neste tipo de modelagem são atribuídos massa, elasticidade e amortecimento para os componentes do sistema fonatório. Dada a natureza distribuída das estruturas, é feita uma aproximação de primeira ordem onde a prega vocal é representada por dois ressonadores mecânicos acoplados por uma mola, conforme mostrado na Figura 2.1, sendo as pregas vocais admitidas bilateralmente simétricas. Estas estruturas são submetidas à pressão subglótica, que atua como força de excitação do sistema.

Neste modelo, as massas m_1 e m_2 têm deslocamento lateral x_1 e x_2, e estas massas são acopladas por uma mola com coeficiente k_c.

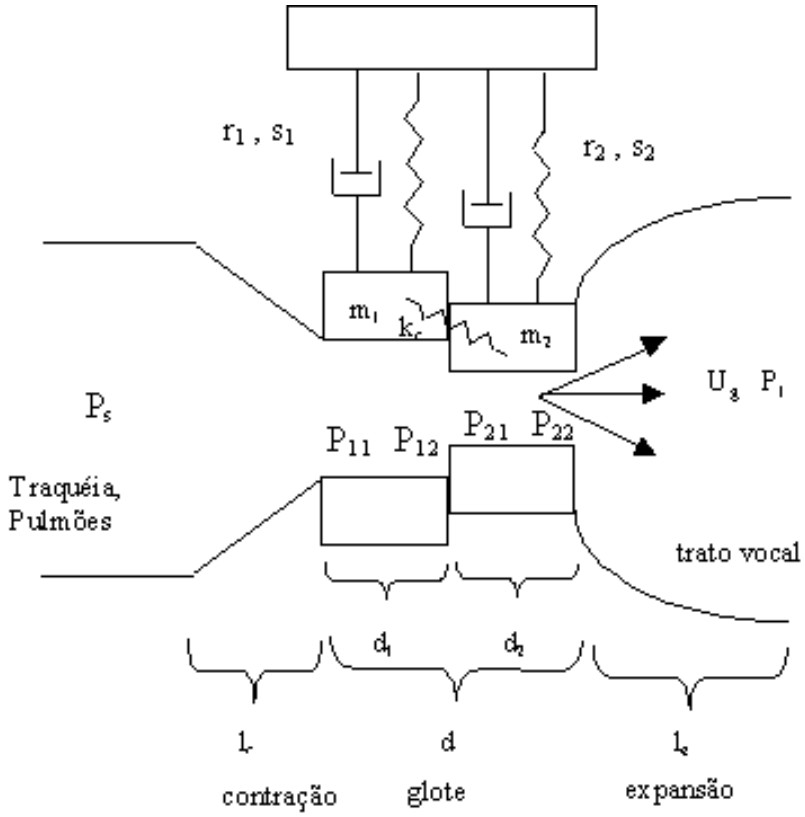

Figura 2.1: Diagrama esquemático do sistema a duas massas. O movimento das massas se dá na vertical, sendo x_1 e x_2, os deslocamentos de m_1 e m_2.

Os parâmetros do modelo são o comprimento glótico efetivo l_g (que compreende o comprimento de expansão, l_e, o comprimento de contração, l_c, e a espessura das massas, d), a espessura das massas d_1 e d_2, as constantes elásticas equivalentes s_1 e s_2, e coeficiente de viscosidade r_1 e r_2. A constante k_c representa um efeito da mudança na flexibilidade na direção lateral das pregas. Esta resulta da variação em espessura e flexibilidade das pregas por ação do músculo tireoaritenóideo. Os coeficientes s_1 e s_2 são uma representação da tensão submetida às pregas vocais. Elas expressam uma variação não-linear, medida experimentalmente, dos tecidos humanos. Esta relação entre a deflexão do ponto de equilíbrio e força necessária para tal é dada por:

$$f_{sj} = k_j x_j (1 + \eta_{kj} x_j^2), \quad j = 1,\, 2,$$

onde f_{sj} é força necessária para produzir x_j, k_j é o coeficiente linear de s_j, e η_{kj} é o coeficiente que descreve a não-linearidade de s_j.

No caso de colisão de m_1 e m_2 com as partes correspondentes, modifica-se a expressão, já que a colisão causa deformação das pregas vocais. A força restauradora pode ser representada pela mola equivalente s_{hj}, $(j = 1, 2)$. Admitindo-se que para $x_j + A_{g0j}/2l_g \leq 0 \quad (j = 1, 2)$:

$$f_{hj} = h_j \left(x_j + \frac{A_{g0j}}{2l_g} \right) \left[1 + \eta_{hj} \left(x_j + \frac{A_{g0j}}{2l_g} \right)^2 \right],$$

onde f_{hj} é a força necessária para produzir uma deformação em m_j durante uma colisão, h_j é o coeficiente linear e η_{hj} representa a não-linearidade das pregas vocais durante o contato. A_{g0j} é a área inicial da glote, no trecho da prega j.

A força resultante, que surge durante o fechamento, é a soma de f_{sj} e f_{hj}.

A perda viscosa será admitida como linear por partes. É conveniente tratar as resistências viscosas em termos de coeficientes de amortecimento: ζ_1 e ζ_2 para os osciladores não acoplados, onde:

$$r_1 = 2\zeta_1\sqrt{m_1 k_1} \quad e \quad r_2 = 2\zeta_2\sqrt{m_2 k_2}. \tag{2.1}$$

Os valores utilizados são $\zeta_1 = 0,1$ e $\zeta_2 = 0,6$ enquanto a glote está aberta. Para a glote fechada, toma-se um valor que tornará o amortecimento crítico, a saber: $\zeta_1 = (1,0 + 0,1)$ e $\zeta_2 = (1,0 + 0,6)$.

A distribuição de pressão ao longo da glote é expressa da seguinte forma:

$$\begin{aligned}
P_s - P_{11} &= 1,37\frac{\rho}{2}\left(\frac{U_g}{A_{g1}}\right)^2 + \frac{dU_g}{dt}\int_0^{l_c}\frac{\rho}{A_c(x)}dx, \\
P_{11} - P_{12} &= 12\frac{\mu l_g^2 d_1}{A_{g1}^3}.U_g + \frac{\rho d_1}{A_{g1}}.\frac{dU_g}{dt}, \\
P_{12} - P_{21} &= \frac{\rho}{2}U_g^2\left(\frac{1}{A_{g2}^2} - \frac{1}{A_{g1}^2}\right), \\
P_{21} - P_{22} &= 12\frac{\mu l_g^2 d_2}{A_{g2}^3}.U_g + \frac{\rho d_2}{A_{g2}}.\frac{dU_g}{dt}, \\
P_{22} - P_1 &= -\frac{\rho}{2}\left(\frac{U_g}{A_{g2}}\right)^2 2\frac{A_{g2}}{A_1}\left(1 - \frac{A_{g2}}{A_1}\right),
\end{aligned} \tag{2.2}$$

sendo ρ a densidade do ar, U_g o fluxo glótico volumar, A_{gj}, a área da glote na seção correspondente a massa j e A_c a área na região da contração.

Com base nas relações dadas em (2.2), os elementos de impedância acústica do orifício glotal formam o circuito análogo elétrico da Figura 2.2 para a glote, onde os fluxos são tratados como corrente elétrica e a pressão como tensão. Os elementos do circuito estão em série, e são dados por:

$$R_c = 1,37\frac{\rho}{2}\frac{|U_g|}{A_{g1}^2}, \quad L_c = \int_0^{l_c}\frac{\rho}{A_c(x)}dx,$$

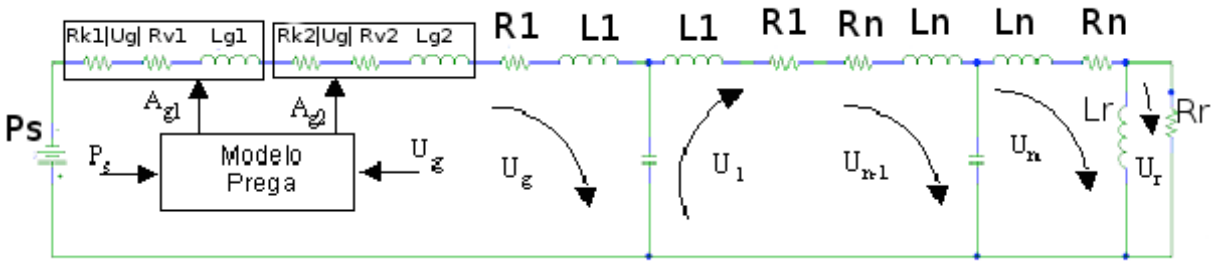

Figura 2.2: Esquema do circuito análogo elétrico do modelo a duas massas, incluindo o trato vocal e carga de irradiação.

$$
\begin{aligned}
R_{vi} = 12\frac{\mu l_g^2 d_i}{A_{gi}^3}, \quad L_{gi} = \frac{\rho d_i}{A_{gi}} \quad \text{com} \quad i = 1, 2, \\
R_e = -\frac{\rho}{2}\frac{2}{A_{g2}A_1}\left(1 - \frac{A_{g1}}{A_1}\right)|U_g|, \quad R_{12} = \rho\left(\frac{1}{A_{g2}^2} - \frac{1}{A_{g1}^2}\right)|U_g|, \\
R_{k1} = \frac{0,19\rho}{A_{g1}^2}, \quad R_{k2} = \frac{\rho\left[0,5 - \frac{A_{g2}}{A_1}\left(1 - \frac{A_{g2}}{A_1}\right)\right]}{A_{g2}^2}.
\end{aligned} \tag{2.3}
$$

As equações do movimento para as pregas vocais podem ser escritas da seguinte maneira:

$$
m_1\frac{d^2x_1}{dt^2} + r_1\frac{dx_1}{dt} + s_1(x_1) + k_c(x_1 - x_2) = F_1, \tag{2.4}
$$

$$
m_2\frac{d^2x_2}{dt^2} + r_2\frac{dx_2}{dt} + s_2(x_2) + k_c(x_2 - x_1) = F_2, \tag{2.5}
$$

sendo

$$
\begin{aligned}
A_{gi} &= (A_{g0i} + 2l_g x_i) \quad i = 1, 2, \\
s_i(x_i) &= k_i(x_i + \eta_{ki}x_i^3) \text{ se } \quad x_i > -\frac{A_{g0i}}{2l_g}, \\
s_i(x_i) &= k_i(x_i + \eta_{ki}x_i^3) + h_i\left\{\left(x_i + \frac{A_{g0i}}{2l_g}\right) + \eta_{ki}\left(x_i + \frac{A_{g0i}}{2l_g}\right)^3\right\} \text{ se } x_i \leq -\frac{A_{g0i}}{2l_g}.
\end{aligned} \tag{2.6}
$$

A força atuante sobre as massas é dada pela Tabela 2.1, que trata as condições de colisão:

Tabela 2.1: Força atuante sobre as massas, com tratamento de colisão:

x_1	x_2	F_1	F_2
$x_1 > x_{1_{min}}$	$x_2 > x_{2_{min}}$	$P_{m1}l_g d_1$	$P_{m2}l_g d_2$
$x_1 \leq x_{1_{min}}$	$x_2 > x_{2_{min}}$	$P_s l_g d_1$	0
$x_1 > x_{1_{min}}$	$x_2 \leq x_{2_{min}}$	$P_s l_g d_1$	$P_s l_g d_2$
$x_1 \leq x_{1_{min}}$	$x_2 \leq x_{2_{min}}$	$P_s l_g d_1$	0

Os termos de pressão P_{m1} e P_{m2} são obtidos através das relações estabelecidas em (2.2), são detalhados abaixo:

$$P_{m1} = \frac{1}{2}(P_{11} + P_{12}) = P_s - 1.37\frac{\rho}{2}\left(\frac{U_g}{A_{g1}}\right)^2 - \frac{1}{2}\left(R_{v1}U_g + L_{g1}\frac{dU_g}{dt}\right), \tag{2.7}$$

$$P_{m2} = \frac{1}{2}(P_{21} + P_{22}) = P_{m1} - \frac{1}{2}\left\{(R_{v1} + R_{v2})U_g + (L_{g1} + L_{g2})\frac{dU_g}{dt}\right\} - \frac{\rho}{2}U_g^2\left(\frac{1}{A_{g2}^2} - \frac{1}{A_{g1}^2}\right). \tag{2.8}$$

Com a descrição do sistema correspondente a fonte, passamos à descrição do sistema correspondente ao filtro, o trato vocal (Fant, 1960).

Nesta etapa descreve-se a representação do trato vocal por um sistema de parâmetros distribuídos, cujo análogo elétrico é uma linha de transmissão. Para tanto, são mantidas as analogias fluxo-corrente e pressão-tensão. Aproximando o trato vocal por um sistema formado por n seções de paredes rígidas, cujos elementos acústicos são dependentes de seu comprimento l_i e área da seção A_i (Fant, 1960; Ishizaka e Flanagan, 1972), os componentes do circuito elétrico análogo de cada seção i são:

1. A indutância é : $L_i = \rho l_i / 2A_i$.
2. A capacitância é: $C_i = (l_i A_i / \rho c^2)$, sendo c a velocidade do som.
3. A resistência é $R_i = (S_i/A_i)\sqrt{\rho\mu\omega/2}$, onde S_i é a circunferência da i-ésima seção, ω a freqüência em radianos. Para efeitos de simulação, utiliza-se $\omega = \sqrt{k_1/m_1}$, que corresponde à freqüência natural de oscilação da massa m_1.

A terminação do trato é feita nos lábios. Como modelo, utiliza-se uma carga que consiste em uma resistência em paralelo com uma indutância, dadas pelas relações: $L_R = 8\rho/3\pi\sqrt{\pi A_n}$ e $R_R = (128\rho c/9\pi^2 A_n)$. Isto corresponde a uma abertura (boca) num "grande" plano refletor (face) (Flanagan, 1972).

Um esquema para o circuito descrito está mostrado na Figura 2.2.

Diferentes configurações de tratos permitem a produção de diferentes vogais. A extração das áreas, originalmente feita por raio-X (Fant, 1960), hoje em dia é feita por ressonância magnética (Story et al., 1996). Inicialmente, nesta discussão, o trato vocal estará restrito a um tubo uniforme de comprimento $l = 16\,cm$ e seção transversal de $5\,cm^2$.

Na Figura 2.3, tem-se o resultado da simulação utilizando os parâmetros das tabelas 2.2, 2.1.1 e 2.1.1.

A freqüência de oscilação determinada é de $165,75\,Hz$. A Figura 2.3 mostra três gráficos. O primeiro representa o fluxo glótico U_g obtido, o segundo é a posição das pregas ao longo do tempo, sendo a linha contínua a posição de m_1 e a linha tracejada a posição de m_2. Por

Tabela 2.2: Parâmetros de simulação modelo a duas massas:

Parâmetro	**Símbolo**	**Valor**
densidade do ar	ρ	$1,14 \times 10^{-3}\, g/cm^3$
coeficiente de viscosidade	μ	$1,8610^{-4}\, dyn - s/cm^2$
pressão subglótica	P_s	$8,0\, cmH_2O$
velocidade do som (ar, $37°C$)	c	$3,5 \times 10^4\, cm/s$
comprimento da glote	L_g	$1,40\, cm$
coeficiente de acoplamento	k_c	$25 \times 10^3\, dyn/cm$

Tabela 2.3: Parâmetros para m_1:

Parâmetro	**Valor**
d_1	$0,25\, cm$
h_1	$240\, kdyn/cm$
η_{h_1}	500
k_1	$80\, kdyn/cm$
η_{k_1}	100
m_1	$0,125\, g$
ζ_1	$0,1$
$\zeta_{fechado}$	$1 + 0,1$
Ag_{01}	$0,05\, cm^2$

Tabela 2.4: Parâmetros da massa m_2:

Parâmetro	**Valor**
d_2	$0,05\, cm$
h_2	$24\, kdyn/cm$
η_{h_2}	500
k_2	$8\, kdyn/cm$
η_{k_2}	100
m_2	$0,025\, g$
ζ	$0,6$
$\zeta_{fechado}$	$1 + 0,6$
Ag_{02}	$0,05\, cm^2$

fim, o último gráfico representa o fluxo U_r, que passa pelo ramo R_r da terminação labial, conforme a Figura 2.2, e se relaciona com a pressão irradiada pela expressão $P_r = R_r U_r$. Valores negativos para x_1 e x_2 indicam fechamento da glote, justificado pela continuidade do movimento do centro de massa, mesmo quando a estrutura encontra-se fechada.

Os valores de diferença de fase entre as massas e razão de abertura da glote são consistentes com os valores encontrados experimentalmente através de técnicas de visualização em alta velocidade das pregas vocais e com a filtragem inversa (que será detalhada na Seção 2.2).

A forma de onda do fluxo glotal, U_g, possui uma assimetria característica, isto é, uma diminuição mais abrupta que o aumento do fluxo. Isto se deve à maior velocidade da fase de fechamento em relação à fase de abertura do ciclo glótico. Já o gráfico da posição de x_1 e x_2

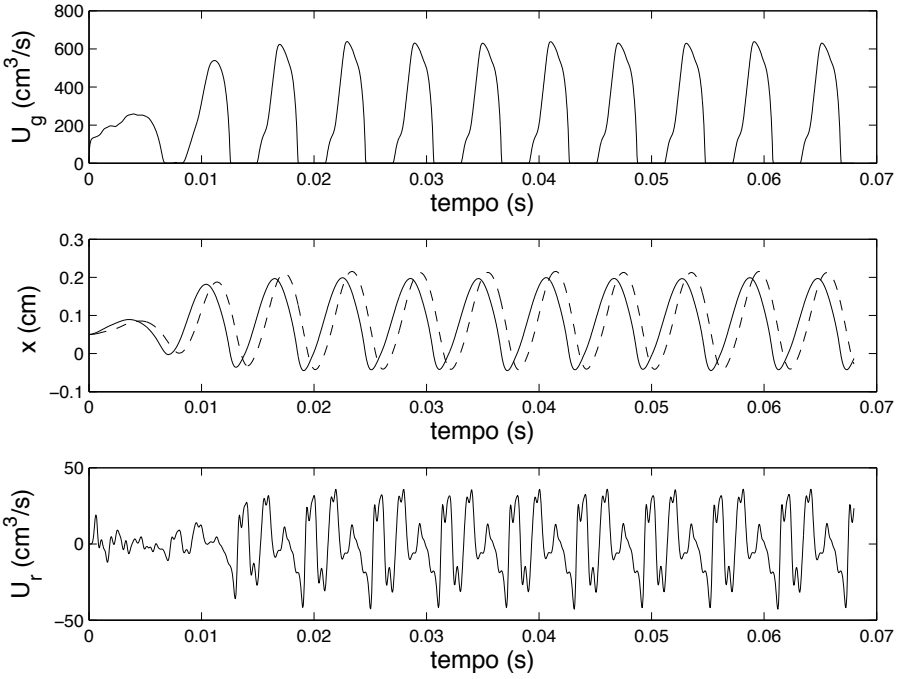

Figura 2.3: Simulação do modelo a duas massas, trato vocal uniforme, $l = 16\,cm$, $A = 5\,cm^2$.

é mais suave e simétrico, não reproduzindo detalhes da interação acústica com o fluxo.

Os modos normais de um tubo uniforme semi-aberto de comprimento l estão em $f_n = (2n + 1)l/4c$, com $n = 0, 1, 2, ...$ Estas freqüências de ressonância têm aproximadamente o valor de $f_1 = 500\,Hz$, $f_2 = 1500\,Hz$, $f_3 = 2500\,Hz$, e assim em diante. As manifestações destes picos no espectro do sinal de saída são chamadas de formantes. Na Figura 2.4, tem-se a densidade espectral de potência, calculada pelo método de Welch (Marple, 1987) (janela de FFT[1] de 1024 amostras com sobreposição de 512), da pressão sonora irradiada. Nota-se nesta figura a presença dos harmônicos da freqüência fundamental, $f_0 = 165,75\,Hz$, após serem submetidos à função de transferência do trato vocal, onde são enfatizadas às freqüências próximas as ressonâncias do trato. A linha tracejada indica a densidade espectral de potência calculada por LPC [2] através do método da co-variância (com $M = 20$). Através dela, pode-se perceber três picos claros, indicativos dos formantes, o primeiro em $F_1 = 500\,Hz$, o segundo em $F_2 = 1500\,Hz$, e o terceiro em $F_3 = 2300\,Hz$.

O modelo a duas massas, tal como descrito e dentro de seus limites de estabilidade não apresenta perturbações, apresenta freqüência de vibração regular e amplitude constante na existência de uma fonte de pressão constante. Desta maneira, a voz gerada soa artificial, já que não apresenta a irregularidade e a variabilidade existentes na voz natural.

Modelos com múltiplas massas, por possuírem um maior grau de liberdade, podem gerar voz sintética com maior naturalidade que um modelo mais simples como o modelo a duas massas (Titze, 1973; Titze e Strong, 1975; Wong et al., 1991).

[1] *Fast Fourier Transform*, Transformada Rápida de Fourier
[2] Linear Predictive Coding - Codificação Linear Preditiva (Markel e Gray Jr, 1976)

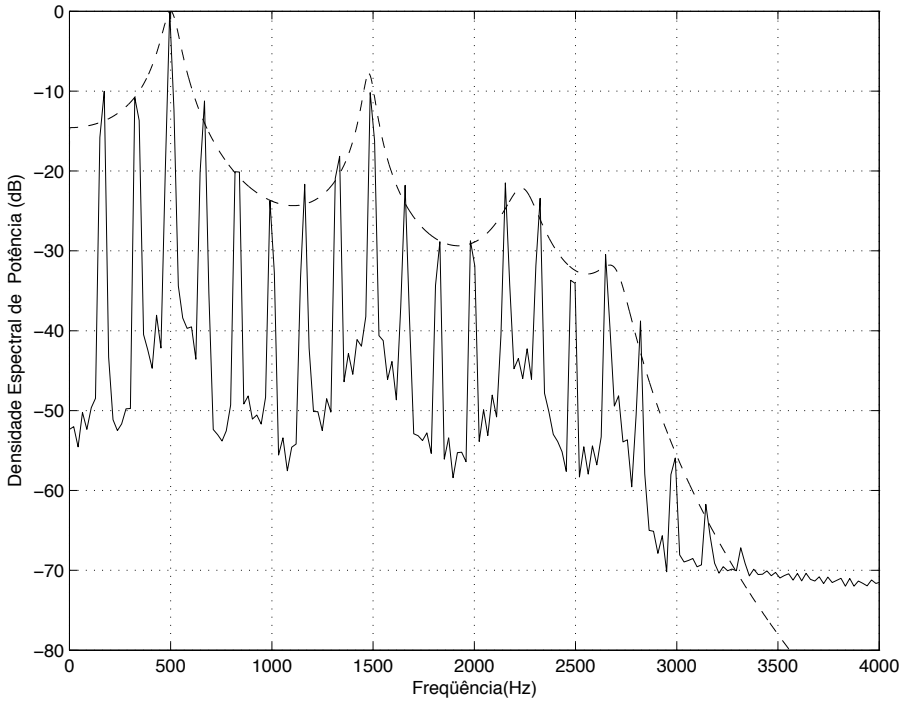

Figura 2.4: Densidade espectral de potência para pressão sonora obtida na saída do trato vocal uniforme, $l = 16\,cm$, $A = 5\,cm^2$.

Toma-se como base o modelo a duas massas para a construção de um modelo a quatro massas, sendo assim possível a introdução de assimetrias entre as pregas vogais.

Em condições simétricas, o modelo a quatro massas simétrico tem comportamento equivalente ao de duas massas, conforme mostra a Figura 2.5.

As assimetrias no sistema introduzem perturbações na forma de onda. Assimetrias de massa e elasticidade alteram as freqüências naturais de oscilação do sistema, podendo causar perturbações na freqüência do fluxo aéreo e, por conseqüência, na voz gerada. No caso mostrado na Figura 2.6, temos um parâmetro Q que escalona inversamente a massa e a espessura das pregas vocais, enquanto escalona diretamente a rigidez, de modo a variar linearmente a freqüência de oscilação. Neste caso, escalona-se o lado A por $Q_A = 1,4$ e o lado B por $Q_B = 0,8$. Nota-se, em relação ao caso simétrico anterior: diminuição na freqüência fundamental, menor amplitude de oscilação do lado A em relação ao lado B e diminuição da amplitude do fluxo glótico.

Perturbações também podem ser inseridas através de componentes aleatórias no modelo. Um alternativa possível é a inclusão de uma parcela aleatória no termo da pressão subglótica. Este tipo de abordagem tipicamente influencia na amplitude de cada ciclo da oscilação (*shimmer*), conforme mostra o painel inferior da Figura 2.7.

Apesar de reproduzirem diversos fenômenos através de variações paramétricas, estes modelos têm aplicação limitada quando é necessário o controle preciso das perturbações a serem simuladas. Por exemplo, ao alterar a pressão subglótica altera-se necessariamente a

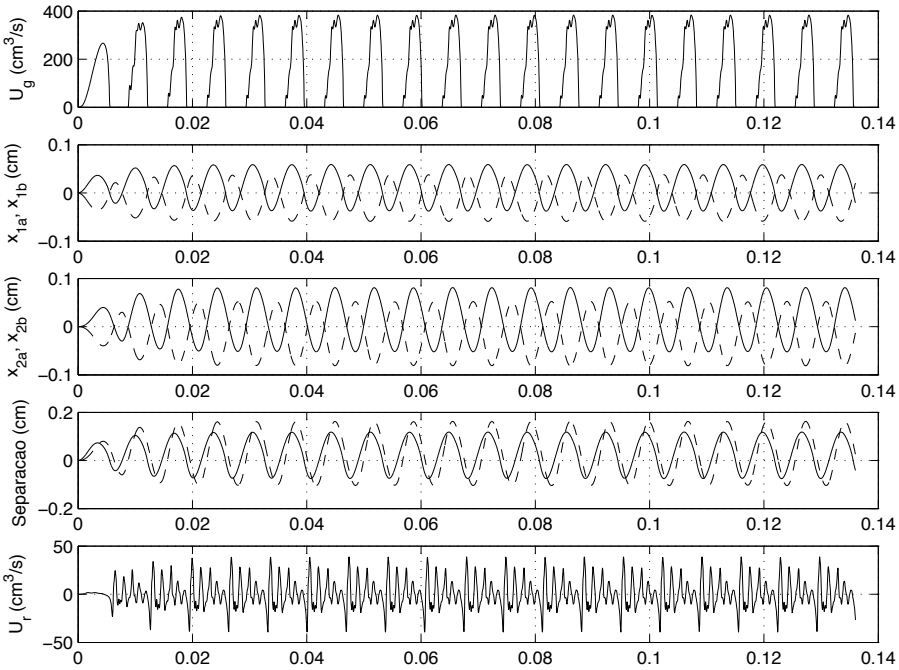

Figura 2.5: Simulação do modelo a 4 massas simétrico, $l = 17\,cm$, vogal /a/. O eixo horizontal indica o tempo em segundos.

freqüência de oscilação do sistema. No restante deste capítulo, estabeleceremos um modelo onde será possível o controle mais preciso da forma de onda para se ter um sinal de referência de voz.

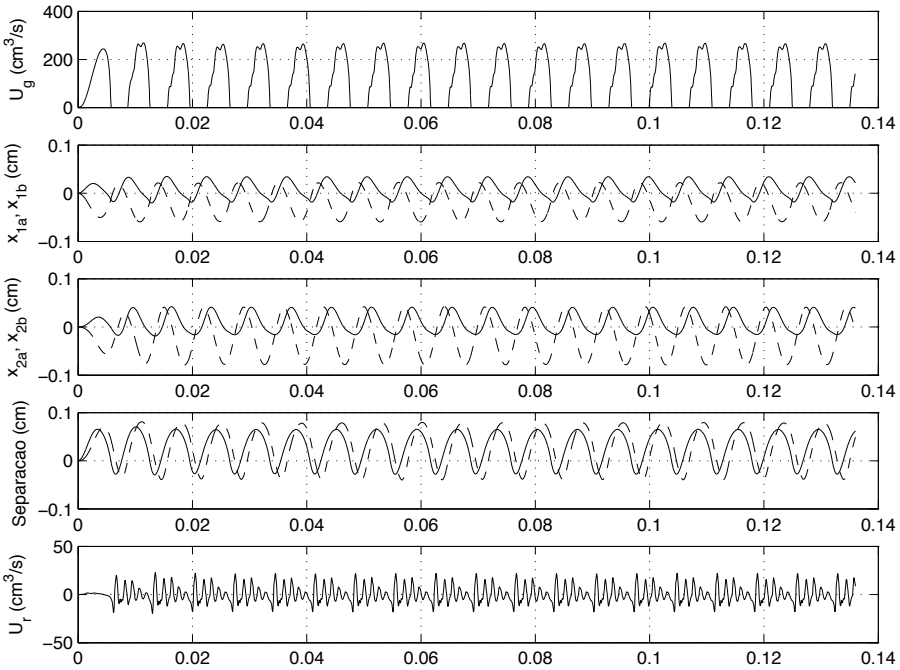

Figura 2.6: Simulação do modelo a 4 massas assimétrico, $l = 17\,cm$, vogal /a/, $Q_A = 1,4$ e $Q_B = 0,8$. O eixo horizontal indica o tempo em segundos.

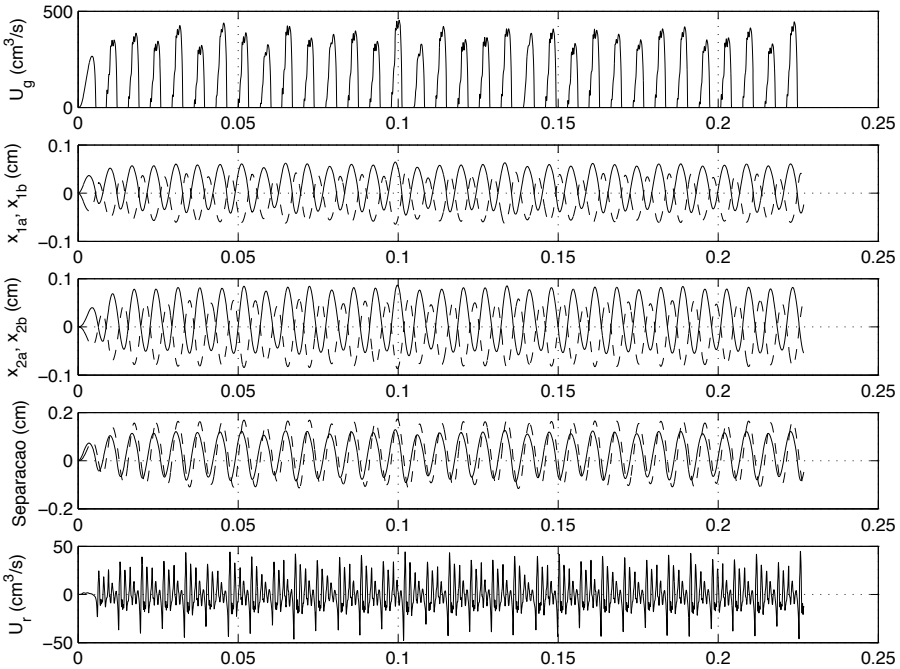

Figura 2.7: Simulação do modelo a 4 massas simétrico, ruído adicionado a P_s, $l = 17\,cm$, vogal /a/, $Q_A = 1,0$ e $Q_B = 1,0$. O eixo horizontal indica o tempo em segundos.

2.1.2 Modelos paramétricos para o fluxo glótico

Matematicamente, existem três tipos básicos de métodos para excitação da glote. Estes são (1) excitação por impulso seguido de um filtro com simulação da resposta glótica; (2) formas de onda do fluxo glótico obtidas de filtragem inversa ou forma de onda da área glótica, conforme mostrado no item anterior; e ainda (3) modelos de excitação de forma de onda, como o modelo de Rosenberg e o modelo *LF* (Fant et al., 1985). Por permitirem ajuste paramétrico das características da forma de onda do pulso glótico, modelos de forma de onda são mais flexíveis e simples de se utilizar para geração de som natural.

A excitação do trato vocal pode ser modelada por um trem periódico de pulsos de fluxo de ar. Desta forma, o espectro consiste de harmônicos da freqüência fundamental e sua amplitude é determinada pelo envelope da magnitude do espectro do pulso.

Como visto na seção anterior, o pulso do fluxo de ar tem uma inclinação para a direita, porque a fase de fechamento é mais rápida que a fase de abertura da glote. Fant (1979) modelou o pulso utilizando o *cosseno levantado*:

$$\text{Fase de Abertura: } q(t) = q_{max}\frac{1}{2}\left[1 - \cos\frac{\pi(t - T_1)}{T_2}\right], \quad T_1 \leq t \leq T_2, \tag{2.9}$$

$$\text{Fase de Fechamento: } q(t) = q_{max}\left[K\cos\frac{\pi(t - T_2)}{T_2} - K + 1\right], \quad T_2 < t \leq T_3, \tag{2.10}$$

onde q é o fluxo volumar glótico (m^3/s), t é tempo, q_{max} é o valor de pico do fluxo, K é fator de inclinação ($K = 0, 5, \cdots, \infty$), T_1, T_2 e T_3 os parâmetros a serem definidos. Neste modelo, a energia espectral está praticamente concentrada em torno do chamado *formante glótico* $F_g = [2(T_2 - T_1)]^{-1}$, praticamente independente do fator K, que determina o decaimento do espectro. Variando-se K entre $0,51$ e $4,0$ temos decaimentos que vão de $-18\,dB/oitava$ a $-12\,dB/oitava$, correspondendo aos ajustes modal e falsete, respectivamente.

Este modelo, pela simplicidade e flexibilidade dos parâmetros, será usado na síntese das vogais.

2.2 Modelos do Filtro

Na seção anterior, utilizou-se um modelo de linha de transmissão para a caracterização do trato vocal. Os dados utilizados foram obtidos através de medições de imagem do raio-X (Fant, 1960). Estes parâmetros podem ser aproximados através de métodos de inversão a partir do sinal da fala, conforme Rothenberg (1973); Wong et al. (1979). Este processo permite também a obtenção do fluxo glótico, ou seja da fonte.

2.2.1 Modelagem matemática

Pela teoria de Fonte-Filtro citada anteriormente, e tratando as equações no tempo discreto, com suas respectivas transformadas Z, $E(z) \leftrightarrow e(n)$ é o modelo de excitação glotal, $U_G(z) \leftrightarrow u_G(n)$, fluxo glótico volumar, $U_L(z) \leftrightarrow u_L(n)$ é o fluxo volumar labial e $S(z) \leftrightarrow s(n)$ é pressão sonora irradiada da fala. O modelo $e(n)$ serve como entrada para o modelo do filtro glotal $G(z)$, que irá gerar o sinal $U_G(z)$. Normalmente, $e(n)$ é tomado como um trem periódico de impulsos.

O modelo do trato vocal $V(z)$ é assumido como um modelo somente de pólos, da forma:

$$V(z) = \left[1 + \sum_{i=1}^{K} c_i z^{-i}\right]^{-1}, \tag{2.11}$$

sendo K inteiro par e c_i os coeficientes ao filtro correspondente ao trato vocal com K pólos, pode haver até $K/2$ formantes em voz, caso todos os pólos sejam complexos conjugados.

A pressão sonora está relacionada com o fluxo volumar labial por uma derivada. Em baixas freqüências, modelamos, a menos de uma constante de proporcionalidade:

$$R(z) = \frac{S(z)}{U_L(z)} = 1 - z^{-1}. \tag{2.12}$$

Baseado no modelo acima exposto, a expressão para a transformada do fluxo glotal é dada por:

$$U_G(z) = \frac{S(z)}{V(z)R(z)}. \tag{2.13}$$

A equação depende do modelo $V(z)$, que será estimado através da análise LPC pelo método de co-variância.

Definindo uma função de excitação efetiva $Q(z) \leftrightarrow q(n)$, por:

$$Q(z) = U_G(z)R(z) \leftrightarrow q(n) = u_G(n) * r(n). \tag{2.14}$$

Devido à característica de passa-altas de $r(n)$, o sinal $q(n)$ deve ter média nula. Admitindo condições de fechamento estável para a glote, os pontos $n = L_c$ e $n = L_o$ são, respectivamente, os pontos de abertura e fechamento. Se $u_G(n) = 0$ para $L_c \leq n < L_o$, então $q(n) = 0$ para $L_c + 1 \leq n < L_o$.

Considerando o modelo de predição linear, tem-se :

$$s(n) = \sum_{i=1}^{K} c_i s(n-i) + q(n). \tag{2.15}$$

No instante de fechamento da glote, tem-se $q(n) = 0$, logo:

$$s(n) = -\sum_{i=1}^{K} c_i s(n-i) \quad (L_c + 1 \leq n < L_o). \tag{2.16}$$

Uma amostra após o fechamento glotal, a forma de onda $s(n)$ se torna uma oscilação com decaimento livre (soma de exponenciais complexas) que é uma função das resonâncias especificadas pelos coeficientes c_i e as condicões iniciais $s(L_c), \cdots, s(L_c - K + 1)$.

Admite-se o filtro de ordem M na forma:

$$A(z) = \sum_{i=0}^{M} a_i z^{-1} \quad (a_o = 1). \tag{2.17}$$

onde $M \geq K$ é para ser obtido. Se $s(n)$ é passado através deste filtro, a saída é definida pelo resíduo ou sinal de erro, dado por:

$$\epsilon(n) = s(n) + \sum_{i=1}^{M} a_i s(n-i). \tag{2.18}$$

No intervalo onde ocorre o fechamento da glote, $L_c + 1 \leq n < L_o$, aplicando a expressão (2.16) a (2.18), tem-se:

$$\epsilon(n) = \sum_{i=1}^{M} (a_i - c_i) s(n-i). \tag{2.19}$$

Se as condições:

$$a_i = \begin{cases} c_i & \text{se} \quad i = 1, \cdots, K, \\ 0 & \text{se} \quad i = K+1, \cdots, M, \end{cases} \tag{2.20}$$

são satisfeitas, então $\epsilon(n) = 0$ para $L_c + 1 \leq n < L_o$. Obtendo A_z pelo método da co-variância, conforme Markel e Gray Jr (1976), para um janela de análise de $s(n-M)$ a $s(n+N-M-1)$, o erro quadrático $\alpha_M(n)$ é calculado como:

$$\alpha_M(n) = \sum_{j=n}^{n+N-M-1} \epsilon^2(j). \tag{2.21}$$

Os coeficientes do filtro são obtidos pela minimização de $\alpha_M(n)$. Teoricamente, este valor é nulo no período de glote fechada. Na prática, este valor informa a erro do modelo durante este período.

2.2.2 Aplicação do método

Nesta seção, aplica-se o método da filtragem inversa a sinais de voz sintéticos e reais. O primeiro exemplo mostrado é a inversão do sinal gerado pelo modelo a 4 massas com perturbação em amplitude (Figura 2.7). Nas Figuras 2.8 e 2.9 temos o fluxo glótico recuperado, o sinal acústico original e a resposta em freqüência do trato no trecho considerado. O critério para a determinação do trecho para o cálculo do filtro é a minimização da ondulação na fase fechada. Nota-se que não é possível a recuperação da componente contínua do sinal do fluxo glótico e este é apresentado em unidades normalizadas. A ordem do filtro é $M = 14$, de modo

a detectar 7 formantes (o filtro original possuía 7 seções cilíndricas).

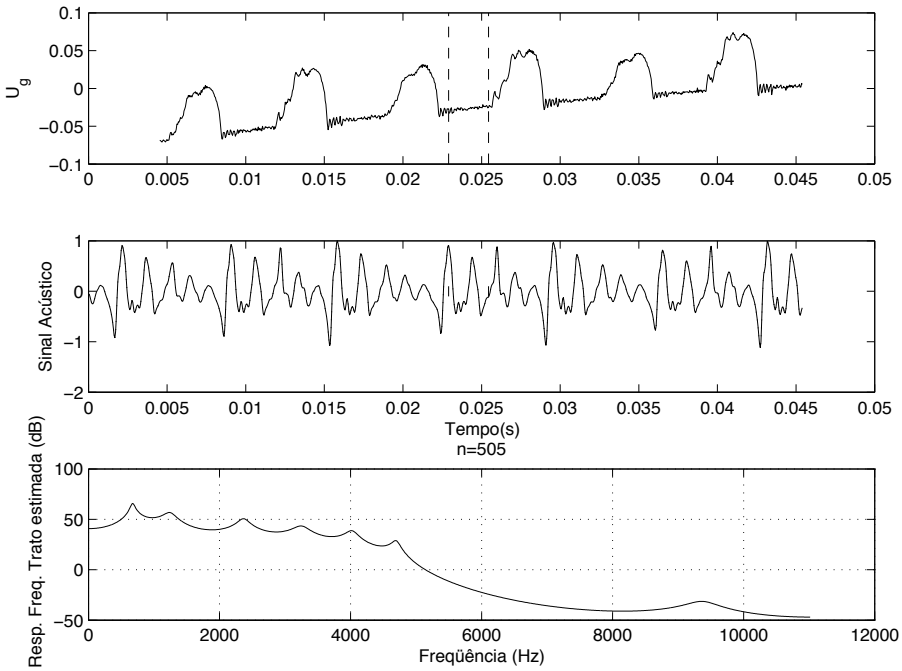

Figura 2.8: Filtragem Inversa de voz sintetizada pelo modelo a 4 massas, vogal /a/. O par de linhas tracejadas indica o trecho utilizado para a filtragem inversa.

Para a voz real, repetiu-se o procedimento em uma gravação de voz feminina, amostrada a $22050\,Hz$, retirada da base de dados coletada na *Royal Infirmary of Edinburgh* (Vieira, 1997). Utilizando a mesma ordem do filtro, notam-se nas Figuras 2.10 e 2.11 as seguintes características: fase fechada mais curta devido à maior freqüência fundamental, o que implica em maior dificuldade de tomar um trecho de tamanho $5M$ que comporte apenas a fase fechada. Neste caso também, percebe-se uma maior simetria da fase de abertura com a fase de fechamento, que diminui o conteúdo harmônico do sinal de voz, fato que pode ser constatado ao comparar o sinal acústico com o exemplo anterior, da Figura 2.9.

A Figura 2.12 mostra a aplicação do método à voz masculina, amostrada como no exemplo anterior. Neste caso, nota-se uma fase de fechamento bastante abrupta, além de uma fase fechada mais longa que a fase aberta.

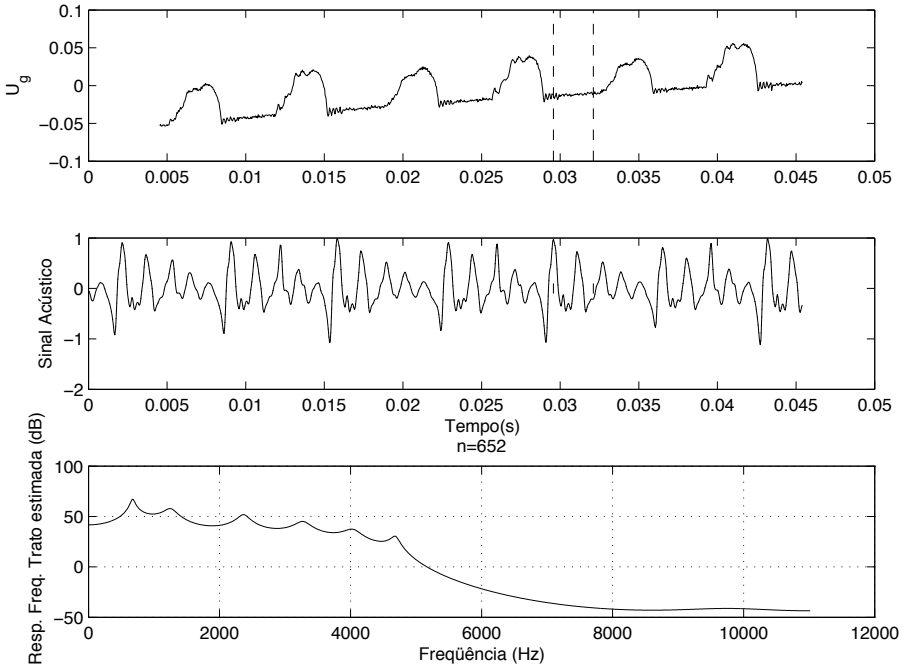

Figura 2.9: Filtragem Inversa de voz sintetizada pelo modelo a 4 massas, vogal /a/. O par de linhas tracejadas indica o trecho utilizado para a filtragem inversa, processando-se a mesma gravação anterior, em diferente trecho.

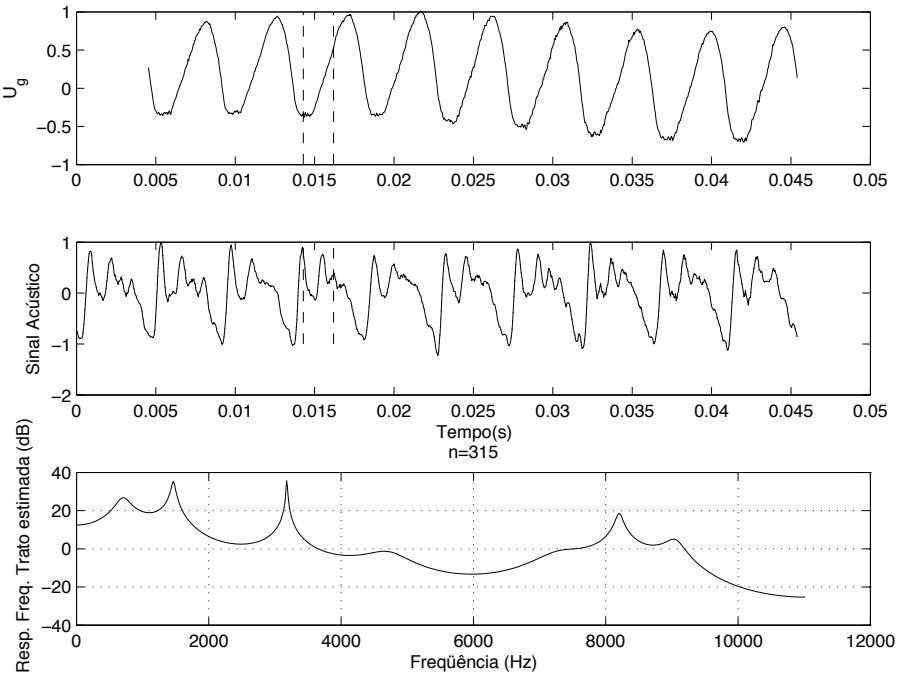

Figura 2.10: Filtragem Inversa de voz real feminina, vogal /a/.

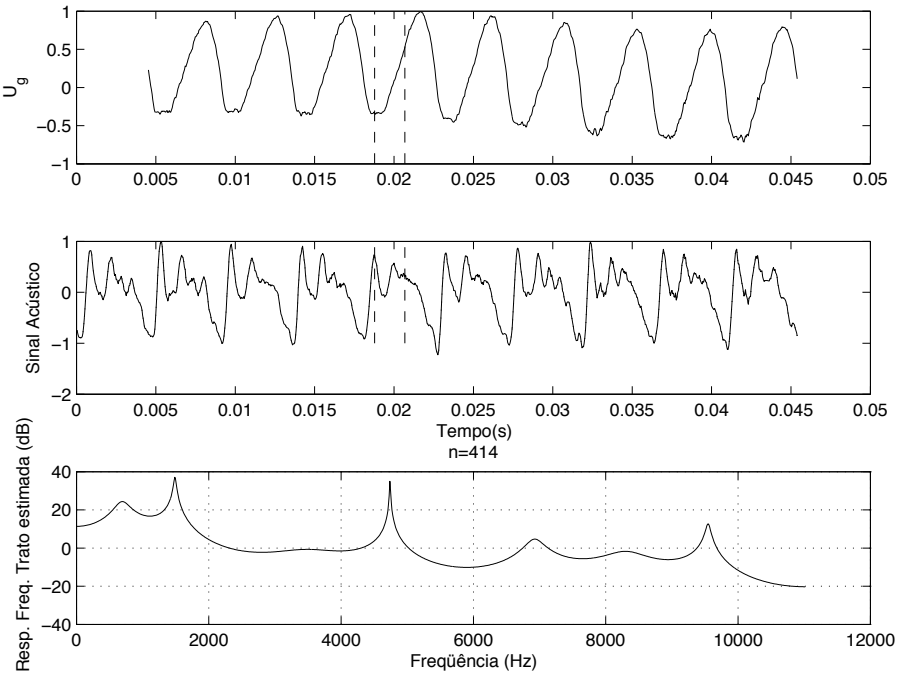

Figura 2.11: Filtragem Inversa de voz real feminina, vogal /a/.

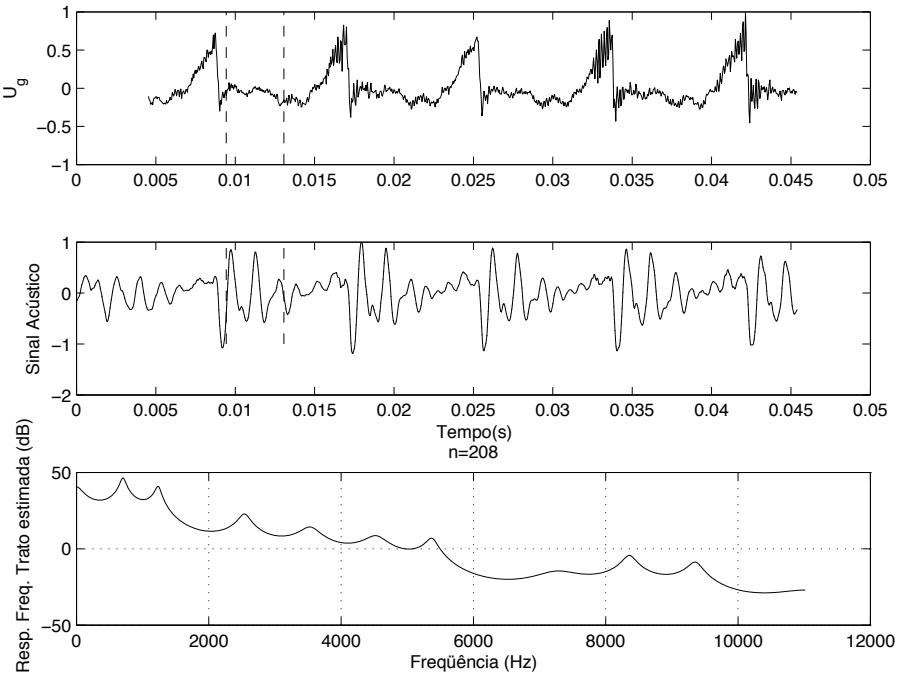

Figura 2.12: Filtragem Inversa de voz real masculina, vogal /a/.

2.2.3 Filtragem inversa a partir do fluxo oral

Um sinal alternativo ao acústico (captado via microfone) para a realização da filtragem inversa é o fluxo aéreo labial, que tem a vantagem de não ser necessária a compensação do termo de radiação labial. Para a captação deste fluxo, utilizou-se a máscara pneumatográfica da Glottal Enterprises (Rothenberg, 1973). Nas Figuras 2.13, 2.14, 2.15, 2.16 e 2.17 tem-se o resultado da filtragem inversa das vogais /a/,/ɛ/,/i/,/ɔ/,/u/ gravadas através deste equipamento. Nestas figuras, o primeiro gráfico mostra o valor do fluxo glótico (U_g) resultante da filtragem inversa do sinal do fluxo aéreo labial (U_L), adquirido através da máscara, mostrado no segundo gráfico. O terceiro gráfico mostra a resposta em freqüência obtida para o trato vocal através da filtragem inversa.

Comparando ao procedimento anterior, que tem o sinal acústico como entrada, o valor de U_g nestes casos parece mais suavizado que no caso do sinal de pressão de voz (microfone) por duas razões: a menor faixa de passagem do sensor de fluxo em relação ao microfone e a menor influência do efeito derivativo da radiação labial. O filtro inverso precisa de ordens mais baixas e desta forma um menor trecho para o cálculo é necessário, sendo maior a probabilidade de se tomar um trecho de glote completamente fechada.

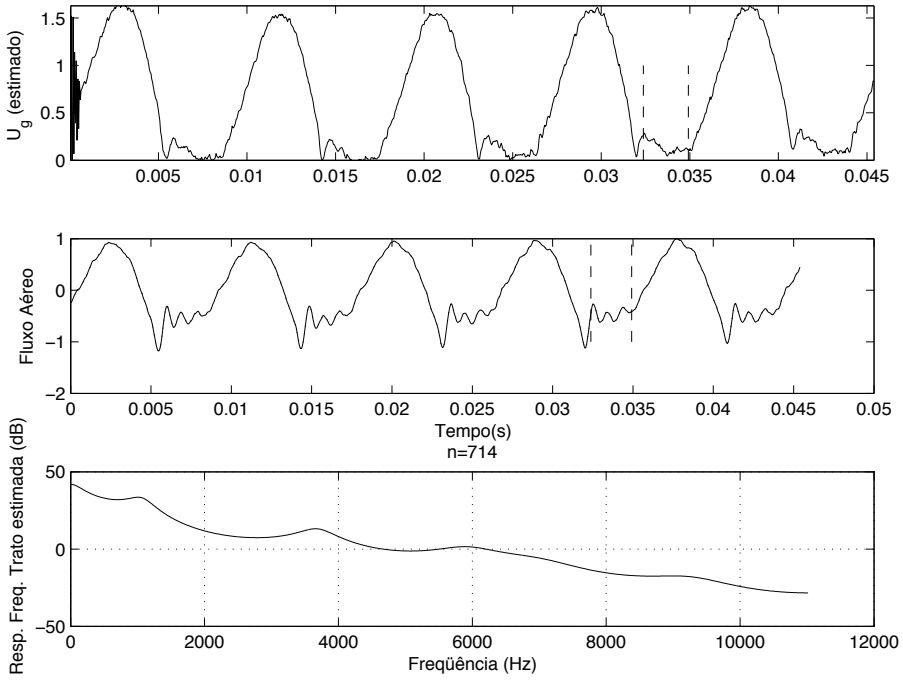

Figura 2.13: Filtragem Inversa de fluxo oral, vogal /a/. O fluxo aéreo foi obtido numa unidade arbitrária. O equipamento utilizado permite uma calibração do nível de tensão (V) em fluxo (ml/s), mas isto não foi feito neste caso.

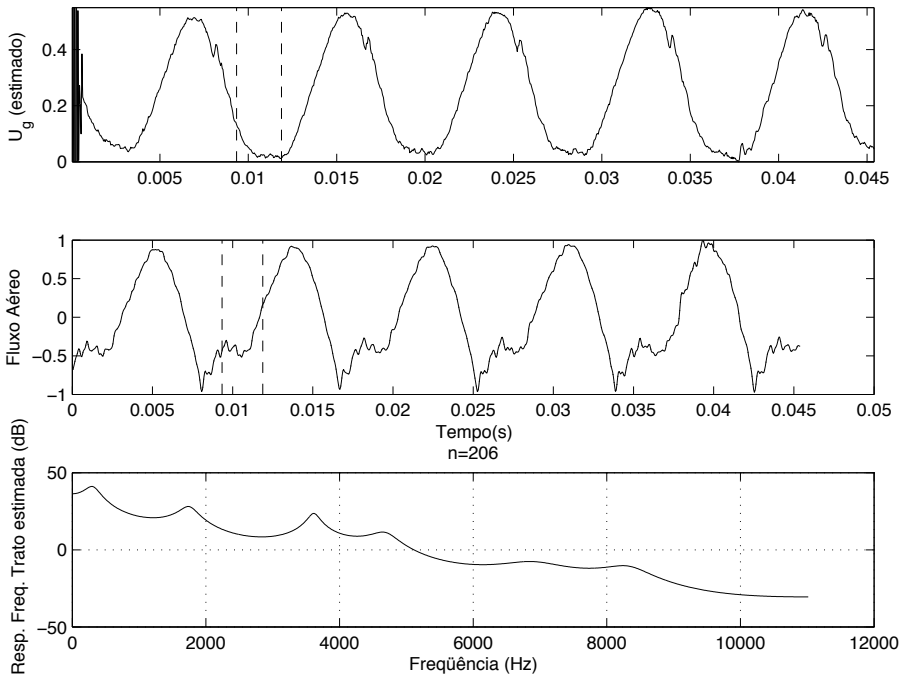

Figura 2.14: Filtragem Inversa de fluxo oral, vogal /ε/. O fluxo aéreo foi obtido numa unidade arbitrária, vide legenda da Figura 2.13.

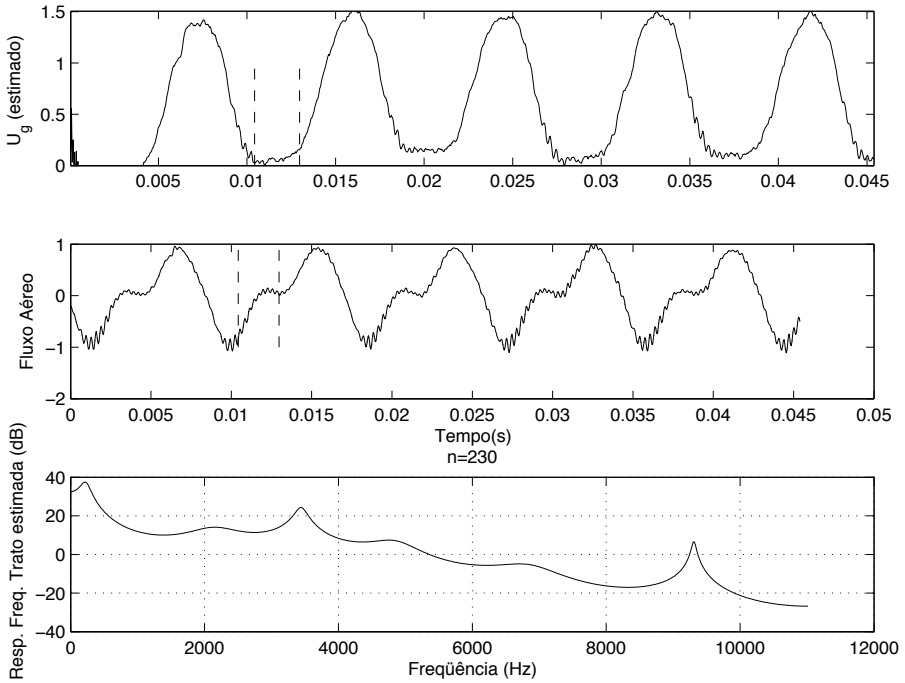

Figura 2.15: Filtragem Inversa de fluxo oral, vogal /i/. O fluxo aéreo foi obtido numa unidade arbitrária, vide legenda da Figura 2.13.

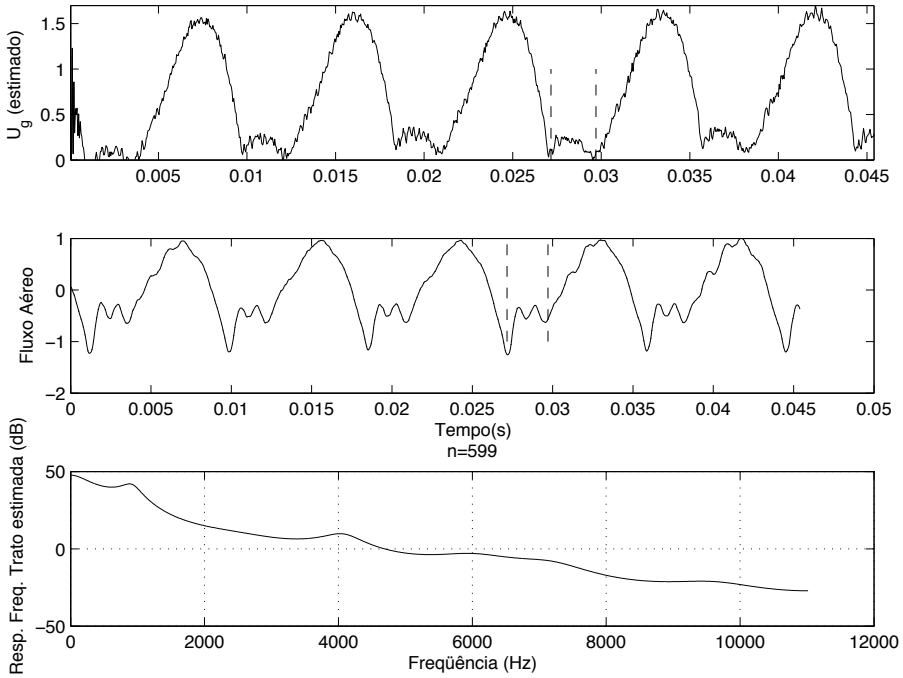

Figura 2.16: Filtragem Inversa de fluxo oral, vogal /ɔ/. O fluxo aéreo foi obtido numa unidade arbitrária, vide legenda da Figura 2.13.

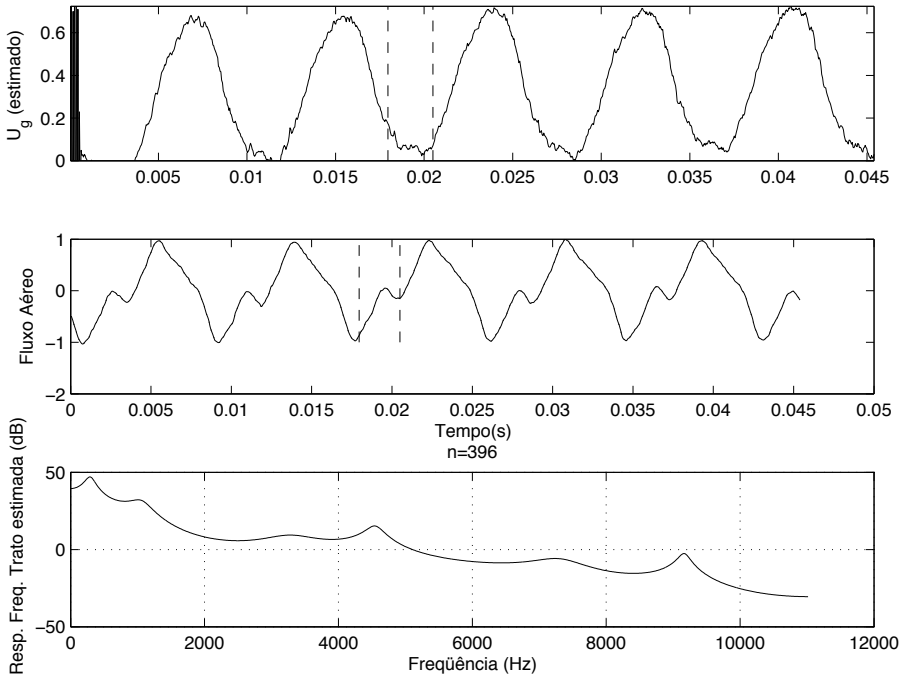

Figura 2.17: Filtragem Inversa de fluxo oral, vogal /u/. O fluxo aéreo foi obtido numa unidade arbitrária, vide legenda da Figura 2.13.

2.2.4 Síntese de sinais de voz para testes

A determinação do filtro em casos reais é feita através da determinação do coeficientes LPC de um trecho englobando alguns períodos da vogal. Este método oferece melhor síntese que utilizando apenas o filtro obtido durante a fase fechada da filtragem inversa. Nesta condição, o filtro teria participação do trato subglótico, que não é englobado na filtragem inversa.

Com as técnicas descritas na Subseção 2.1.2, é possível realizar a síntese de voz de diferentes vogais controlando uma série de parâmetros, tais como freqüência fundamental, velocidade de fechamento, relação de ciclo (*duty cycle*), amplitude e ruído. As Figuras 2.18, 2.19 e 2.20 mostram o sinal base U_g, que é submetido ao filtro encontrado através da LPC. Mostra-se o sinal original, o sinal sintetizado e a comparação dos espectros. Observa-se que os espectros dos sinais sobrepõem-se principalmente na parte inferior do espectro. Já as formas de onde apresentam maior diferença, pois o sinal gerado resultou de uma excitação (fluxo sintético) diferente do fluxo glótico que produziu o sinal original. Além disso, a resposta de fase do filtro usado na síntese não é a mesma da resposta de fase do filtro original, pois a técnica LPC aproxima a densidade espectral de potência.

A ordem do filtro foi escolhida através da regra empírica de um par de pólos por kHz, isto é, um formante por kHz. Desta forma temos $M = 20$ com a amostragem de $F_s = 22050\,Hz$. A voz gravada é de um sujeito do sexo masculino, 25 anos. A gravação foi feita com microfone capacitivo da marca *Shure* através de placa de som dedicada, marca *Turtle Beach*, garantindo relação sinal ruído superior a $60\,dB$.

Nas seções seguintes descrevem-se métodos para introduzir perturbações na forma do pulso glótico sintetizado. Estas perturbações serão usadas para avaliar o desempenho dos algoritmos de medição desenvolvidos neste trabalho.

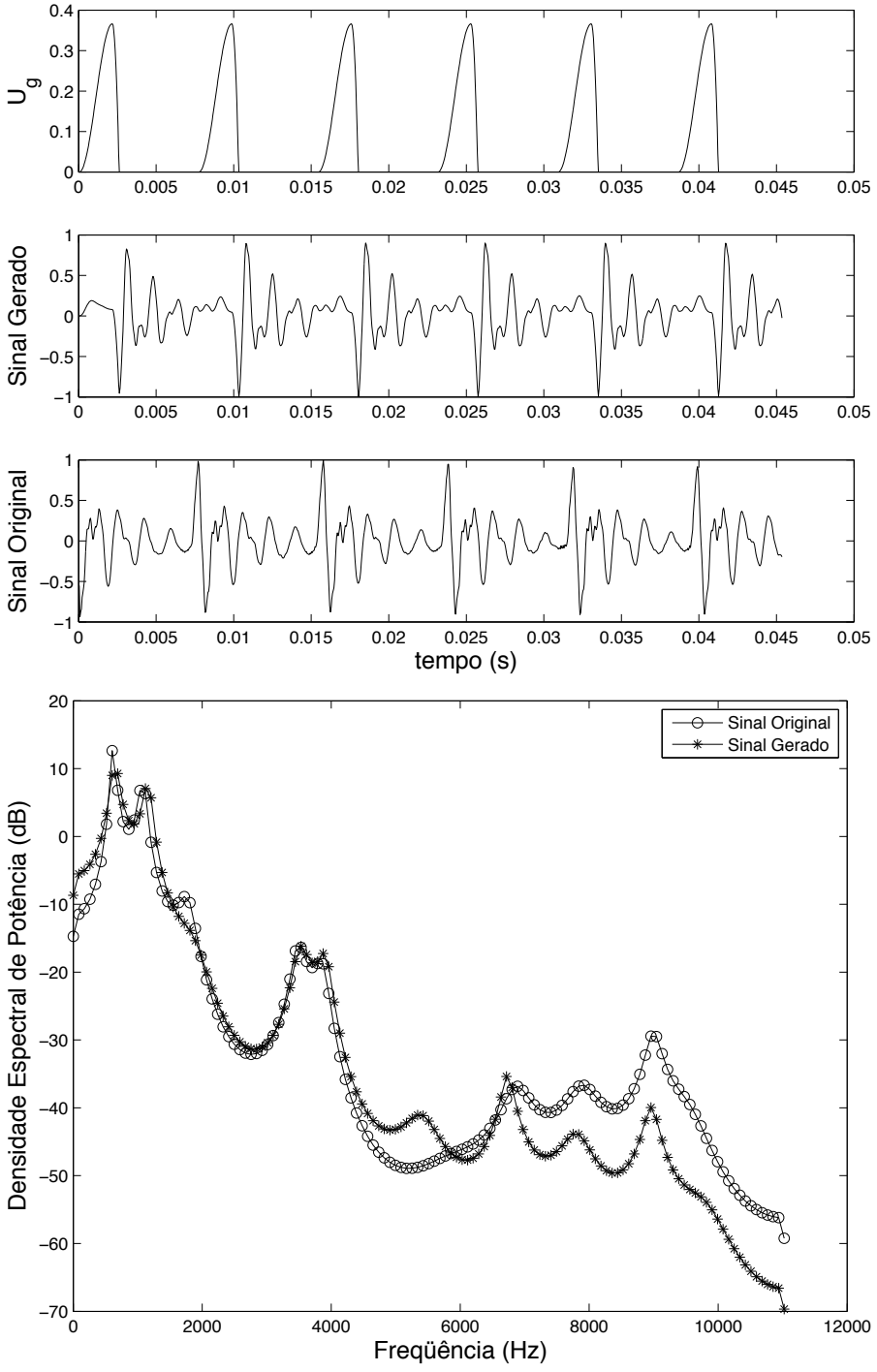

Figura 2.18: Obtenção do filtro a partir de vogal /a/ real, comparação com vogal gerada com fluxo glótico sintético.

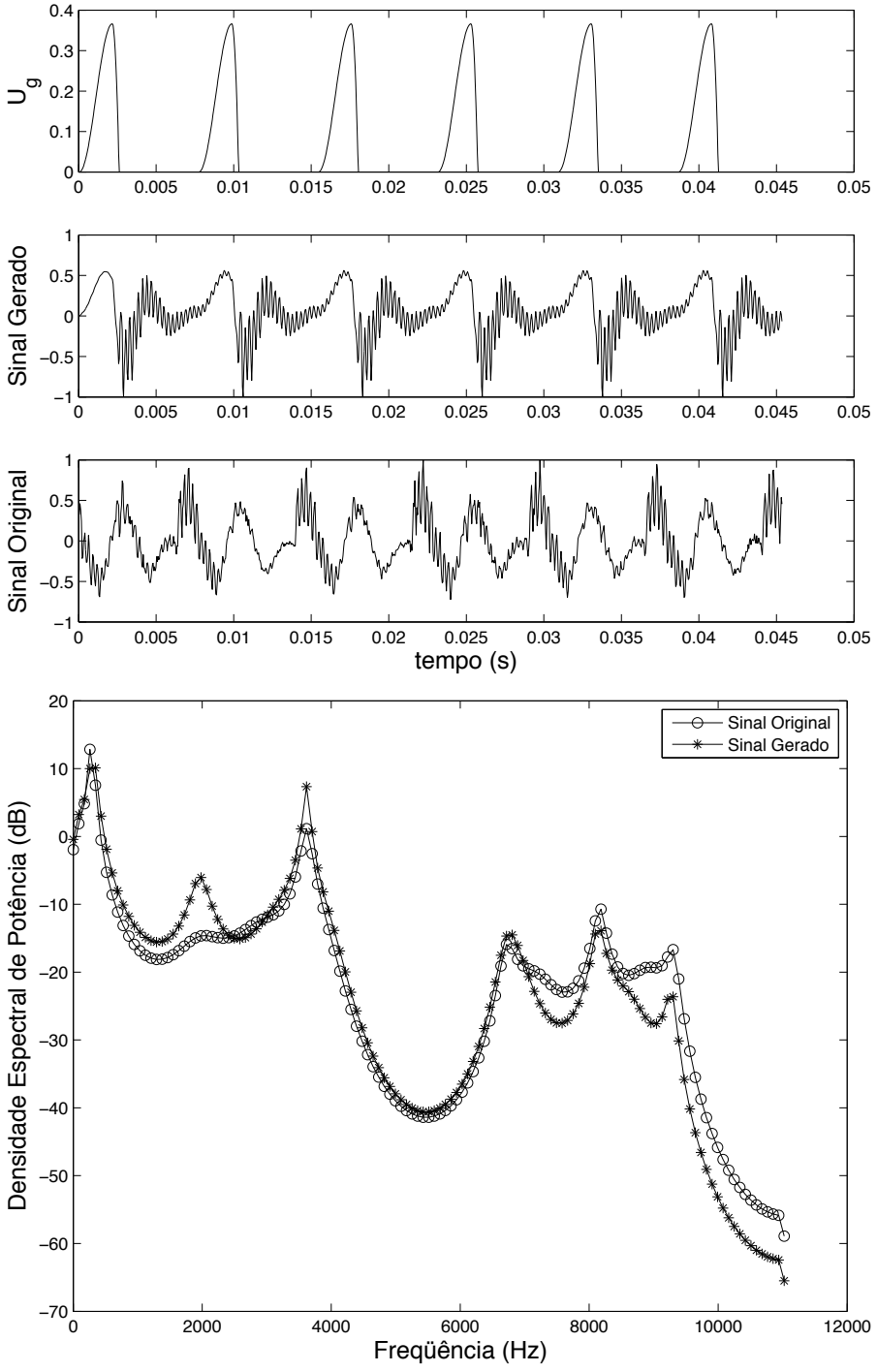

Figura 2.19: Obtenção do filtro a partir de vogal /i/ real, comparação com vogal gerada com fluxo glótico sintético.

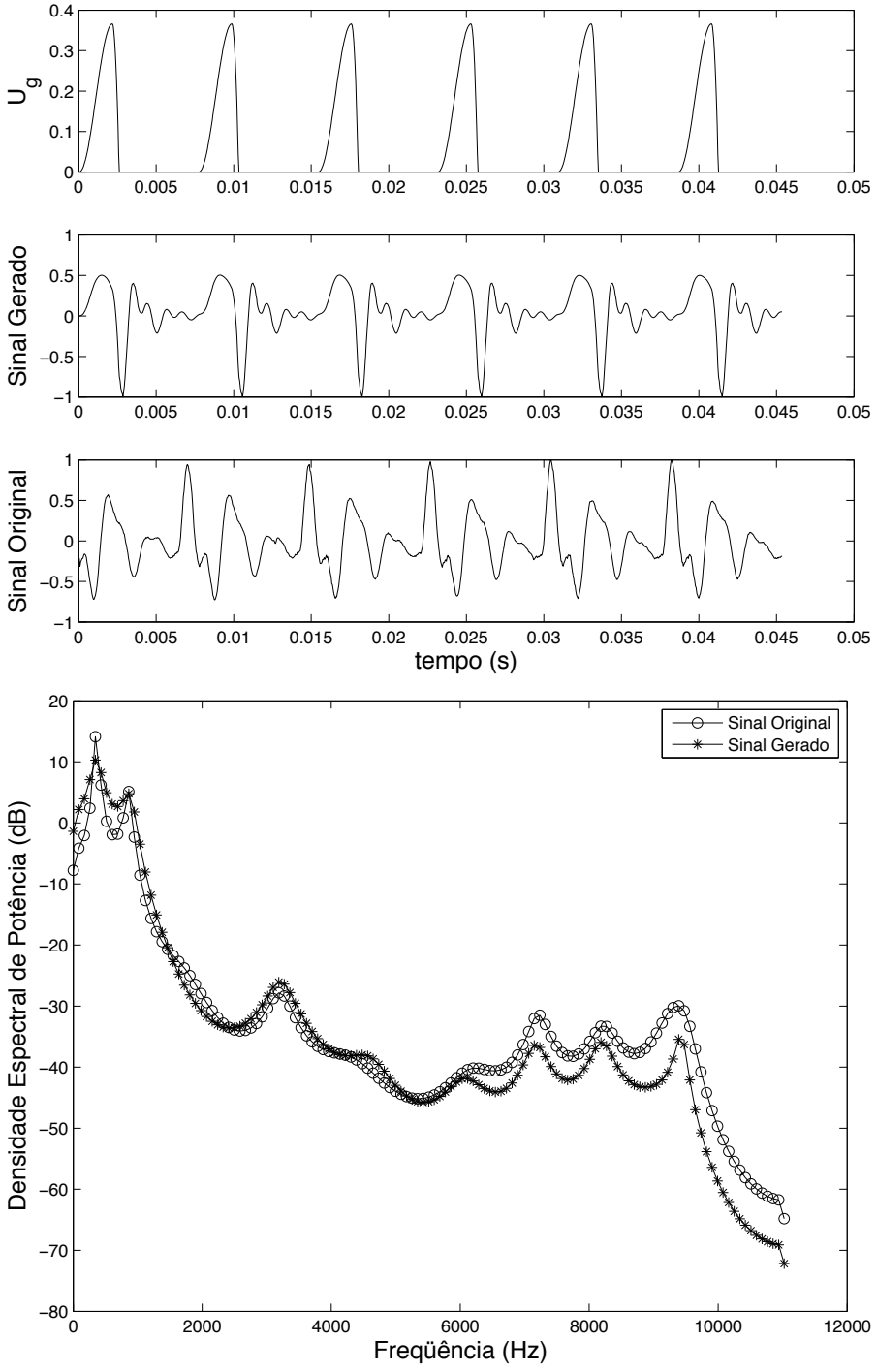

Figura 2.20: Obtenção do filtro a partir de vogal /u/ real, comparação com vogal gerada com fluxo glótico sintético.

2.3 Geração de perturbações

2.3.1 *Jitter*

2.3.1.1 Definição e fontes de *jitter*

Segundo Baken e Orlikoff (2000, Capítulo 6), *jitter* é a perturbação ciclo-a-ciclo em freqüência (ou período). Se o sistema fonatório fosse uma mecanismo ideal e estável, não haveria este tipo de perturbação. Na voz humana, o *jitter* ocorre pela diminuição dos controles neuro-motor e aerodinâmico, além de alterações nos parâmetros do tecido (Hemler et al., 1997) e possivelmente da ação do *feedback* auditivo no controle do sistema fonatório (Burnett et al., 1997).

Basicamente, a fontes deste tipo de perturbação são (Baken e Orlikoff, 2000, Capítulo 6): neurogênicas (ex: variações no controle da tensão muscular), aerodinâmicas, mecânicas (ex: alterações nas propriedades bio-mecânicas do tecido) e estilísticas (ex: vibrato).

A presença de *jitter* é normal na voz. No entanto, níveis aumentados do *jitter* podem ser indicativos de distúrbios da voz, ainda que não possa ser especificamente ligados a patologias laríngeas (Baken e Orlikoff, 2000, Capítulo 6).

2.3.1.2 Simulação de *jitter*

O *jitter* instantâneo, definido de acordo com a função de perturbação de primeira ordem (Titze, 1995), pode ser expresso matematicamente por:

$$J(k) = \frac{[P + \Delta(k)] - [P + \Delta(k-1)]}{\frac{1}{2}\{[P + \Delta(k)] + [P + \Delta(k-1)]\}}. \tag{2.22}$$

Onde $J(k)$ é o valor do jitter no k-ésimo ciclo, P é período fundamental sem *jitter* e $\Delta(k)$ é a perturbação. O período fundamental é portanto $T(K) = P + \Delta(k)$.

Define-se $\overline{J}$ como *jitter* médio, que será obtido através da expressão a seguir:

$$\overline{J} = \frac{1}{N-1} \sum_{k=2}^{N} \frac{|\hat{T}(k) - \hat{T}(k-1)|}{\frac{1}{2}[\hat{T}(k) + \hat{T}(k-1)]}, \tag{2.23}$$

onde $\hat{T}(k)$ é k-ésimo período fundamental de uma dada amostra e N é o número de períodos fundamentais. Logo $J(k)$ pode ser gerada como uma varíavel aleatório de distribuição uniforme entre $\pm\overline{J}$ (a simetria em torno de zero é necessária para reduzir tendências na freqüência fundamental sintetizada). Devido ao sinal de absoluto na equação 2.23, espera-se que o *jitter* esteja distribuído uniformente entre $[0, +2\overline{J}]$. A expressão seguinte, derivada de 2.22 (Vieira, 1997) pode ser utilizada para geração recursiva de $\Delta(k)$,

$$\Delta(k) = \Delta(k-1)\frac{2 + J(k)}{2 - J(k)} + 2P\frac{J(k)}{2 - J(k)}. \tag{2.24}$$

Este valor $\Delta(k)$ é então adicionado ao período nominal P para fornecer o período com *jitter*, $T(k) = P + \Delta(k)$. Na implementação computacional, os períodos instântaneos estão confinados a $0,8P \leq T \leq 1,2P$. Sem as limitações, a freqüência fundamental pode diminuir ou aumentar sem limite, dado que $J(k)$ é uma seqüência aleatória.

2.3.2 *Shimmer*

2.3.2.1 Definição e fontes de *shimmer*

O *shimmer* é a medida da perturbação em amplitude da voz (Baken e Orlikoff, 2000). Como no caso do *jitter*, os valores de *shimmer* servem para quantificar a instabilidade ciclo-a-ciclo do processo fonatório.

A relação da perturbação em amplitude com anormalidades específicas da função glótica ou a distúrbios mais globais da voz não apresenta um consenso entre os pesquisadores (Baken e Orlikoff, 2000). Estudos preliminares indicam que o *shimmer* possa ser inversamente (mas não linearmente) proporcional a intensidade vocal (Orlikoff e Kahane, 1991) e que *jitter* e *shimmer* tendem a covariar (Wong et al., 1995). O *shimmer* pode ocorrer quando há variações ciclo-a-ciclo na velocidade de fechamento e isto pode ocorrer quando há assimetrias entre o lado esquerdo e direito das pregas vocais (Wong et al., 1991).

2.3.2.2 Simulação de *shimmer*

A dedução da expressão do *shimmer* é semelhante à do *jitter*. Sendo A a amplitude nominal, e $S(k)$ o shimmer do k-ésimo período, calculamos $\Delta_s(k)$ como

$$\Delta_s(k) = \Delta_s(k-1)\frac{2+S(k)}{2-S(k)} + 2A\frac{S(k)}{2-S(k)}. \tag{2.25}$$

O valor $\Delta_s(k)$ calculado é então adicionado à amplitude nominal A para fornecer a amplitude com shimmer, $A_s(k) = A + \Delta_s(k)$. As mesmas considerações de limitação de amplitude são tomadas neste caso, sendo as limitações feitas entre $0,2A \leq A_s \leq 1,8A$. Estes limites foram obtidos em Vieira (1997).

2.3.3 Ruído

2.3.3.1 Fontes de ruído turbulento durante a fonação

Segundo Stevens (1998), em laringes normais quando a glote está aduzida ou parcialmente aduzida, ocorre um fluxo aéreo através da constrição da glote. Este fluxo forma um jato que pressiona as paredes do trato vocal ligeiramente acima da glote. O fluxo turbulento gerado nas paredes é considerado como a fonte dominante do ruído no trato vocal, admitindo que as vias aéreas supra-glotais não estejam constritas nas regiões faríngea e oral. Esta fonte de

ruído está localizada entre 1, 0 a 2, 5 cm acima da glote, ou no nível das pregas ventriculares, que estão 0, 5 cm acima das pregas vocais. As fontes de ruído de cada um destes lugares pode ser representada como uma fonte de pressão sonora. Além destas fontes, pode ocorrer uma fonte adicional de ruído decorrente de flutuações aleatório no fluxo glótico. No caso de patologias, pode haver fendas glóticas entre as pregas vocais mesmo na fase de (maior) fechamento e o fluxo nesta constrição será turbulento (Hirano, 1981).

O espectro para fontes de ruído é apresentado em Stevens (1998, p.116). Na voz modal, a componente de fluxo turbulento se aproxima em amplitude da componente periódica em freqüência altas ($f > 3\,kHz$). A componente gerada das flutuações é negligível em relação à componente periódica e praticamente desaparece para $f > 3\,kHz$. A sensação perceptiva do ruído glótico é usualmente denominada de soprosidade.

No caso da voz soprosa, a inclinação da resposta em freqüência da componente periódica é maior que da voz sem escape de ar. Com isto, a componente de fluxo turbulento excede a periódica principalmente em altas freqüências.

2.3.3.2 Simulação do ruído

O ruído é simulado (Vieira, 1997) adicionando uma seqüência $e(n)$, com média nula e uniformente distribuída em cada ciclo, de forma que:

$$x(n) = s(n) + e(n), 0 \leq n < T. \tag{2.26}$$

Admitindo a ergodicidade da seqüência, a largura M da função densidade de probabilidade é ajustada de acordo com a relação sinal-ruído desejada e da potência do período sem ruído, da seguinte maneira:

$$SNR = 10 \log \left[\frac{\frac{1}{T} \sum_{n=0}^{T-1} s^2(n)}{\frac{1}{T} \sum_{n=0}^{T-1} e^2(n)} \right] = 10 \log \left[\frac{\frac{1}{T} \sum_{n=0}^{T-1} s^2(n)}{\frac{M^2}{12}} \right]. \tag{2.27}$$

Desta forma, M é determinado. E o respectivo ruído é adicionado à vogal.

2.4 Conclusão

Neste capítulo foram apresentadas as ferramentas necessárias para a construção de um banco de vogais sintéticas parametrizáveis quanto ao tipo de vogal, freqüência fundamental, e perturbações possíveis de serem aplicadas de forma controlável, tais como relação sinal ruído, perturbações em amplitude (*shimmer*) e perturbações em freqüência (*jitter*).

Nas simulações do próximo capítulo, utilizaremos como modelo de excitação glótica o modelo baseado em *cossenos levantados*, em conjunção com filtros determinados para as vogais /a/, /i/, /u/. A estes serão aplicados o ruído de forma controlada (conforme a Subseção 2.3.3), a perturbação em freqüência (conforme a Subseção 2.3.1) e a perturbação em amplitude

(conforme a Subseção 2.3.2).

Esta abordagem, é mais simples do que outras mostradas ao longo do capítulo, mas permitem ensaios controlados para que se possa avaliar os algoritmos desenvolvidos.

Capítulo 3

Método para medição da relação sinal ruído pelo processamento da imagem espectrográfica

3.1 Introdução

Neste capítulo é introduzido um novo método para a determinação da relação sinal ruído da voz humana sem a necessidade de estimação direta da freqüência fundamental. Para o cálculo deste índice leva-se em conta o espectrograma de uma vogal sustentada e através de métodos de análise de imagem que serão detalhados extrai-se características acústicas do sinal.

O fechamento incompleto da glote durante a fase fechada pode levar ao aumento do fluxo turbulento e do ruído de banda larga, que é um importante correlato acústico da soprosidade.

Visualmente, o espectrograma de uma vogal sustentada apresenta linhas harmônicas bem definidas. Estas linhas definem a componente de "sinal" a ser calculado. De modo análogo, energia entre as linhas harmônicas, que não apresentam a regularidade, podem ser classificadas como ruído.

Por estas características, um especialista treinado consegue inferir pela imagem do espectrograma aspectos da qualidade da voz. Logo, seria possível implementar um método automático para investigação destas imagens. Para tanto, utilizaremos ferramentas utilizadas no pré-processamento de impressões digitais. Impressões digitais são colhidas através de sensores apropriados (imagem, térmico, contato) e estas imagens são pré-processadas para permitir aplicação de métodos de inteligência computacional (classificação, discriminação, etc).

Pode-se estabelecer algumas semelhanças entre espectrogramas e impressões digitais. A partir do estudo dos métodos utilizados para detecção de cristas e vales em impressões digitais, pode-se estabelecer um método para detecção de linhas espectrais no espectrograma da vogal. Não está sendo sugerido, contudo, que espectrogramas da voz tenham a característica

individual das impressões digitais e que permitam a identificação inequívoca do sujeito.

3.2 Avaliação da relação harmônico ruído da voz

Vários métodos para a medição da relação sinal (ou harmônico) ruído estão disponíveis na literatura (Yumoto et al., 1982; de Krom, 1993; Kasuya et al., 1986a). Neste trabalho, detalham-se os métodos baseados na medida no tempo e introduz-se a medição utilizando a imagem do espectro, S^2NR, acrônimo para *Spectrographic Signal-to-Noise Ratio*.

3.2.1 Métodos usuais para cálculo da SNR

Yumoto e colaboradores (Yumoto et al., 1982; Yumoto, 1983; Yumoto et al., 1984) formularam um método para estimação de SNR, baseado na análise de 50 ciclos glotais consecutivos, que pode ser expresso por:

$$SNR = 10\log\left\{\frac{50\sum_{n=0}^{T_{max}}[\overline{x}(n)]^2}{\sum_{i=1}^{50}\sum_{n=0}^{Tmax}[x_i(n)-\overline{x}(n)]^2}\right\}, \tag{3.1}$$

$$\overline{x}(n) = \frac{1}{50}\sum_{i=1}^{50} x_i(n), \tag{3.2}$$

o valor T_{max} é o máximo período glotal entre os 50 ciclos. Se um dado período T_i é menor que T_{max}, o alinhamento é feito acrescentando zeros à respectiva forma de onda, isto é, $x_i = 0$, $T_i \leq n \leq T_{max}$, $\forall\, T_i < T_{max}$. A escolha do número de ciclos é arbitrária, e em Yumoto et al. (1982) alega-se obter neste trecho um intervalo longo suficientemente estável de fonação de uma vogal sustentada.

Os problemas freqüentemente encontrados na aplicação do método são(Cox et al., 1989): dificuldade da demarcação de ciclos e aumento na estimativa do ruído devido a flutuações na duração dos ciclos (*jitter*).

Kasuya et al. (1986a) propuseram um método para estimativa no domínio do tempo baseado em filtros pente (*comb-filter*). Nesta abordagem, admite-se que o sinal de voz, $x(n)$, consiste de uma componente períodica corrompida por ruído aditivo. Detectam-se os períodos do sinal $x(n)$, e submete-se este sinal a um determinado filtro pente. A saída do filtro, $y(n)$, é uma estimativa da componente harmônica de $x(n)$. Estabelece-se então um índice chamado de NNE (*Normalized Noise Energy*, energia do ruído normalizada), definido por:

$$NNE = 10\log\frac{\sum[x(n)-y(n)]^2}{d\sum x(n)^2}, \tag{3.3}$$

onde n são as amostras do sinal que correspondem aos ciclos considerados e d é uma constante de compensação, que depende do número de ciclos utilizados, no caso do artigo citado, $d = 7$.

Em Kasuya et al. (1986b), propõe-se uma versão para a NNE na freqüência. Neste caso, tomam-se 7 ciclos glóticos e calcula-se a transformada de Fourier do trecho. O número de

ciclos é definido de modo a minimizar efeitos de janelamento (Kasuya et al., 1986b). A energia do ruído é dada pela energia entre os harmônicos. Na seqüência, esta valor é normalizado pelo valor da energia total do espectro, gerando a NNE no domínio da freqüência.

Outros métodos são encontrados na literatura, como uma tentativa baseada em filtros pente cepstrais (de Krom, 1993) ou ainda métodos que tentam melhorar a performance dos citados anteriormente (Awan e Frenkel, 1994). Nos métodos baseados na demarcação periódica, o *jitter* parece ser a maior fonte erros de estimação da SNR (Cox et al., 1989).

O algoritmo temporal que utilizado no presente trabalho para comparação com o método proposto nesta dissertação está descrito em Vieira (1997). O método consiste na estimativa ciclo-a-ciclo da SNR usando filtros pente definidos por:

$$\hat{s}(n) = \frac{1}{2}[x(n) + x(n-T)], \tag{3.4}$$

$$\hat{e}(n) = \frac{1}{2}[x(n) - x(n-T)]. \tag{3.5}$$

Nas expressões acima, $x(n)$ é a n-ésima amostra do sinal de voz, $\hat{s}(n)$ e $\hat{e}(n)$ são as estimativas instantâneas das componentes harmônica e de ruído, respectivamente. T é o período fundamental normalizado ($T = F_s/f_o$, com F_s a freqüência de amostragem e f_o a freqüência fundamental). Pode-se mostrar, conforme Vieira (1997), que a magnitude da resposta em freqüência dos filtros é:

$$\left|\frac{\hat{S}(f)}{X(f)}\right| = |\cos\frac{\pi f}{f_o}|, \tag{3.6}$$

$$\left|\frac{\hat{E}(f)}{X(f)}\right| = |\mathrm{sen}\frac{\pi f}{f_o}|. \tag{3.7}$$

De acordo com a equação 3.6, os máximos encontram-se nos múltiplos de f_o, enquanto neste pontos encontram-se os mínimos da expressão 3.7.

Para segmentação dos períodos é utilizado o método apresentado em Schäfer-Vincent (1983). Este método utiliza um conjunto de regras baseadas no conhecimento prévio da forma de onda de vogais, utilizando correlações empíricas para identificar e encadear pares de pulsos glotais, chamados períodos gêmeos, delimitados por t_1 e t_2 para o primeiro período, t_2 e t_3 para o segundo.

A expressão para o cálculo local da SNR ($LSNR$) é dada por:

$$LSNR(t_1, t_2, \tau) = 10\log\frac{\sum_{i=0}^{0,8N}[x(t_1+i) + x(t_2+\tau+i)]^2}{\sum_{i=0}^{0,8N}[x(t_1+i) - x(t_2+\tau+i)]^2}, -10 \leq \tau \leq 10, \tag{3.8}$$

$$N = \min(t_3 - t_2, t_2 - t_1). \tag{3.9}$$

Na expressão 3.8 fazem-se as seguintes observações: τ é o valor dentro de um pequeno intervalo que maximiza a $LSNR$, os somatórios estão limitados a 80% do período de cada

segmento, artifícios utilizados para compensar erros de alinhamento devido ao *jitter*. Outro artifício utilizado é a utilização do período gêmeo de menor duração de cada par. Estas otimizações estão discutidas em detalhe em Vieira (1997).

A expressão do SNR final é calculada pela média ao longo da amostra. Este valor será chamado neste trabalho de $SNR(t)$, em referência à sua característica temporal.

3.2.2 Avaliação por imagem do espectrograma

O método aqui proposto pretende extrair da imagem do espectrograma as informações do conteúdo harmônico e do ruído utilizando técnicas de processamento de imagem, sem a necessidade da demarcação individual dos ciclos glóticos.

A imagem a ser estudada é o espectrograma de uma vogal sustentada com duração mínima de centenas de milissegundos.

3.2.2.1 Descrição do método

O algoritmo base é inspirado em Hong et al. (1998) e em sua implementação base disponibilizada por Kovesi (2005) para Matlab e GNU/Octave. Os parâmetros foram ajustados de forma empírica para o espectrograma.

Etapas do algoritmo

Para a utilização dos algoritmos de detecção de impressão digital, faz-se uma analogia entre as cristas e vales da pele, com os picos e vales espectrais do espectrograma.

Em linhas gerais, o algoritmo executa as seguintes ações, que serão detalhadas nas seções posteriores e estão demonstradas no diagrama de blocos na Figura 3.1:

- Geração da imagem do Espectrograma: Obtenção do espectrograma a partir da voz;
- Normalização: A imagem é normalizada e passa a ter média e desvio padrão desejados;
- Segmentação: separação das regiões onde existem cristas e vales das regiões onde ocorre apenas ruído;
- Estimação de orientação: Gera-se uma nova imagem que contém a orientação das linhas (harmônicos) a partir da imagem normalizada. Nesta etapa, a imagem é dividida em blocos;
- Estimação de freqüência espacial: verifica-se a freqüência espacial de cristas e vales em determinado sentido de orientação para definir o sentido de orientação mais confiável;
- Filtragem: filtram-se regiões espúrias e artefatos. A filtragem utiliza filtros de Gabor, por blocos. Os parâmetros para sintonizar os filtros são retirados da assinatura (freqüência e orientação das cristas). Os filtros de Gabor são direcionais e seletivos;

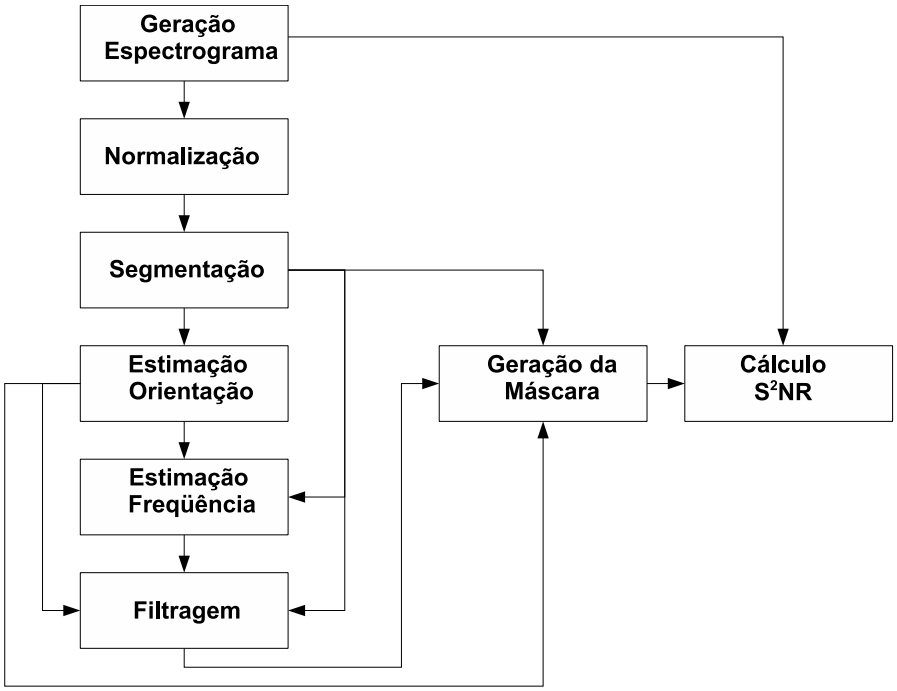

Figura 3.1: **Diagrama de Blocos.** Etapas do cálculo da S^2NR.

- Geração da máscara: gera-se uma máscara para as regiões de sinal e ruído, utilizando as informações das etapas anteriores (resultado da filtragem, segmentação e da orientação);
- Cálculo da S^2NR.

A máscara gerada na última etapa indica a posição onde estão as linhas espectrais (sinal). O inverso desta máscara equivale às regiões de ruído. Desta forma, obtém-se a componente harmônica multiplicando-se a máscara do sinal pelo espectrograma, enquanto a componente de ruído é determinada multiplicando o inverso da máscara binária pelo espectrograma.

A relação sinal-ruído é calculada a partir das imagens das componente harmônica e da componente de ruído. A relação sinal ruído é calculada ao longo do tempo, isto é, em colunas (verticais) do espectrograma. Para tal, soma-se cada coluna da componente harmônica e divide-se pela correspondente na imagem do ruído. Desta forma, obtém-se uma linha com a evolução do sinal-ruído no tempo. Finalmente, passam-se estes dados por um filtro de mediana, evitando grandes flutuações no resultado.

3.2.3 Detalhamento do algoritmo

3.2.3.1 Geração da imagem do espectrograma

Para a geração de imagem do espectrograma, utiliza-se a ferramenta do *toolbox* de Processamento de Sinais do Matlab R2007a (7.4) que permite a escolha do tamanho da janela da FFT e da sobreposição das janelas. A entrada é o arquivo de voz, no formato *WAV* (Murray e vanRyper, 1996) e a saída é uma imagem, onde o eixo x corresponde ao tempo, o eixo y à freqüência e o eixo z é a transformada de Fourier discreta no ponto correspondente. Esta imagem será denominada B, com dimensões $M \times N$, onde N é a metade da janela da FFT utilizada e M depende da sobreposição entre as janelas e da duração da amostra.

A imagem inicialmente tem valor complexo, por se tratar da *STFT* (*Short-time Fourier transform*). Para efeito de cálculo, obtém-se a magnitude de cada ponto e em seguida normaliza-se a imagem de modo que tenha valor mínimo igual a 0 e valor máximo igual a 1.

O cálculo para a obtenção da imagem normalizada, A, é executado da seguinte forma:

$$A = \frac{|B| - \min|B|}{\max|B|}. \tag{3.10}$$

Para auxílio na detecção do sinal, gera-se uma nova imagem, A_{log}, obtida por uma compressão logarítmica da magnitude da imagem original. Esta imagem também é normalizada entre 0 e 1. Pode-se expressar as operações descritas da seguite maneira:

$$C = \log|B|, \tag{3.11}$$

$$A_{log} = \frac{C - \min C}{\max C}. \tag{3.12}$$

3.2.3.2 Etapa de normalização

A próxima etapa corresponde à normalização da imagem logarítmica, A_{log}, de modo a ter média nula e desvio padrão unitário, obtendo a imagem A_{norm}, considerando $\overline{A}_{log}$ o valor médio de A_{log} e σ_A o desvio padrão de A_{log}:

$$A_{norm} = \frac{A_{log} - \overline{A}_{log}}{\sigma_A}. \tag{3.13}$$

A média e o desvio padrão da imagem são calculados por:

$$\overline{A}_{log} = \frac{1}{N^2}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} A_{log}(i,j), \tag{3.14}$$

$$\sigma_A = \sqrt{\frac{1}{N^2}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left(A_{log}(i,j) - \overline{A}_{log}\right)^2}. \tag{3.15}$$

3.2.3.3 Etapa de segmentação

Esta etapa corresponde a uma tentativa inicial de segmentar as regiões onde ocorrem padrões com harmônicos das regiões onde ocorre puramente ruído. Partindo do princípio de que nas regiões onde há presença apenas de ruído a variância é mais baixa do que nas regiões onde ocorre predominantemente o sinal da voz, pode-se segmentar estas duas regiões aplicando um limiar ao desvio padrão (σ_{th}). Primeiro, a imagem é dividida em blocos e o desvio padrão é calculado para cada bloco. Se o desvio padrão é menor que o limiar definido, considera-se como região de ruído. No outro caso, considera-se região potencial de sinal. O desvio padrão de um bloco de tamanho $W \times W$ é definido por:

$$\sigma_{bloco}(k) = \sqrt{\frac{1}{W^2}\sum_{i=0}^{W-1}\sum_{i=0}^{W-1}[A_{norm}(i,j) - M(k)]^2}, \tag{3.16}$$

sendo $\sigma_{bloco}(k)$ o desvio padrão de cada bloco k, $A_{norm}(i,j)$ a imagem calculada no item anterior no pixel (i,j) e $M(k)$ é o valor médio para o bloco k.

A aplicação do limiar resulta numa máscara binária, que daqui em diante será referida como máscara de segmentação (M_s), e será importante no cálculo da máscara final.

$$M_s(i,j) = \begin{cases} \text{verdadeiro}, & \text{se} \quad \sigma_{bloco}(k) \geq \sigma_{th} \\ \text{falso}, & \text{caso contrário} \end{cases}, \tag{3.17}$$

O próximo passo é gerar uma imagem normalizada a partir da máscara de segmentação e de A_{norm}. A normalização, que não afeta as estruturas de cristas e vales da imagem, é utilizada para padronizar os níveis dinâmicos de variação, facilitando os processos seguintes de tratamento da imagem.

Neste caso, deseja-se obter a imagem A_s, a imagem normalizada final, visando média nula e desvio padrão unitário. O cálculo é análogo ao da expressão (3.13) e a única diferença está no cálculo da média e do desvio padrão onde, neste caso, são excluídos os valores em (i,j) que não estejam na máscara de segmentação, quer dizer $M_s(i,j) =$ falso. Portanto, as saídas desta seção são A_s e M_s e os parâmetros de entrada W e σ_{th}. Uma saída típica desta etapa esta na Figura 3.2.

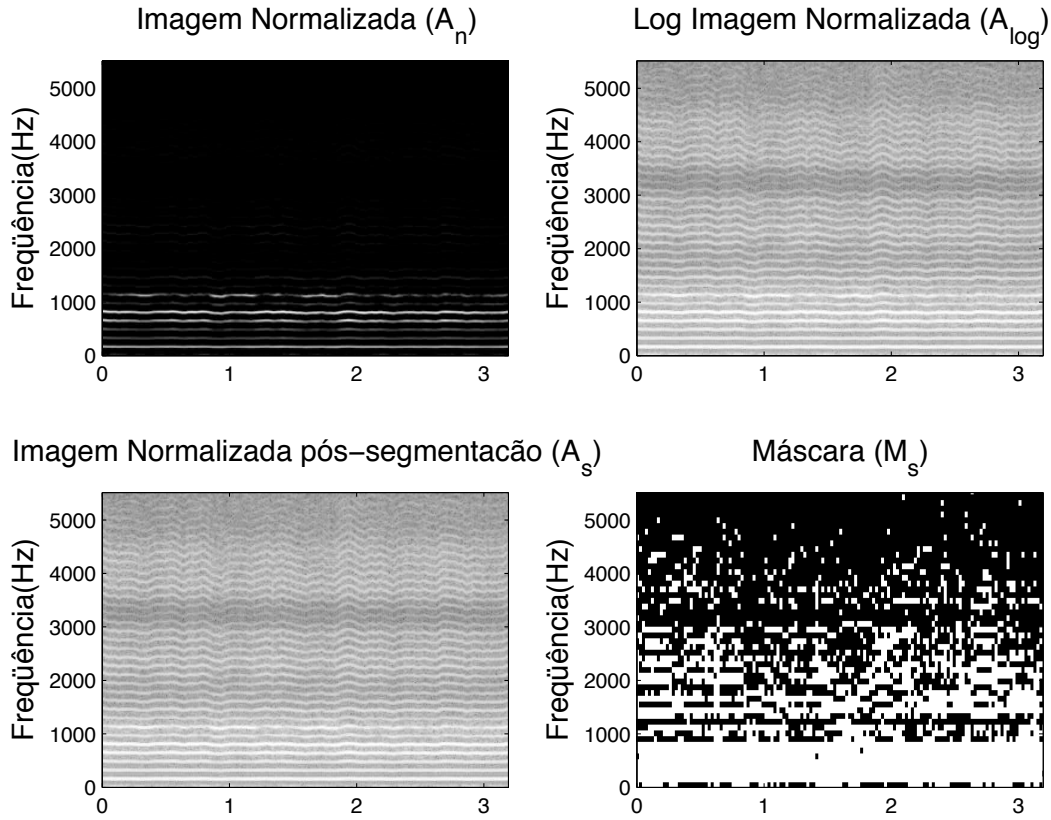

Figura 3.2: **Etapa de segmentação.** Os gráficos apresentam um exemplo de imagens intermediárias para o cálculo da S^2NR. O eixo horizontal é tempo, em segundos. Na escala de tons de cinza, tons mais claros indicam valores maiores.

3.2.3.4 Estimação de orientação

O campo de orientação de uma imagem extraída do espectrograma define a orientação local das cristas do espectro (em analogia às cristas da impressão digital). A estimação da orientação é portanto um passo importante para o processo de tratamento da imagem, já que o próximo estágio (filtro de Gabor) depende desta estimativa de orientação como parâmetro.

Um método de mínimos quadrados baseado nos trabalhos de Hong et al. (1998) e Bazen (2002) foi utilizado. Sabe-se que os gradientes são a orientação no nível dos *pixels*, enquanto os campos de orientação descrevem a orientação das estruturas vale-crista do espectrograma (como na impressão digital), mas em uma escala mais larga. Assim, o campo de orientação pode ser extraído fazendo operações de média dos *pixels* da vizinhança. Esta abordagem permite também estabelecer um índice de confiança para a estimativa da orientação, sendo portanto uma informação a mais na segmentação da imagem na detecção do vozeamento.

Matematicamente, pode-se descrever o método conforme Bazen (2002). Inicialmente, calcula-se o gradiente, ∇G, da imagem, garantindo que este tenha orientação entre $-\pi/2$ e $\pi/2$, da seguinte forma:

$$\begin{bmatrix} G_x(x,y) \\ G_y(x,y) \end{bmatrix} = \text{sinal}(G_x)\nabla I(x,y), \tag{3.18}$$

onde $I(x,y)$ representa uma imagem em tons de cinza, e a função sinal(x) é definida por:

$$\text{sinal}(x) = \begin{cases} +1, x \geq 0 \\ -1, x < 0 \end{cases} \tag{3.19}$$

Fazendo, para o cálculo de G_s, as componentes quadráticas de G,

$$G_{s,x} + jG_{s,y} = (G_x + jG_y)^2 = (G_x^2 - G_y^2) + j(2G_xGy), \tag{3.20}$$

tem-se :

$$\begin{bmatrix} G_{s,x} \\ G_{s,y} \end{bmatrix} = \begin{bmatrix} G_x^2 - G_y^2 \\ 2G_xGy \end{bmatrix}. \tag{3.21}$$

De forma genérica, calcula-se a média de G_s, em uma determinada vizinhança, utilizando uma janela W (incluindo janelas não uniformes):

$$\begin{bmatrix} \overline{G_{s,x}} \\ \overline{G_{s,y}} \end{bmatrix} = \begin{bmatrix} \sum^W G_{s,x} \\ \sum^W G_{s,y} \end{bmatrix} = \begin{bmatrix} \sum^W (G_x^2 - G_y^2) \\ \sum^W 2G_xGy \end{bmatrix} = \begin{bmatrix} G_{xx} - G_{yy} \\ 2G_{xy} \end{bmatrix}. \tag{3.22}$$

As expressões $G_{xx} = \sum^W G_x^2$, $G_{yy} = \sum^W G_y^2$ e $G_{xy} = \sum^W G_xG_y$ são estimativas para a covariância e covariância cruzada de G_x e G_y. O gradiente de direção médio (Φ) é dado por:

$$\Phi = \frac{1}{2}\angle(G_{xx} - G_{yy}, 2G_{xy}), \tag{3.23}$$

sendo operador $\angle(x,y)$ definido por:

$$\angle(x,y) = \begin{cases} \arctan(y/x) & \text{para} \quad x \geq 0 \\ \arctan(y/x) + \pi & \text{para} \quad x < 0 \wedge y \geq 0 \\ \arctan(y/x) - \pi & \text{para} \quad x < 0 \wedge y < 0 \end{cases}, \tag{3.24}$$

A imagem de orientação dada por θ, com $-\frac{1}{2}\pi < \theta \leq \frac{1}{2}\pi$, é perpendicular a Φ:

$$\theta = \begin{cases} \Phi + \frac{1}{2}\pi & \text{para} \quad \Phi \leq 0 \\ \Phi - \frac{1}{2}\pi & \text{para} \quad \Phi > 0 \end{cases}. \tag{3.25}$$

Define-se também o índice de confiança da orientação R. Sua expressão é

$$R = 1 - \frac{I_{min}}{I_{max}}, \tag{3.26}$$

sendo I_{min} e I_{max} os momentos de inércia calculados sobre o eixo de orientação e sobre um

eixo perpendicular. Suas expressões são:

$$I_{min} = \frac{G_{xx}+G_{yy}}{2} - \frac{(G_{xx}-G_{yy})\cos 2\theta}{2} - \frac{G_{xy}\text{sen}2\theta}{2}, \quad (3.27)$$

$$I_{max} = G_{yy} + G_{xx} - I_{min}. \quad (3.28)$$

A medida da confiança é mínima quando a relação I_{min}/I_{max} tende à unidade, isto é, tem-se pouca informação de orientação no ponto.

A implentação computacional do algoritmo sofreu algumas adaptações (Kovesi, 2005), já que $I(x, y)$ não se trata de uma função contínua e sim de um sinal amostrado bi-dimensional. Para tanto, a função gradiente G é implementada através da convolução da imagem com máscaras especiais, geradas pelo gradiente da função gaussiana nas direções x e y. O parâmetro σ_{grad} define o desvio padrão, e também a dimensão da matriz, que deve ter tamanho de aproximadamente $6\sigma_{grad}$ (arredondado para o inteiro ímpar superior), já que a energia da gaussiana está praticamente contida no raio de 3σ. Esta abordagem oferece uma melhor aproximação para a função gradiente do que a simples utilização de um filtro de Sobel 3×3, já que permite o controle de σ_{grad} e por conseqüência o tamanho da matriz de convolução (Nixon e Aguado, 2002).

As somas ponderadas da equação (3.22) são calculadas através da convolução das respectivas matrizes com a matrizes de convolução geradas por funções gaussianas, com desvio σ_{bloco}.

Outra adaptação é a suavização através da utilização de um filtro de gaussiana, com desvio σ_{orient} dos parâmetros de entrada da expressão dada em (3.23), de modo a melhorar a estabilidade do algoritmo.

Logo, pode-se resumir que neste trecho do algoritmo, a partir da imagem A_s, e dos parâmetros σ_{grad}, σ_{bloco} e σ_{orient}, obtêm-se a imagem da orientação local (θ) e uma imagem com a confiança na orientação (R) dos respectivos pontos.

Um exemplo do resultado desta etapa pode ser visto na Figura 3.3, onde vêm-se linhas de orientação sobrepostas ao espectrograma. A imagem R correspondente pode ser vista na Figura 3.4.

Matematicamente, este algoritmo é equivalente à análise em componentes principais (PCA), conforme demonstra Bazen (2002).

3.2.3.5 Estimação de freqüência

Além da imagem da orientação (θ), um outro parâmetro importante para a etapa de filtragem com os filtros de Gabor é a estimativa da freqüência local das cristas (freqüência espacial). O primeiro passo é dividir a imagem (A_s) em blocos de tamanho $W_2 \times W_2$. O próximo passo é projetar os valores dos *pixels* de cada bloco perpendicularmente à respectiva imagem de orientação, θ para o local, conforme pode ser visto na Figura 3.5.

Os picos na projeção correspondem aos harmônicos (cristas) e estes são detectados para o cálculo da freqüência. O cálculo do comprimento de onda (λ) é feito pela distância do

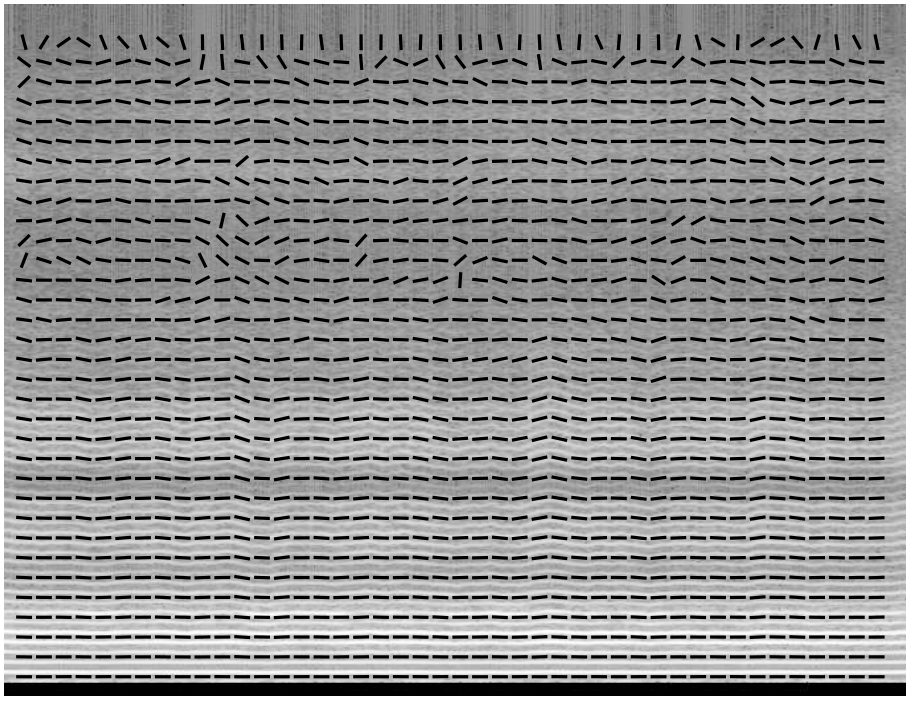

Figura 3.3: **Determinação da orientação.** Espectrograma com a indicação da orientação encontrada. O eixo vertical é a freqüência (0 - 11025 Hz) e o eixo horizontal é o tempo (0 - 3 s).

primeiro ao último pico, divindo pelo número de picos menos 1. Logo, a freqüência é dada por $1/\lambda$.

Na implementação computacional, são necessárias algumas adaptações (Kovesi, 2005), como a limitação do valor possível de λ entre dois valores, λ_{min} e λ_{max}. Aos valores fora desta faixa e em caso onde ocorre apenas um pico no bloco é atribuído um valor nulo para λ.

Logo, este bloco a partir da imagem de orientação θ, da imagem A_s e dos parâmetros W_2, λ_{max} e λ_{min} retorna uma imagem, F, com a freqüência local das cristas na imagem, F.

3.2.3.6 Filtragem

Uma vez que foram estimadas as informações de orientação e freqüência, estes parâmetros são utilizados para o projeto dos filtros de Gabor de simetria par. Um filtro bi-dimensional de Gabor consiste em uma onda senoidal plana com orientação e freqüência pré-determinadas, modulada por um envelope gaussiano, conforme mostra Daugman (1985). Este filtros são utilizados por possuírem propriedades de seletividade em freqüência e orientação. Estas propriedades permitem sintonia ótima do filtro para uma freqüência e orientação específica na imagem. Desta maneira, é possível preservar as estruturas de cristas na imagem e efetivamente reduzir o ruído.

Um filtro de Gabor de simetria par corresponde à parte real da função de Gabor, dada

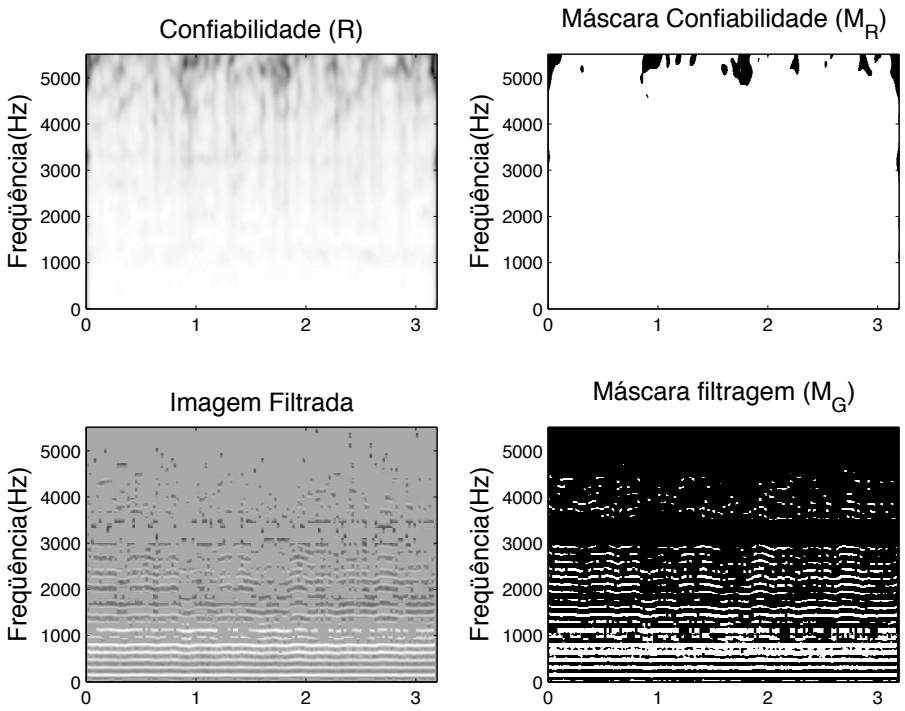

Figura 3.4: **Etapas de Orientação e Filtragem.** Nos gráficos na parte superior da figura, pode-se observar a imagem da confiabilidade R, e a máscara M_R obtida ao aplicar o limiar de confiabilidade. Na linha inferior, observa-se o resultado da filtragem utilizando os filtros de Gabor e a respectiva máscara obtida com a utilização de um limiar nulo. O eixo horizontal indica o tempo, em segundos.

pela onda cossenoidal modulada pela gaussiana. Este filtro é definido por Jain e Farrokhnia (1990) como:

$$G(x, y : \theta, f) = \exp\left\{-\frac{1}{2}\left[\frac{x_\theta^2}{\sigma_x^2} + \frac{y_\theta^2}{\sigma_y^2}\right]\right\}\cos(2\pi f x_\theta), \tag{3.29}$$

$$x_\theta = x\cos\theta + y\sin\theta, \tag{3.30}$$

$$y_\theta = -x\sin\theta + y\cos\theta, \tag{3.31}$$

onde θ é a orientação do filtro de Gabor, f é a freqüência da onda plana cossenoidal e σ_x e σ_y os desvios padrão do envelope gaussiano ao longo dos eixos x e y da imagem.

A freqüência característica do filtro, f, é completamente determinada pela freqüência local de cristas e a orientação é dada pela orientação local das cristas. A seleção de σ_x e σ_y envolve um compromisso, já que com estes valores elevados, os filtros são mais robustos ao ruído, sendo, no entanto, mais propensos a criar cristas e vales inexistentes. No outro sentido, um valor baixo impede o surgimento de cristas e vales espúrios, mas permite a passagem do ruído. Os valores para σ_x e σ_y são determinados empiricamente.

O filtro de Gabor é aplicado à imagem através de convolução. Isto requer, para cada

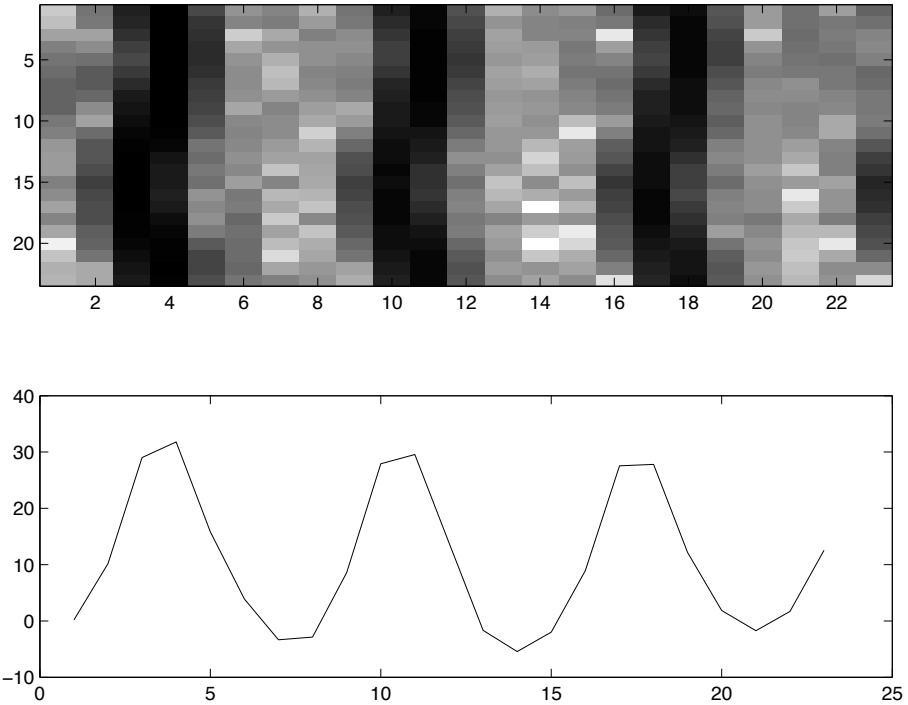

Figura 3.5: **Estimação de freqüência.** A figura mostra um bloco da imagem, rotacionado de modo que suas linhas de espectro estejam ortogonais à imagem de orientação local, e a respectiva projeção sobre o eixo, facilitando desta forma a contagem dos picos. Os eixo horizontal e vertical indicam o número da amostra do bloco de imagem, que está rotacionado. Nesta figura, a escala de tons de cinza implica em tons escuros para maiores valores.

pixel (i, j), a orientação correspondente em $\theta(i, j)$ e o valor da freqüência espacial em $F(i, j)$. A imagem filtrada é dada por:

$$E(i,j) = \sum_{u=-\frac{w_x}{2}}^{\frac{w_x}{2}} \sum_{v=-\frac{w_y}{2}}^{\frac{w_y}{2}} G(u, v; \theta(i,j), F(i,j)) A_s(i-u, j-v), \tag{3.32}$$

sendo $\theta(i, j)$ a imagem de orientação, $F(i, j)$ a imagem de freqüência espacial, A_s a imagem do espectrograma normalizada, w_x e w_y a largura e o comprimento da máscara dos filtros de Gabor, respectivamente. Devido ao fato de a energia destes filtros estar em maior parte contida em $[-3\sigma, 3\sigma]$, limita-se w_x e w_y ao valor inteiro superior a 6σ, sendo $\sigma = \max(\sigma_x, \sigma_y)$.

Como os filtros de Gabor têm componente contínua nula, a imagem filtrada tem média zero. A nova imagem, E, é em seguida "binarizada" (limiarizada) , utilizando um limiar nulo, gerando a imagem binária M_G. Um exemplo de aplicação do filtro pode ser visto na Figura 3.4, com o resultado da filtragem e da aplicação do limiar. A Figura 3.6 mostra o Filtro de Gabor com $\sigma_x = \sigma_y = 2$, com orientação horizontal.

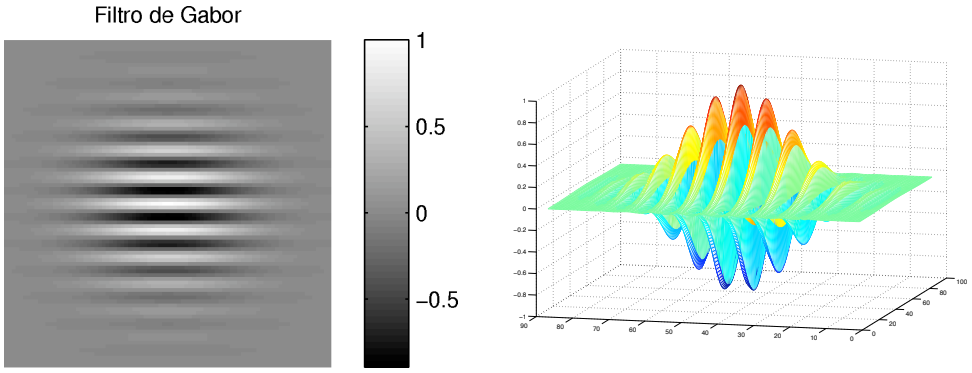

Figura 3.6: **Filtro de Gabor.** A figura mostra um exemplo do filtro de Gabor, com orientação nula e $\sigma_x = \sigma_y = 2$. Visualizações em tons de cinza e em três dimensões.

3.2.3.7 Geração da máscara

Após seguidas etapas de processamento, tem-se uma primeira estimativa para as regiões onde ocorre o sinal, dados na máscara binária, M_s, obtida da imagem original. Tem-se também a imagem de confiança da orientação, R, à qual aplicamos um limiar (σ_R), e obtém-se uma nova máscara binária, M_R. Em conjunto com a máscara binária, M_G, gerada após a filtragem de Gabor, têm-se meios de obter uma máscara para o sinal, M_{sinal}, através da intersecção das três máscaras: $M_{sinal} = M_s \cap M_R \cap M_G$.

A imagem complementar a M_{final} é a máscara do ruído $M_{ruido} = \overline{M}_{final}$. Com estas duas máscaras, pode-se proceder ao cálculo da S^2NR. Um exemplo de máscara está na Figura 3.7.

3.2.3.8 Cálculo da S^2NR

Dada a imagem do espectrograma inicial, imagem A, devemos gerar uma imagem com o sinal e a imagem complementar com o ruído. Para tanto, multiplicamos a imagem inicial pelas máscaras obtidas na seção anterior, M_{sinal} e M_{ruido}. Nas imagens resultantes, A_{sinal} e A_{ruido}, as colunas estão relacionadas com a variação temporal, enquanto as linhas têm relação com as freqüências. O cálculo do sinal da energia, em um instante de tempo j, é dado por:

$$S_{sinal}(j) = \sum_{i=0}^{N-1} [A_{sinal}(i,j)]^2, \tag{3.33}$$

sendo N a metade da duração da janela da FFT e $A_{sinal}(i,j)$ são os valores da imagem do sinal. De modo análogo, o energia do ruído é dada por:

$$S_{ruido}(j) = \sum_{i=0}^{N-1} [A_{ruido}(i,j)]^2. \tag{3.34}$$

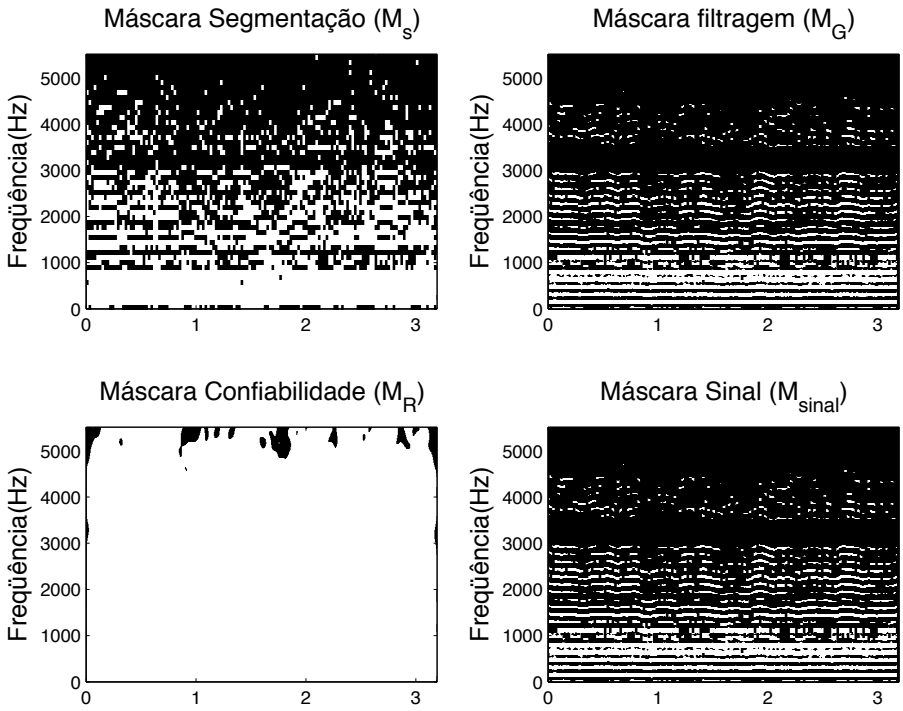

Figura 3.7: **Geração da máscara do sinal.** Esta figura mostra a composição das máscaras para a formação da máscara do sinal. A intersecção das máscaras, M_R, M_G e M_S, gera a máscara do sinal M_{sinal}. O eixo horizontal é o tempo, em segundos.

A S^2NR (*spectrographic signal-to-noise ratio*) é definida por:

$$S^2NR(j) = 10 \log_{10} \frac{S_{sinal}(j)}{S_{ruido}(j)}. \tag{3.35}$$

O gráfico de saída típico está mostrado na Figura 3.8, com o espectrograma correspondente.

3.3 Detalhes da implementação computacional

3.3.1 Parâmetros numéricos

Os parâmetros que podem ser modificados pelo usuário estão listados abaixo, junto com a seção referente onde estão definidos.

O treinamento dos parâmetros foi feito através da vogal /a/, com vogal sintética de relação sinal ruído controlada, de modo a minimizar o desvio em relação à SNR de referência.

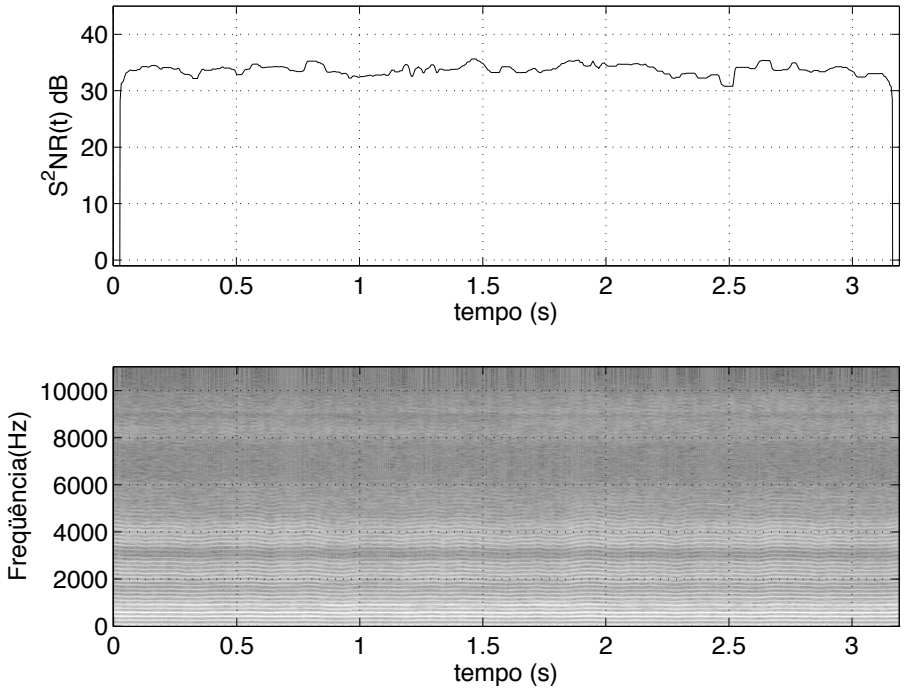

Figura 3.8: **Cálculo da S^2NR.** Gráfico da relação S^2NR ao longo do tempo.

Tabela 3.1: Parâmetros numéricos do algoritmo S^2NR:

Parâmetro	**Símbolo**	**Descrição**
N	N	Tamanho da janela da FFT, Seção 3.2.3.1
NOVERLAP	-	Sobreposição entre as janelas, Seção 3.2.3.1
blksze1	W	Tamanho do bloco da segmentação, Seção 3.2.3.3
blksze2	W_2	Tamanho do bloco para estimação de freqüência, Seção 3.2.3.5
thresh	σ_{th}	Limiar de segmentação, Seção 3.2.3.3
gradientsigma	σ_{grad}	Parâmetro da função gradiente, Seção 3.2.3.4
blocksigma	σ_{bloco}	Parâmetro das somas ponderadas, Seção 3.2.3.4
orientsmoothsigma	σ_{orient}	Parâmetro de suavização da orientação, Seção 3.2.3.4
windsze	-	Parâmetro para estimação de freqüência espacial, Seção 3.2.3.5
minWaveLength	λ_{max}	Comprimento de onda máximo, Seção 3.2.3.5
maxWaveLength	λ_{min}	Comprimento de onda mínimo, Seção 3.2.3.5
kx	σ_x	Desvio padrão do envelope gaussiano em x, Seção 3.2.3.6
ky	σ_y	Desvio padrão do envelope gaussiano em y, Seção 3.2.3.6
rthresh	σ_R	Limiar de segmentação de confiabilidade, Seção 3.2.3.7

Capítulo 4

Resultados

4.1 Introdução

Neste capítulo mostram-se os resultados obtidos na aplicação do método S^2NR. Em um primeiro momento, descreve-se os testes de calibração do algoritmo utilizando vogais sintéticas com relação sinal-ruído e perturbações na amplitude (*shimmer*) e freqüência (*jitter*) controladas.

Na seqüência, aplica-se o método à voz real, em amostras com diferentes graus de soprosidade, classificadas perceptivamente, e compara-se a classificação com as medições do método proposto.

Finalmente, discute-se a utilização do método em fala articulada.

4.2 Aplicação em voz sintetizada

Esta seção tem como objetivo mostrar a calibração do algoritmo para determinação da S^2NR frente a uma relação sinal-ruído pré-determinada. Este passo é importante para determinar se o método retorna números coerentes com os pré-estabelecidos, antes de se aplicar à voz real não controlada.

4.2.1 Metodologia

Para testar o algoritmo, foram geradas vogais sintéticas utilizando os algoritmos descritos no capítulo 2. Adicionou-se ruído de potência conhecida, conforme descrito em 2.3.3, e com *jitter* e *shimmer* controlados, conforme a expressões dadas em (2.3.1) e (2.3.2). Inicialmente, os efeitos de *jitter* e *shimmer* no cálculo foram estudados separadamente. Em seguida, foram incluídos simultaneamente. Variou-se também a freqüência fundamental sintetizada, com $f_o = 120\,Hz$ e $f_o = 220\,Hz$, simulando vozes masculinas e femininas respectivamente. Em todos os casos, as amostras tiveram duração de 3 segundos. As vogais sintéticas utilizadas

foram /a/, /i/ e /u/, obtidas conforme a Seção 2.2.4, com freqüência de amostragem de 22050 Hz. Os níveis de relação sinal-ruído de referência tiveram valores de 5 a 35 dB, com passos de 5 dB.

Para fins de comparação, fez-se também a medição utilizando o algoritmo de análise temporal discutido na Seção 3.2.1, que será referido por $SNR(t)$.

Os valores das medições apresentadas nos gráficos são o valor médio obtido ao longo do processamento da amostra. A saída do programa também oferece a resposta ao longo do tempo, como mostrado na Figura 3.8.

Os demais parâmetros de geração de espectrograma foram os seguintes:

1. Tamanho da janela $N = 1024$
2. Superposição (*Overlap*) das janelas 90 %

Os parâmetros numéricos (definidos na seção 3.3.1) do método utilizado nas simulações foram:

Tabela 4.1: Parâmetros numéricos utilizados nas simulações:

Parâmetro	**Valor**
blksze1	5
blksze2	16
thresh	10^{-1}
gradientsigma	1
blocksigma	5
orientsmoothsigma	5
windsze	5
minWaveLength	1
maxWaveLength	1
kx	0,15
ky	0,15
rthresh	0,6

Tais parâmetros foram determinados empiricamente de modo a melhorar a precisão das medidas tanto em casos de SNR baixa (evitando elevado *offset* nestas condições) quanto em condições de SNR alta.

Estes parâmetros foram utilizados em todas as simulações. Os parâmetros que mais influenciam a medição foram *rthresh*, o limiar de confiança, junto com *thresh*, limiar de geração de máscara, que gera a primeira estimativa para a posição das cristas.

4.2.2 Simulações com *jitter* controlado

4.2.2.1 Voz masculina sintética ($f_o = 120\,Hz$)

As amostras foram geradas com *jitter* controlado, de 0,0 a 3,0%, com passo de 0,5% entre elas. Estes valores extremos são aproximadamente os encontrados em voz humana (Vieira, 1997). Foram geradas 49 formas de onda por vogal.

As Figuras 4.1, 4.3 e 4.2 mostram os resultados com $f_o = 120\,Hz$, para as vogais /a/, /i/ e /u/. Nestas figuras, o gráfico à esquerda representa a S^2NR, enquanto à direita temos a $SNR(t)$. Os quadrados sólidos representam as medições efetuadas. Para facilitar a visualização, foram traçados segmentos de reta entres os pontos com mesmo nível de SNR de referência (SNR_{ref}), que é indicado à direita destas curvas.

Analisando o resultado obtido para a vogal /a/ mostrado na Figura 4.1, vê-se que o método da S^2NR apresenta espaçamento praticamente constante entre as curvas de SNR de referência. Nota-se, no entanto, que em condições extremas de *jitter* e com a SNR elevada, a medição decai ligeiramente, começando com aproximadamente $36\,dB$ com $0,0\%$ de *jitter*, chegando a aproximadamente $35\,dB$ com $3,0\%$ de perturbação.

No entanto, na comparação com a $SNR(t)$ tem-se: menor *offset*, pois vê-se a linha de referência de $5\,dB$ com média $6\,dB$, contra $9\,dB$ do método temporal, além de maior robustez com o aumento da perturbação, já que a medição no tempo satura-se em $25\,dB$ com *jitter* maior ou igual a $1,5\%$.

Para a vogal /u/, tem-se comportamento parecido com a vogal /a/, conforme mostra a Figura 4.2. A medida de S^2NR mostrou-se bastante estável até a referência de $25\,dB$, enquanto a medida baseado no tempo teve pior desempenho em níveis acima $20\,dB$. A medição da S^2NR saturou-se em torno de $31\,dB$, iniciando em $37\,dB$, na referência de $35\,dB$, enquanto $SNR(t)$ ficou em $25\,dB$, iniciando em $33\,dB$. Pode-se considerar o comportamento da S^2NR melhor neste exemplo.

A terceira vogal testada, vogal /i/, tem seu resultado mostrado na Figura 4.3. Ao contrário das vogais anteriores, a medição pelo S^2NR mostrou-se menos estável e robusta que a $SNR(t)$, principalmente com SNR de referência acima de $15\,dB$. A medida da $SNR(t)$ ficou estável até a faixa de referência de $20\,dB$. Já a faixa de referência de $25\,dB$ satura em $23\,dB$, e as faixas superiores em $25\,dB$, em linha com os exemplos anteriores, dentro das limitações do algoritmo para a vogal.

Até a conclusão desta dissertação, não se conseguiu explicar o comportamento limitado do S^2NR no caso da vogal /i/ sintética. Uma possibilidade levantada relaciona-se com as características espectrais da vogal /i/ sintetizada, mostradas na Figura 2.19. Aparentemente, a presença de um vale profundo entre F_1 e F_2 seria capaz de reduzir a amplitude dos harmônicos nesta faixa ao ponto de serem rejeitados na etapas de segmentação.

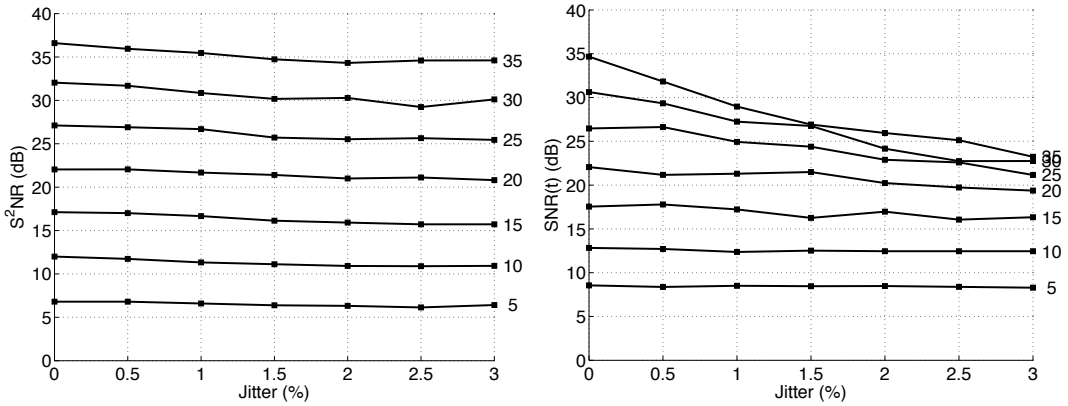

Figura 4.1: **Estimativas de SNR com variação do *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 120\,Hz$, vogal /a/ sintética.

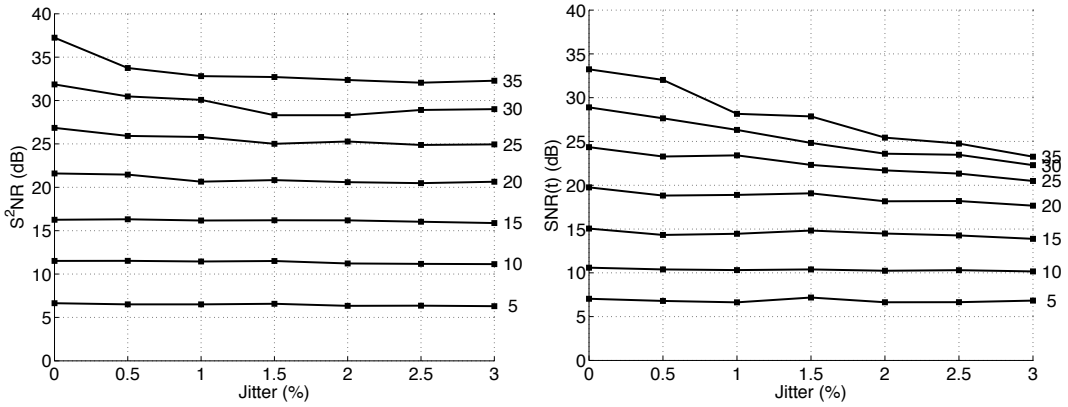

Figura 4.2: **Estimativas de SNR com variação do *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 120\,Hz$, vogal /u/ sintética.

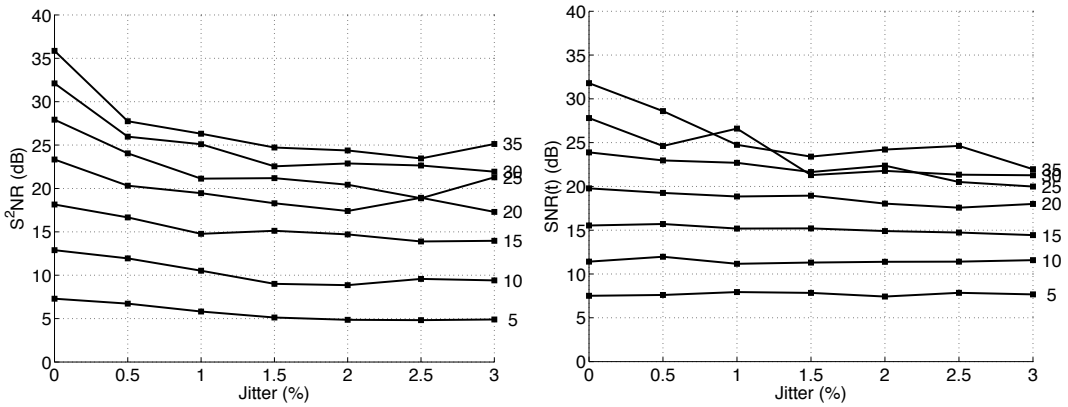

Figura 4.3: **Estimativas de SNR com variação do *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 120\,Hz$, vogal /i/ sintética.

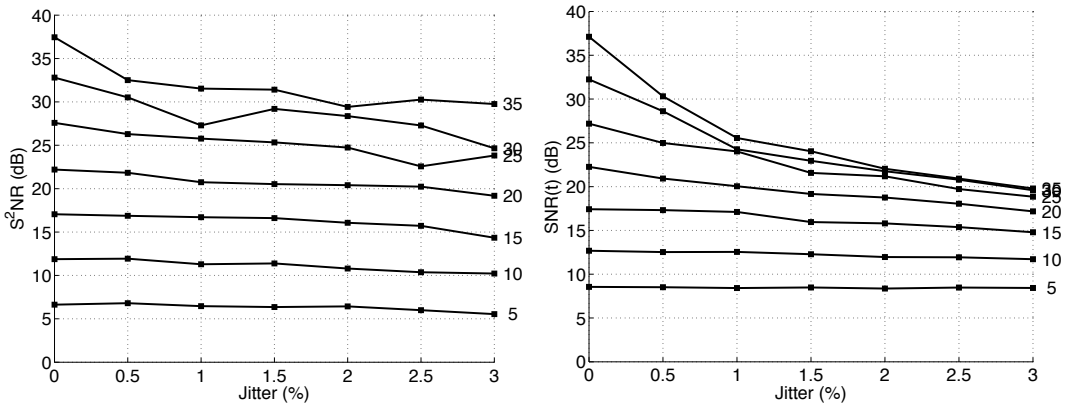

Figura 4.4: **Estimativas de SNR com variação do *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /a/ sintética.

4.2.2.2 Voz sintética feminina ($f_o = 220\,Hz$)

Com as mesmas configurações de trato vocal e parâmetros para o cálculo da S^2NR, repetiram-se as medições alterando a freqüência fundamental para $f_o = 220\,Hz$. Os resultados são apresentados nas Figuras 4.4, 4.5 e 4.6.

Para a vogal /a/, conforme a Figura 4.4, teve-se um comportamento robusto de S^2NR até os valores de referência de $25\,dB$, com *jitter* de $2,0\%$. Neste caso, a medição com S^2NR decaiu mais lentamente com o aumento de *jitter* e ruído.

Para a vogal /i/, o comportamento de ambas medições mostrou-se sensível ao espectro da vogal e a medida ficou abaixo da referência com *jitter* alto.

A vogal /u/, da Figura 4.6, mostrou a limitação do S^2NR acima de $20\,dB$. Em termos de robustez, a medida no tempo mostrou-se mais estável. No entanto, $SNR(t)$ apresentou grande polarização na medida (maior que $4\,dB$), mas foi robusta até a referência de $25\,dB$.

Percebem-se as limitações para o caso da voz feminina, principalmente nas condições de maior dificuldade do algoritmo, isto é, altos SNR e *jitter* (maior que $1,5\,\%$).

A vogais /a/ e /u/ apresentaram aplicação mais limitada no caso feminino. Uma hipótese que pode explicar a limitação é maior freqüência fundamental, pois uma melhor resolução no tempo seria necessária. Considerando o maior espaçamento do espectro, é possível abdicar de resolução na freqüência para obtermos melhor resolução temporal. Para testar isto, diminuiu-se a janela da FFT para N = 512. O resultado é apresentado na Figura 4.7. Nota-se melhora considerável quando comparado à medição apresentada na Figura 4.4. A medição com *jitter* de $3,0\%$ aproxima-se agora de $34\,dB$, contra $27\,dB$ apresentados anteriormente, além de maior estabilidade na medição. A vogais /i/ e /u/ com a janela diminuída estão nas Figuras 4.8 e 4.9 onde o melhor desempenho do algoritmo também é constatado. A saturação na vogal /i/ passa de $15\,dB$ para $25\,dB$ e a vogal /u/ apresenta resultado semelhante a vogal /a/.

Uma observação deve ser feita: um novo conjunto de dados foi gerado, portanto as amostras não são idênticas às do teste com $N = 1024$.

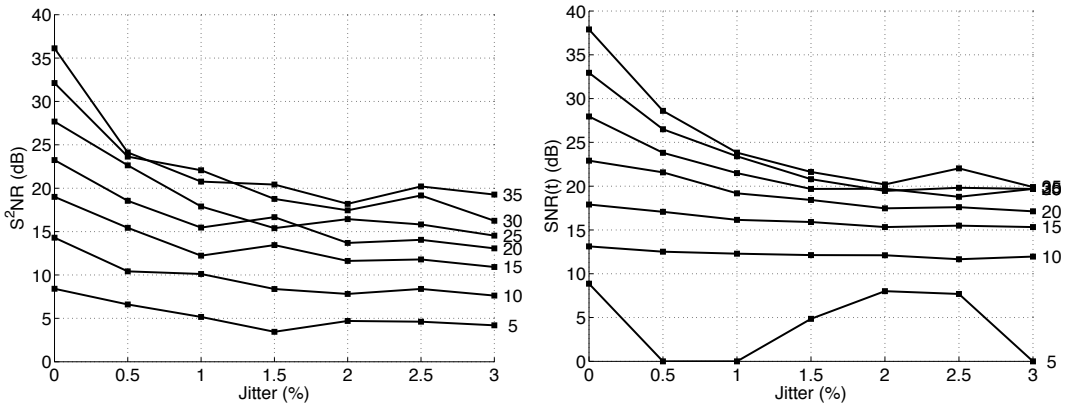

Figura 4.5: **Estimativas de SNR com variação do *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /i/ sintética. Não foi possível a medição da $SNR(t)$ na referência de $5\,dB$ com valores de *jitter* de $0,5\%$, $1,0\%$ e $3,0\%$.

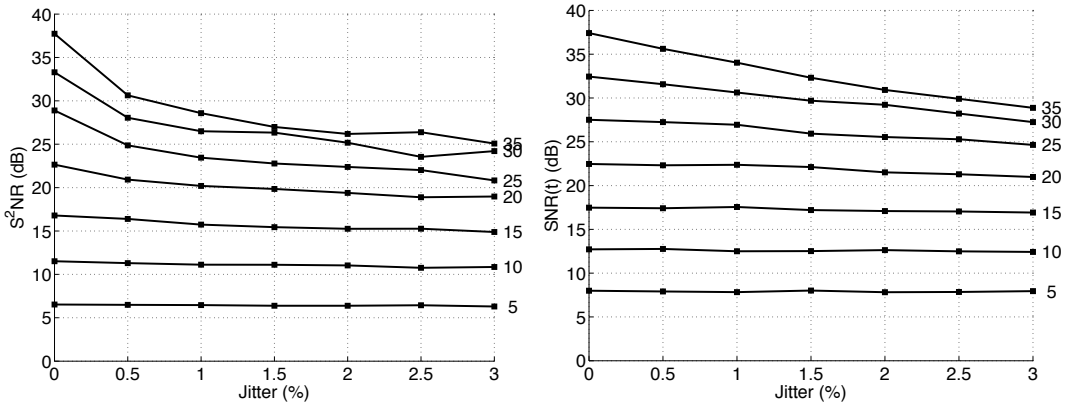

Figura 4.6: **Estimativas de SNR com variação do *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /u/ sintética.

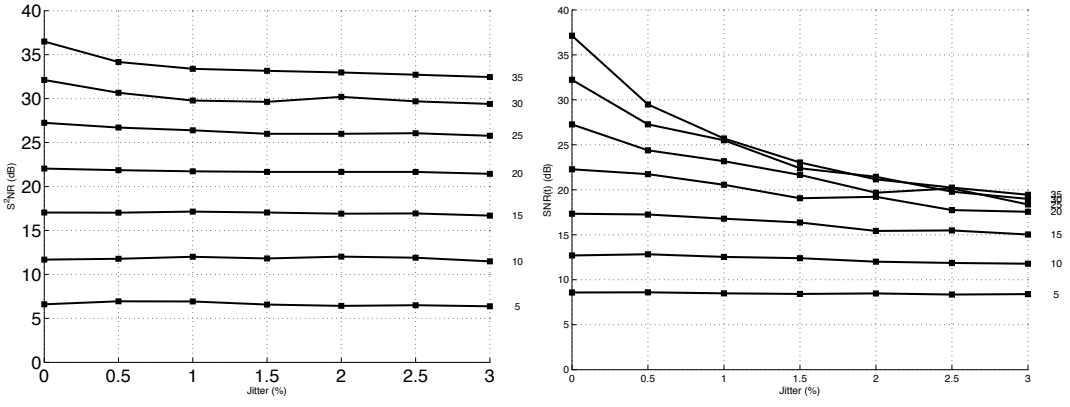

Figura 4.7: **Estimativas de SNR com variação do *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /a/ sintética, reduzindo janela para $N = 512$.

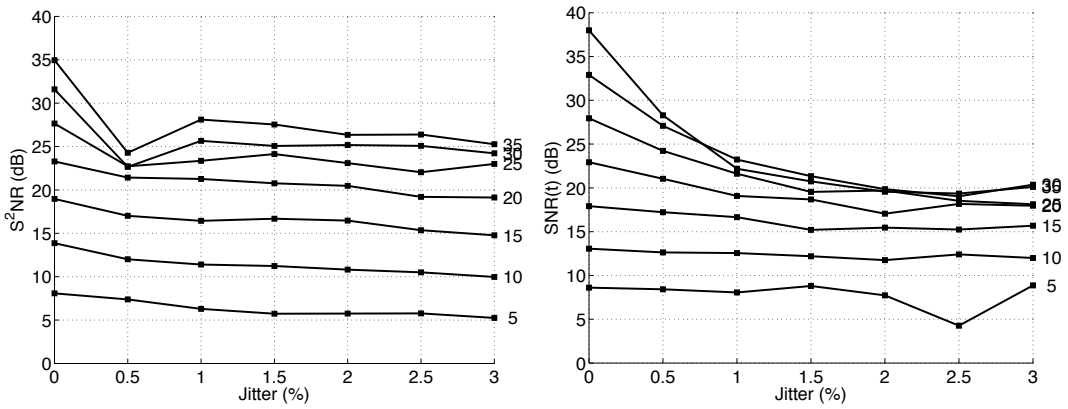

Figura 4.8: **Estimativas de SNR com variação do *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /i/ sintética, reduzindo janela para $N = 512$.

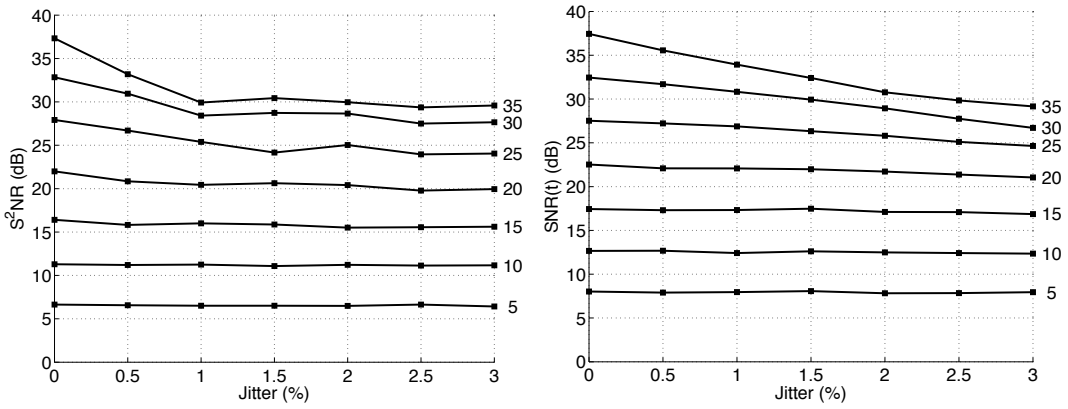

Figura 4.9: **Estimativas de SNR com variação do *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /u/ sintética, reduzindo janela para $N = 512$.

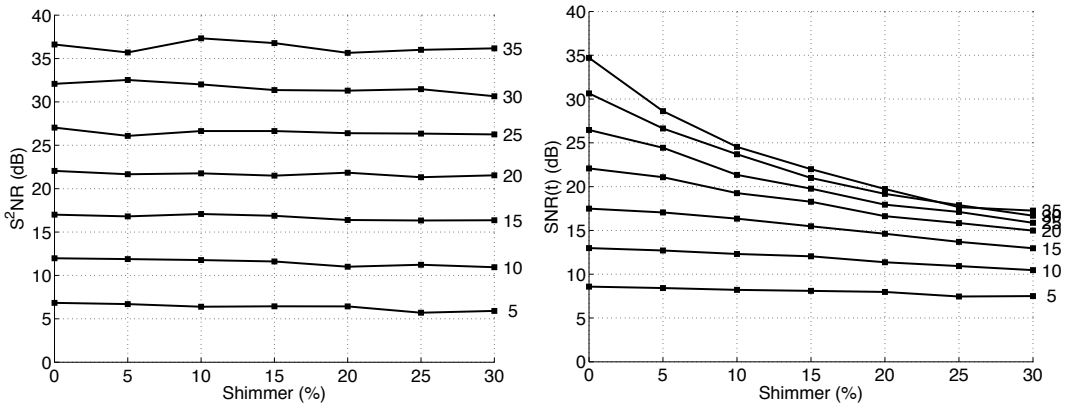

Figura 4.10: **Estimativas de SNR com variação do *shimmer*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 120\,Hz$, vogal /a/ sintética.

4.2.3 Simulações com *shimmer* controlado

4.2.3.1 Voz sintética masculina ($f_o = 120\,Hz$)

Nestas simulações, variou-se a intensidade de *shimmer* de 0,0% a 30%, com passos de 5%. Novamente, utilizaram-se as três vogais, /a/, /i/ e /u/, com $f_o = 120\,Hz$ e com duração de três segundos, com os mesmos parâmetros utilizados anteriormente, com janela de FFT $N = 1024$.

Os resultados são apresentados nas figuras 4.10, 4.11 e 4.12. Nota-se pelo gráficos a robustez do método do S^2NR no caso da perturbação em amplitude, principalmente nas vogais /a/ e /u/. A medida por análise temporal, $SNR(t)$, decai muito rapidamente com o aumento do *shimmer*, e nos três casos, mostrou-se inadequada para a medição.

A medida com S^2NR é robusta ao *shimmer* devido ao fato de não haver um sinal de referência médio, como no método temporal, que computa as variações de amplitude como componentes de ruído. Como estas variações estão sobrepostas ao sinal, elas surgem como variação na amplitude das componentes harmônicas. Portanto, é natural que o método da S^2NR tenha maior facilidade na discriminação das componentes de sinal e ruído.

Nota-se, um *offset* médio de até $2\,dB$, na medidas da S^2NR em /a/ e /u/, novamente, com a vogal /i/ sendo a menos estável das três.

4.2.3.2 Voz sintética feminina ($f_o = 220\,Hz$)

Da mesma forma, repetiram-se os testes para o caso de $f_o = 220\,Hz$.

Ao contrário do que aconteceu-se com o *jitter* na voz feminina, o comportamento do S^2NR foi robusto em relação ao *shimmer*. Com freqüência fundamental fixa, a limitação da resolução temporal não afeta a medição da S^2NR de forma significativa.

Já o algoritmo baseado no tempo, $SNR(t)$, mostrou-se bastante sensível à perturbação,

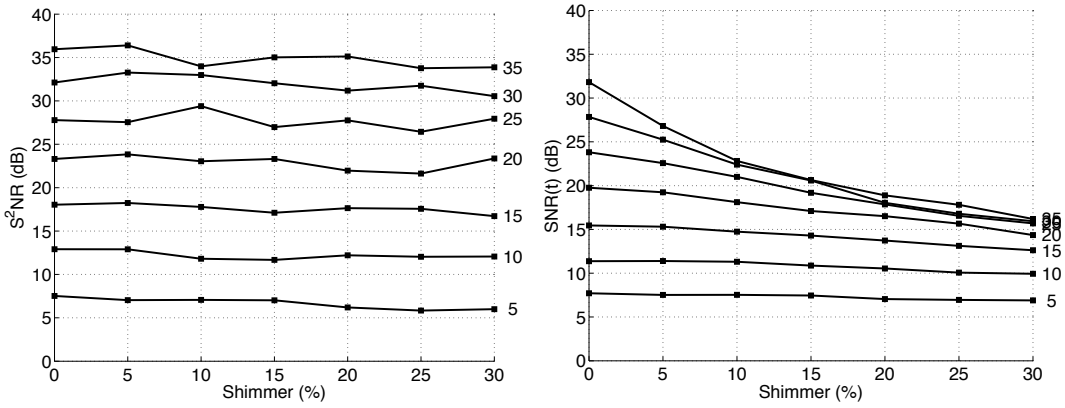

Figura 4.11: **Estimativas de SNR com variação do *shimmer*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 120\,Hz$, vogal /i/ sintética.

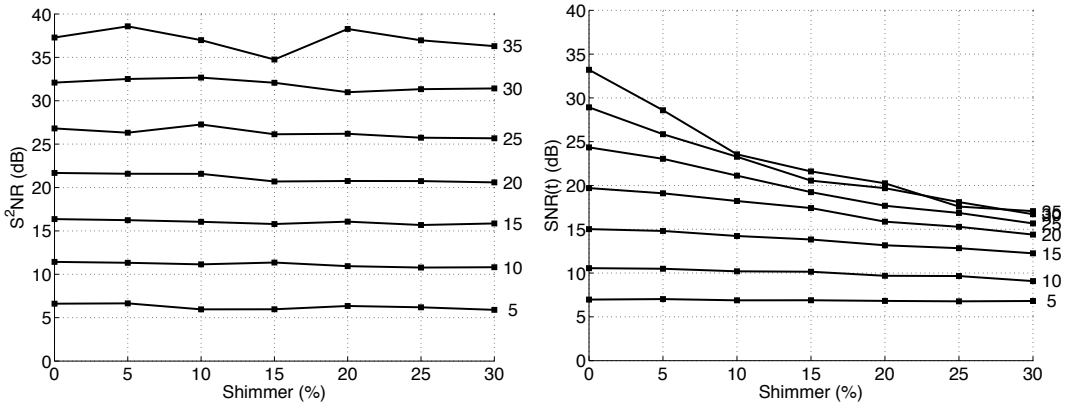

Figura 4.12: **Estimativas de SNR com variação do *shimmer*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 120\,Hz$, vogal /u/ sintética.

como anteriormente, com $f_o = 120\,Hz$. Os resultados estão nas Figuras 4.13, 4.15 e 4.14. Nota-se, para S^2NR, novamente o ligeiro *offset* nas vogais /a/ e /u/, que foram, contudo, mais estáveis como na medição masculina. A vogal /i/ foi novamente a mais crítica, afetando a medição nas amostras com SNR e perturbações altas (*shimmer* $> 15\%$, SNR em torno de $32\,dB$).

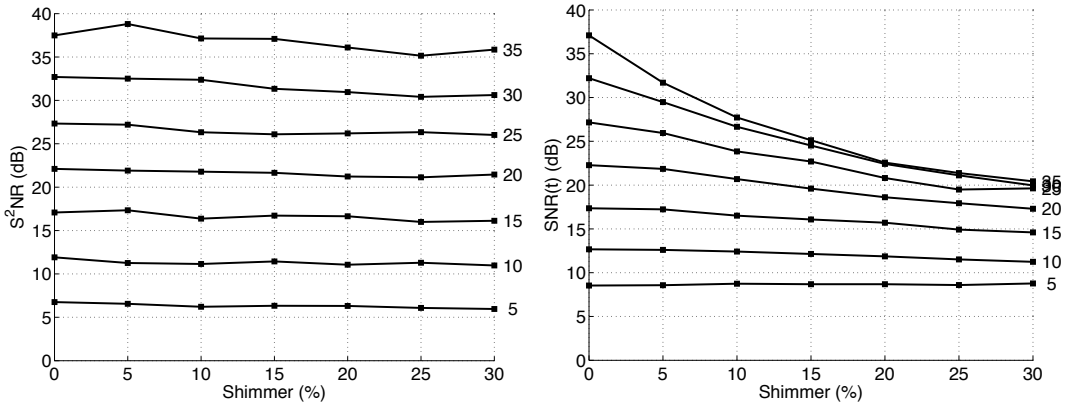

Figura 4.13: **Estimativas de SNR com variação do *shimmer*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /a/ sintética.

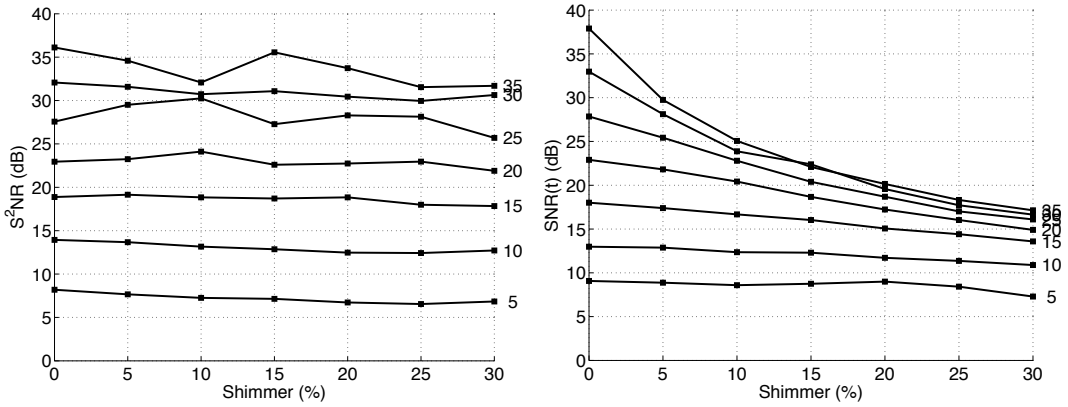

Figura 4.14: **Estimativas de SNR com variação do *shimmer*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /i/ sintética.

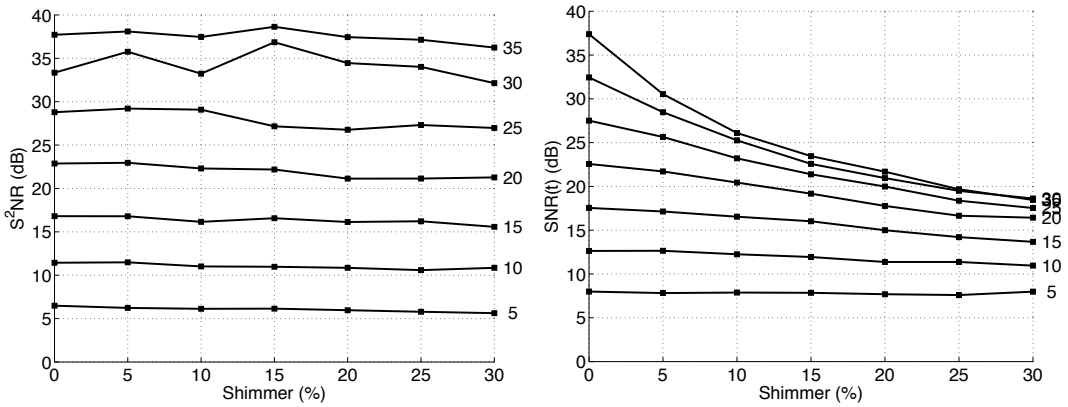

Figura 4.15: **Estimativas de SNR com variação do *shimmer*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /u/ sintética.

4.2.4 Simulações com *jitter* e *shimmer* controlados

Na voz real, as perturbações de amplitude e freqüência freqüentemente ocorrem simultaneamente. Nesta seção, são descritas tentativas para simular estas situações, onde adicionamos *jitter* e *shimmer* na mesma amostra.

4.2.4.1 Voz sintética masculina ($f_o = 120\,Hz$)

Simulando voz masculina, variou-se a relação sinal-ruído de $5\,dB$ a $30\,dB$, com passos de $5\,dB$, e *jitter* de $0,0$ a $3,0\,\%$, e *shimmer* de 0 a $30\,\%$. Os resultados são apresentados a seguir, com o nível de *jitter* nas ordenadas e o *shimmer* nas abscissas, estando a medida de SNR como a cor (escala em tons de cinza, com tom mais claro para SNR mais alta). Novamente, para fins de comparação apresentamos o algoritmo baseado em análise temporal. A vogal estudada foi /a/ e cada figura tem um nível fixo de SNR de referência (SNR_{ref}).

Os testes com *jitter* e *shimmer* combinados são consistentes com os testes de perturbação individuais, no caso do S^2NR e, como nos casos anteriores, as limitações do algoritmo surgiram com SNR alta e perturbação em freqüência elevadas. *A presença simultânea das perturbações não limitou a capacidade de medição do algoritmo.*

Para vogal /a/, o algoritmo se mostrou bastante robusto e apresentou desempenho melhor que o algoritmo baseado na análise temporal da forma de onda, fato que pode ser concluído através da observação da variação de cor nas imagens, que é mais homogênea na S^2NR. Pode-se tomar como exemplo a Figura 4.21. Notam-se cores claras em quase toda área do gráfico, exceto na região onde *shimmer* e *jitter* estão próximos de 30% e 3% respectivamente. No caso do $SNR(t)$ a área é clara próxima da região sem perturbação e escurece com o aumento do *shimmer* e mais discretamente com o aumento do $jitter$.

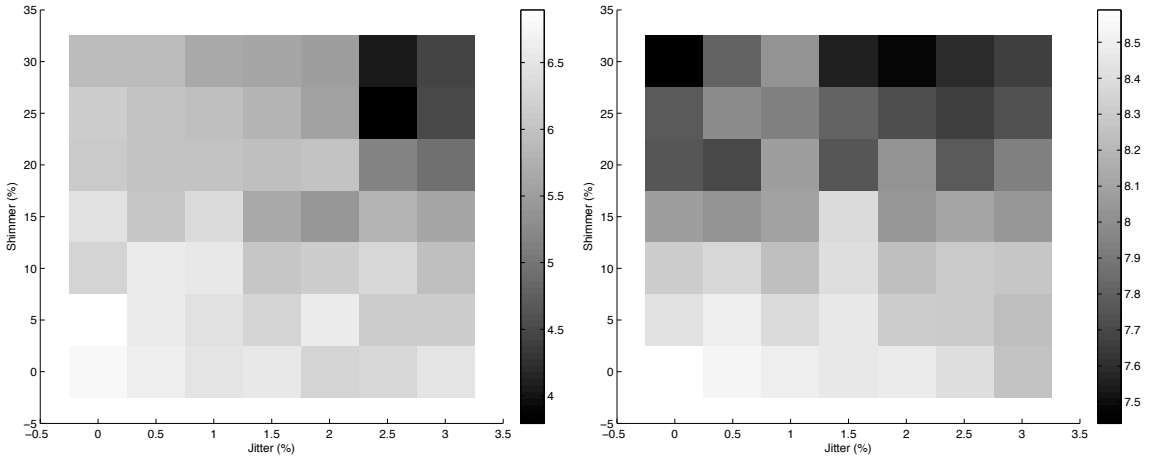

Figura 4.16: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR (esquerda) e $SNR(t)$ (direita), com $f_o = 120\,Hz$, vogal /a/ sintética, $SNR_{ref} = 5\,dB$.

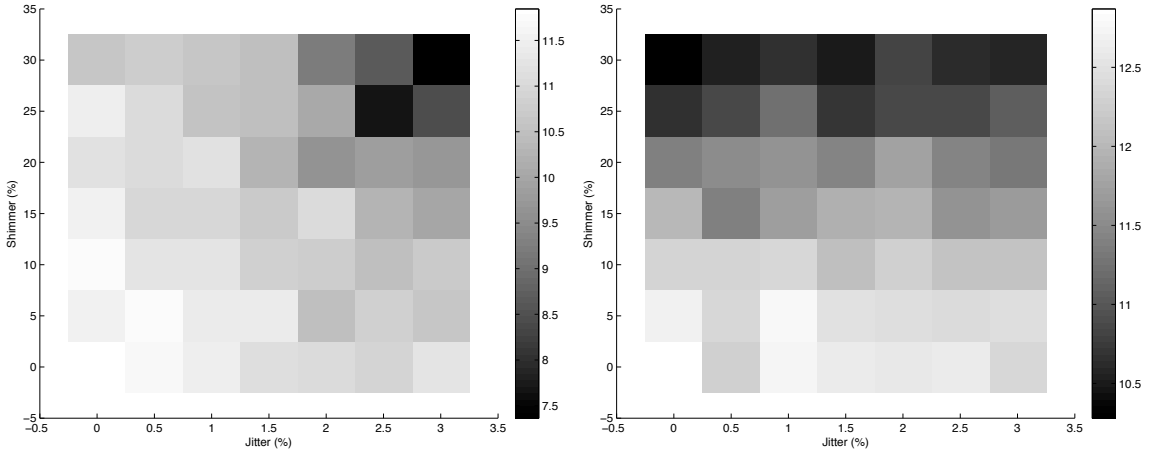

Figura 4.17: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 120\,Hz$, vogal /a/ sintética, $SNR_{ref} = 10\,dB$.

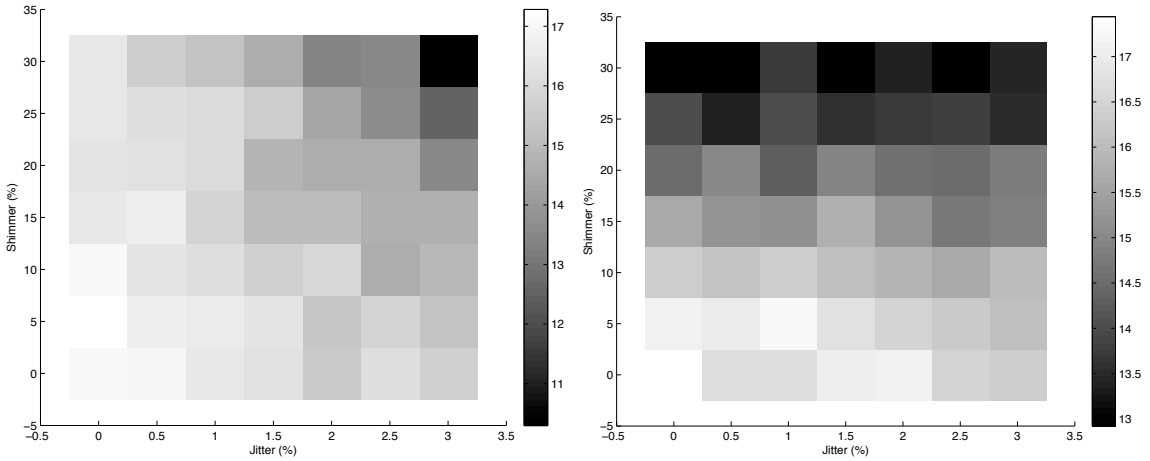

Figura 4.18: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 120\,Hz$, vogal /a/ sintética, $SNR_{ref} = 15\,dB$.

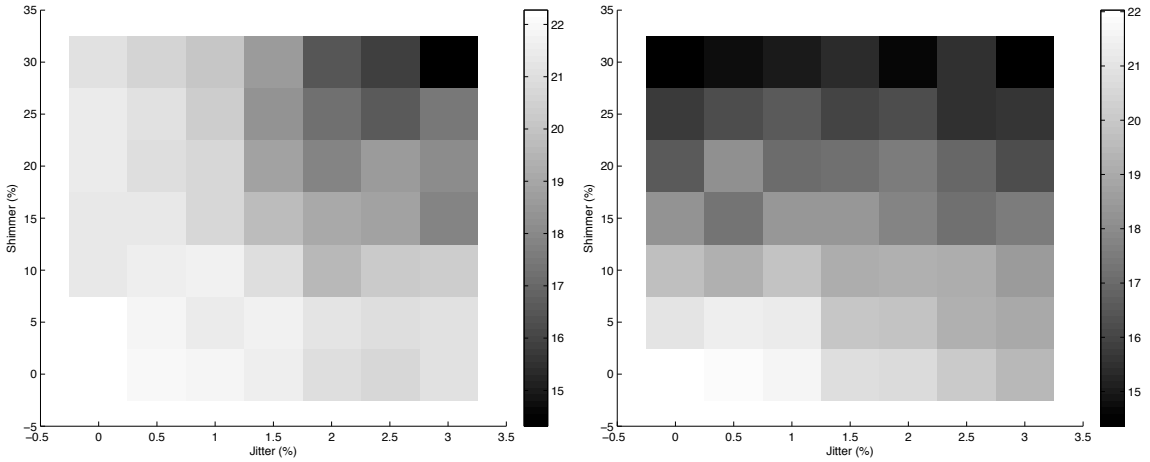

Figura 4.19: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 120\,Hz$, vogal /a/ sintética, $SNR_{ref} = 20\,dB$.

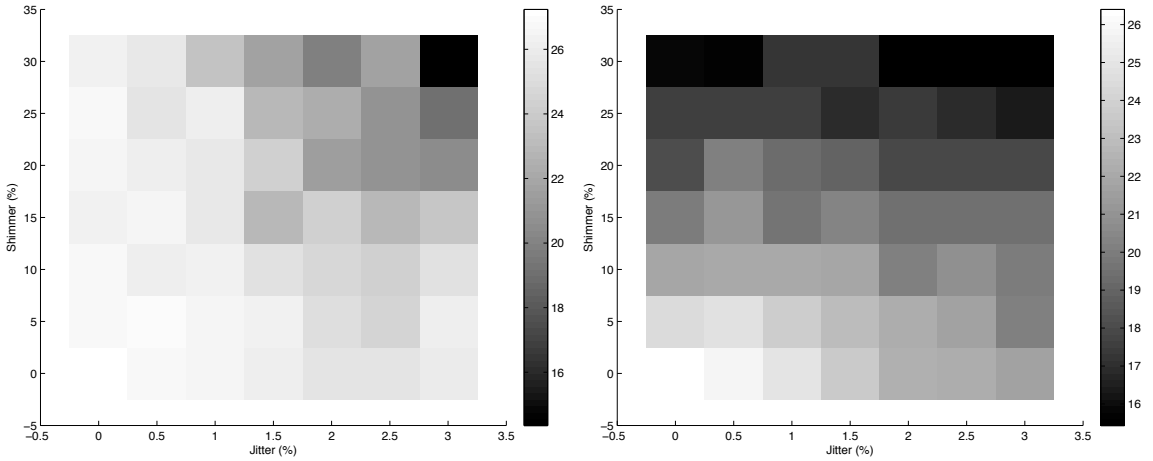

Figura 4.20: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 120\,Hz$, vogal /a/ sintética, $SNR_{ref} = 25\,dB$.

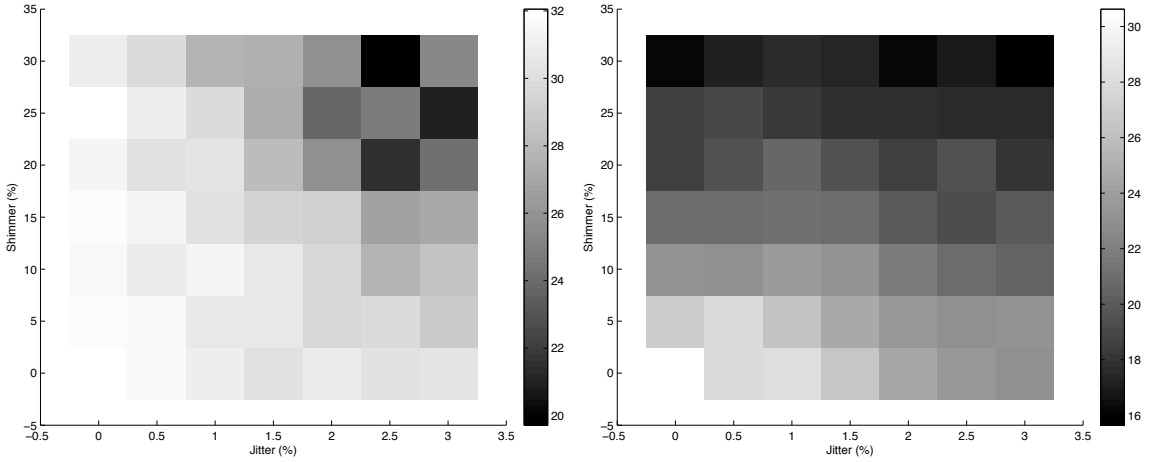

Figura 4.21: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 120\,Hz$, vogal /a/ sintética, $SNR_{ref} = 30\,dB$.

4.2.4.2 Voz sintética feminina ($f_o = 220\,Hz$)

Utilizando o mesmo procedimento adotado anteriormente, obtiveram-se os resultados para voz feminina, novamente para a vogal /a/. Os resultados são coerentes com os testes individuais de perturbação, mostrando a limitação do algoritmo com SNR alta e perturbarção em freqüência. A uniformidade na resposta foi maior para S^2NR, assim como no caso de voz masculina.

Novamente, utilizaram-se as mesmas janelas de $N = 1024$ e $N = 512$. Notam-se os efeitos na medição quando ocorre maior perturbação em freqüência, logo a robustez é menor que no caso masculino, principalmente nos casos onde a referência (SNR_{ref}) é mais alta para janela $N = 1024$. No caso de $N = 512$, o comportamento foi parecido com o caso masculino com a janela maior.

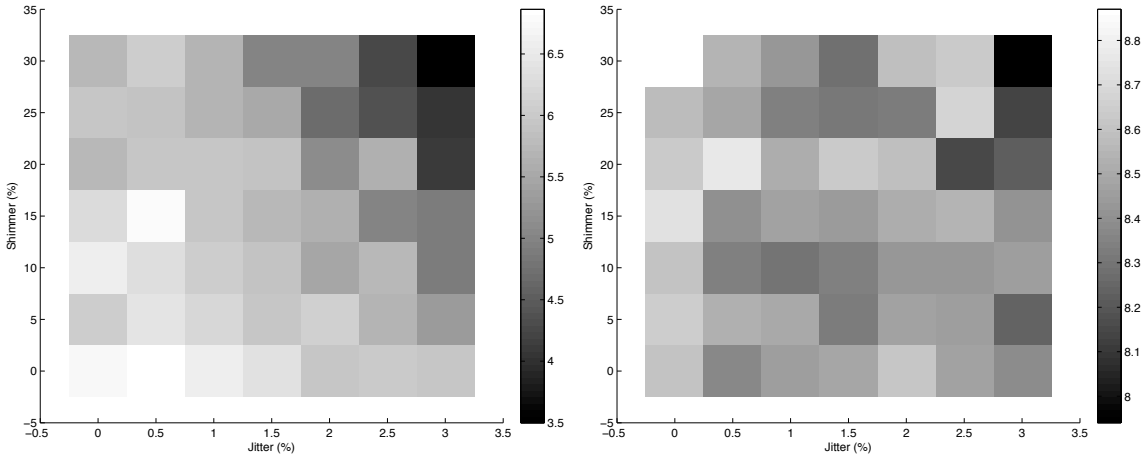

Figura 4.22: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /a/ sintética, $SNR_{ref} = 5\,dB$, $N = 1024$.

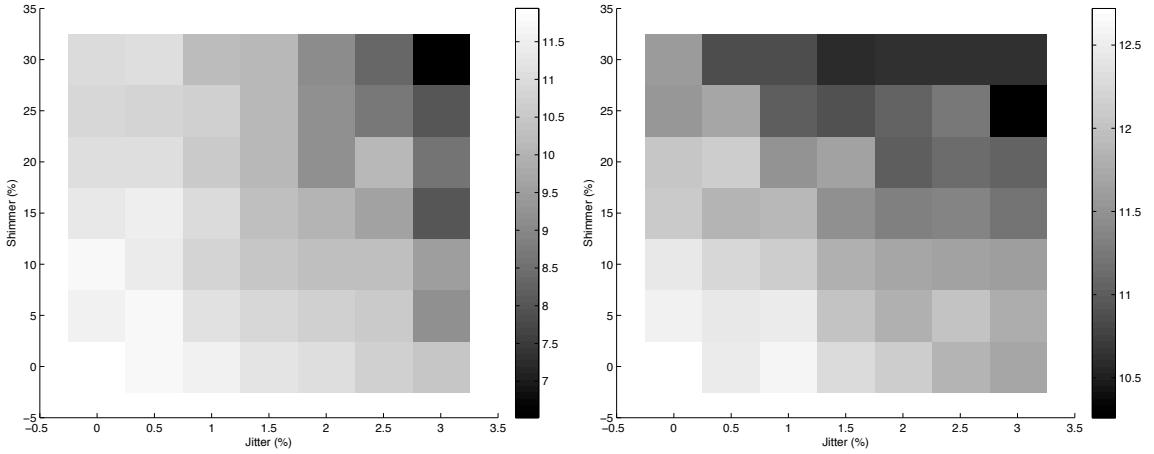

Figura 4.23: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /a/ sintética, $SNR_{ref} = 10\,dB$, $N = 1024$.

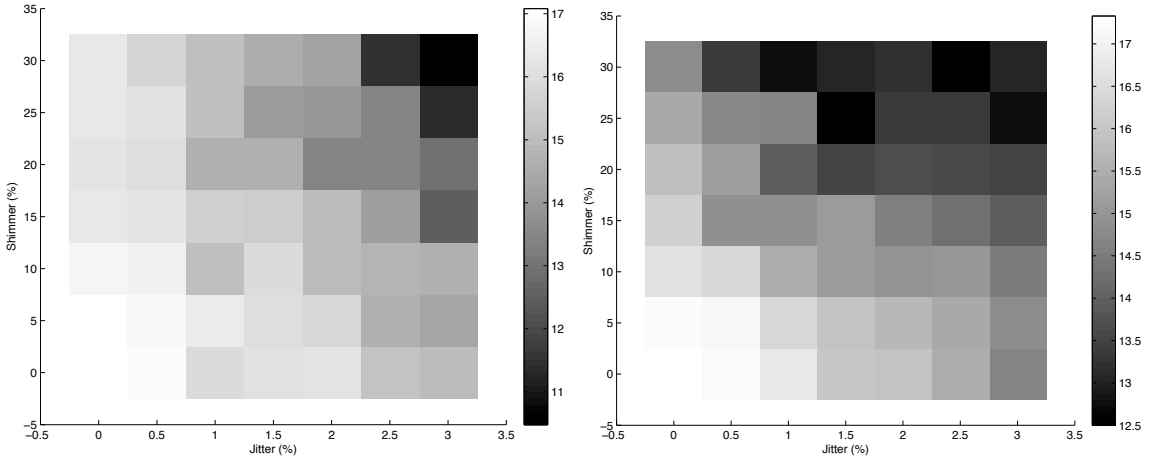

Figura 4.24: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /a/ sintética, $SNR_{ref} = 15\,dB$, $N = 1024$.

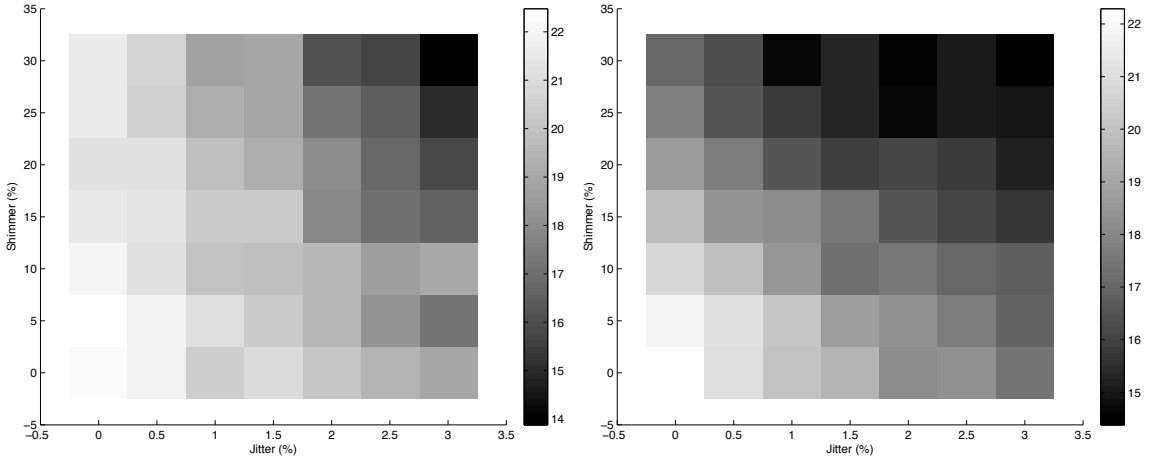

Figura 4.25: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /a/ sintética, $SNR_{ref} = 20\,dB$, $N = 1024$.

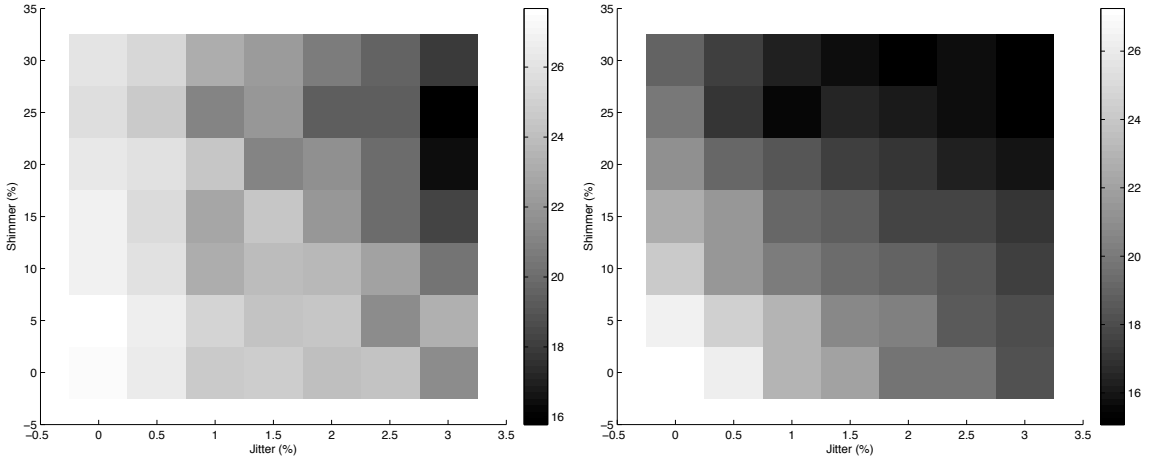

Figura 4.26: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /a/ sintética, $SNR_{ref} = 25\,dB$, $N = 1024$.

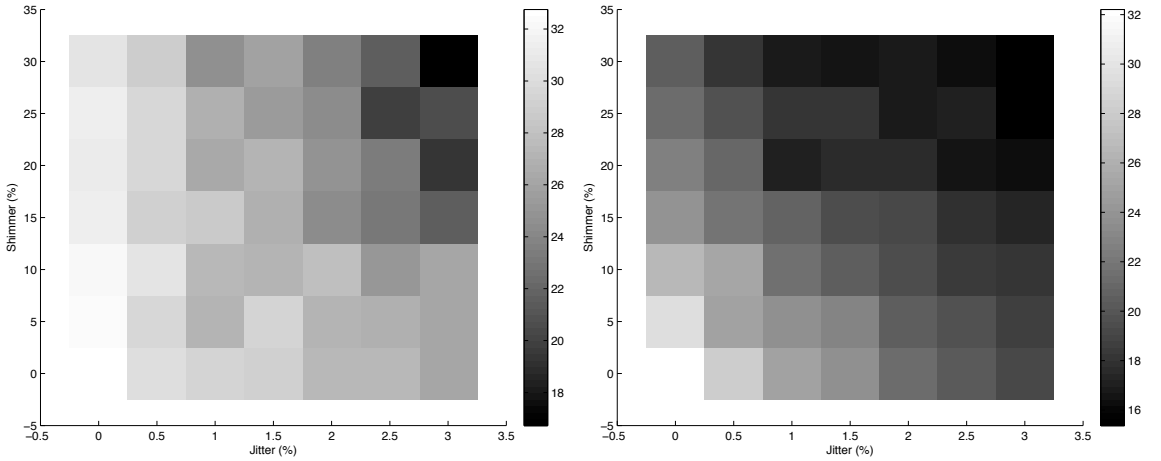

Figura 4.27: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /a/ sintética, $SNR_{ref} = 30\,dB$, $N = 1024$.

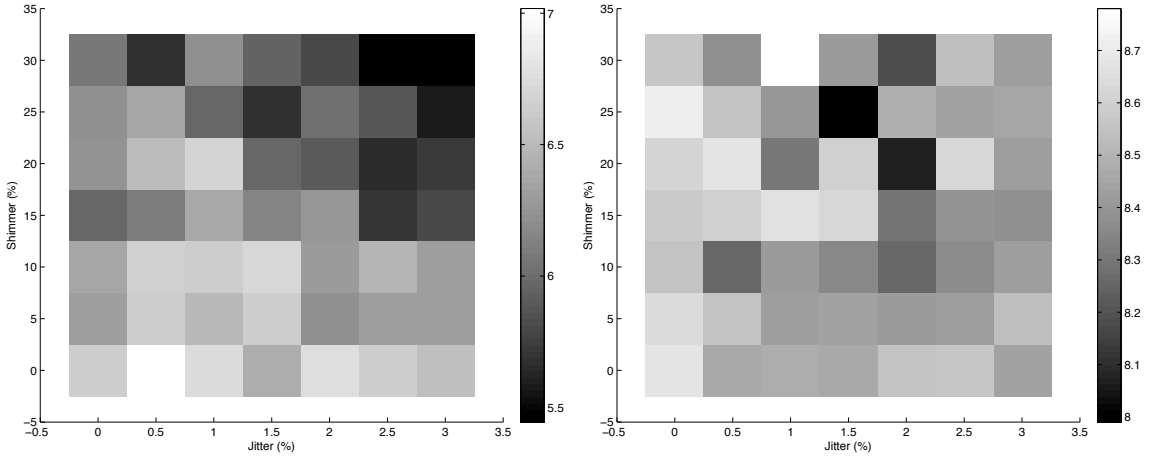

Figura 4.28: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /a/ sintética, $SNR_{ref} = 5\,dB$, $N = 512$.

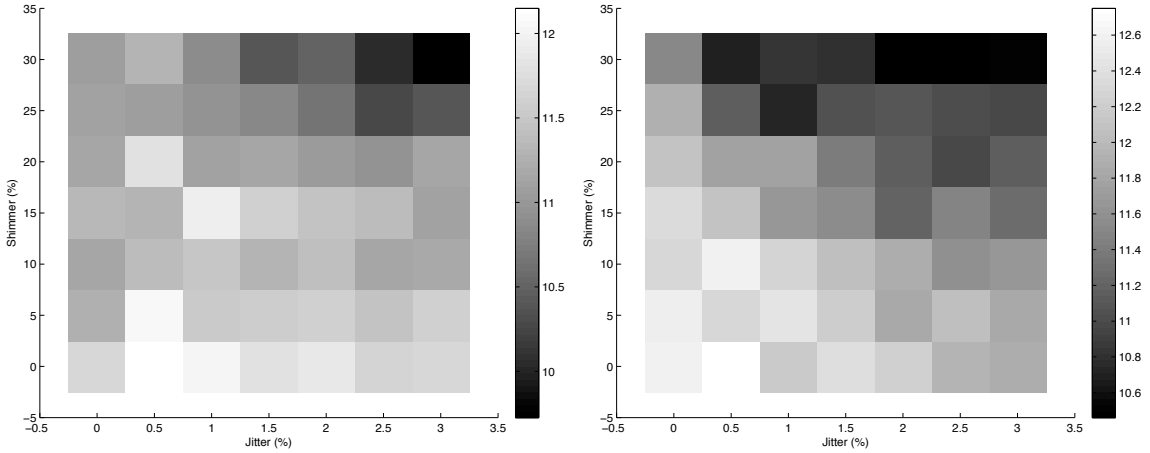

Figura 4.29: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /a/ sintética, $SNR_{ref} = 10\,dB$, $N = 512$.

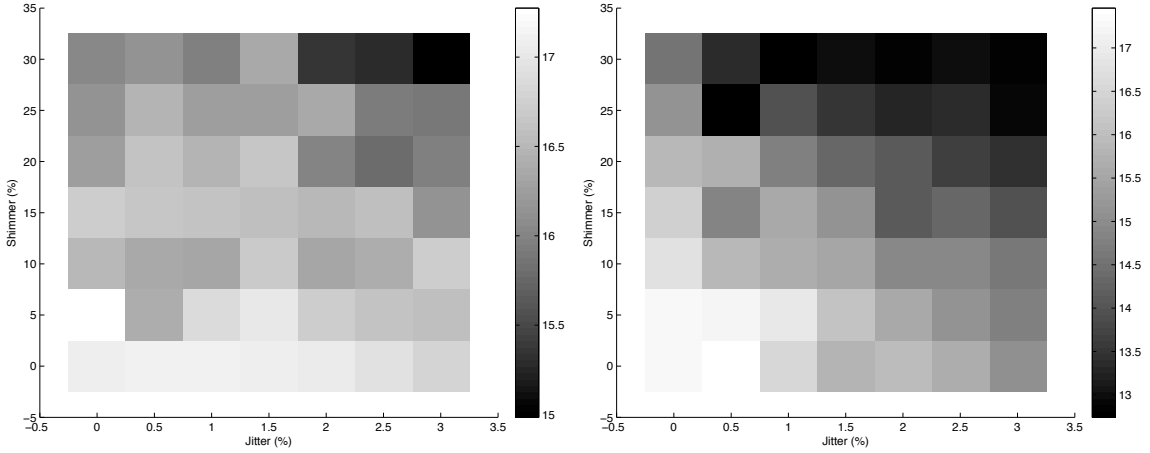

Figura 4.30: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /a/ sintética, $SNR_{ref} = 15\,dB$, $N = 512$.

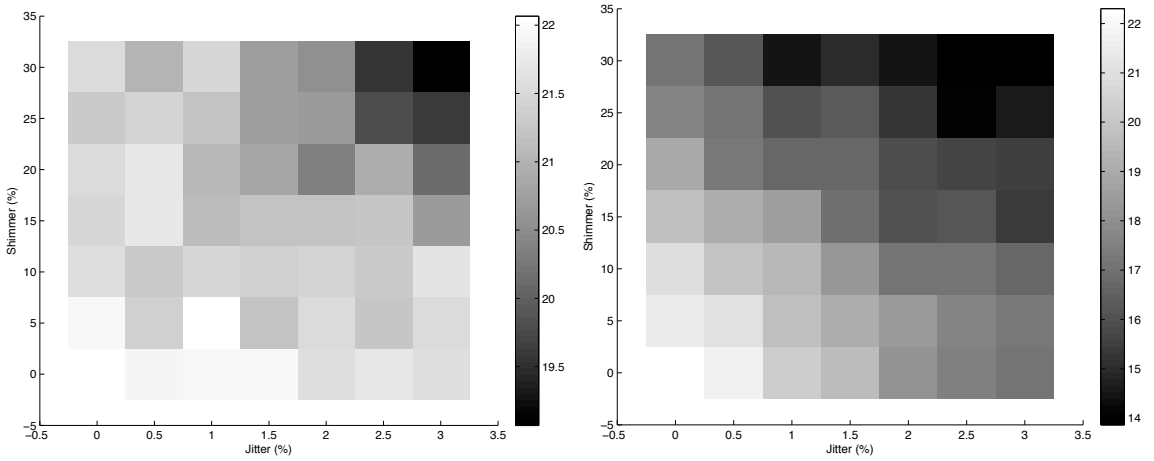

Figura 4.31: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /a/ sintética, $SNR_{ref} = 20\,dB$, $N = 512$.

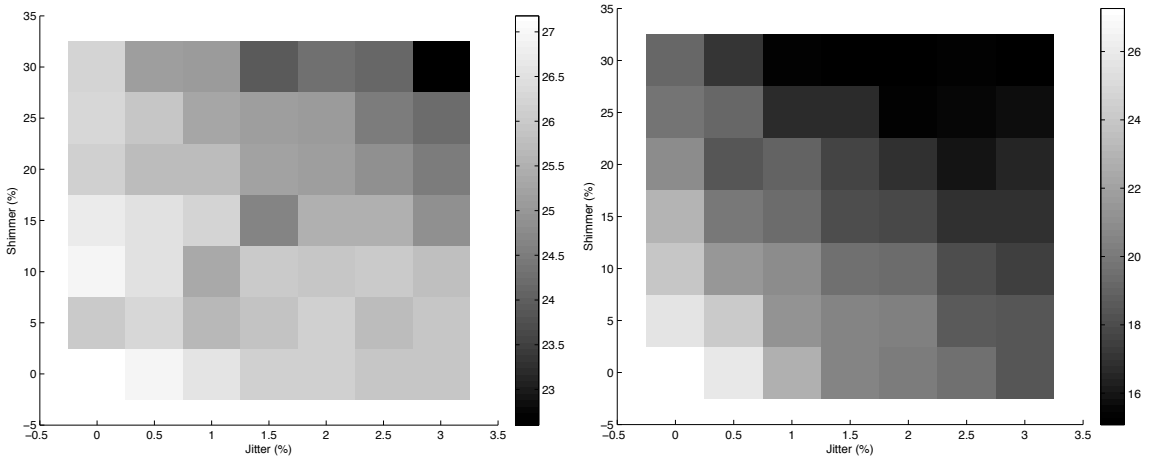

Figura 4.32: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /a/ sintética, $SNR_{ref} = 25\,dB$, $N = 512$.

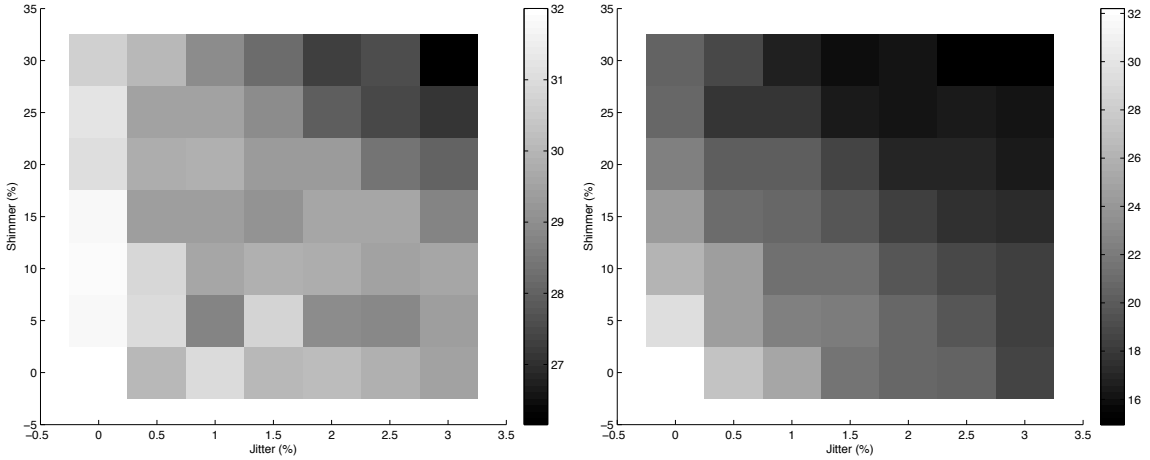

Figura 4.33: **Estimativas de SNR com variação do *shimmer* e *jitter*.** Comparativo das medidas de S^2NR e $SNR(t)$, com $f_o = 220\,Hz$, vogal /a/ sintética, $SNR_{ref} = 30\,dB$, $N = 512$.

4.3 Resultados com voz real

4.3.1 Metodologia

Dada a associação da soprosidade (Yumoto et al., 1984) com a relação sinal ruído da voz, avaliou-se o potencial de utilização da S^2NR como um medidor objetivo para o grau de soprosidade na voz humana. Para testar esta possibilidade, foram utilizadas 21 amostras da vogal /a/ de uma base de dados maior (Vieira, 1997). Estas amostras foram classificadas perceptivamente utilizando uma escala de 7 pontos (3 amostras por grau), 0-3, com intermediários (0,0; 0,5; 1,0; 1,5; 2,0; 2,5; 3,0), efetuamos o cálculo da S^2NR e da $SNR(t)$. A classificação foi realizada por uma dupla de estudantes de Fonoaudiologia da UFMG, ordenando comparativamente (por consenso) as amostras por grau de soprosidade.

4.3.2 Medidas

A Figura 4.34 mostra o resultado para ambos algoritmos. Constata-se a relação monotônica entre a SNR e o índice de soprosidade. Através de uma linearização, mostra-se esta tendência, ressaltada pela linha cheia no gráfico de ambos. Não se pode afirmar, no entanto, *a priori*, uma relação linear entre a soprosidade e a relação sinal ruído.

O algoritmo do S^2NR indicou uma separação de $-10,9\,dB$ por grau de soprosidade. Já o algoritmo baseado no tempo $-4,9\,dB$ por grau. Isto se deve à maior capacidade do algoritmo S^2NR de medir em condições variadas de perturbação, que normalmente surgem na voz real, mesmo em condições normais.

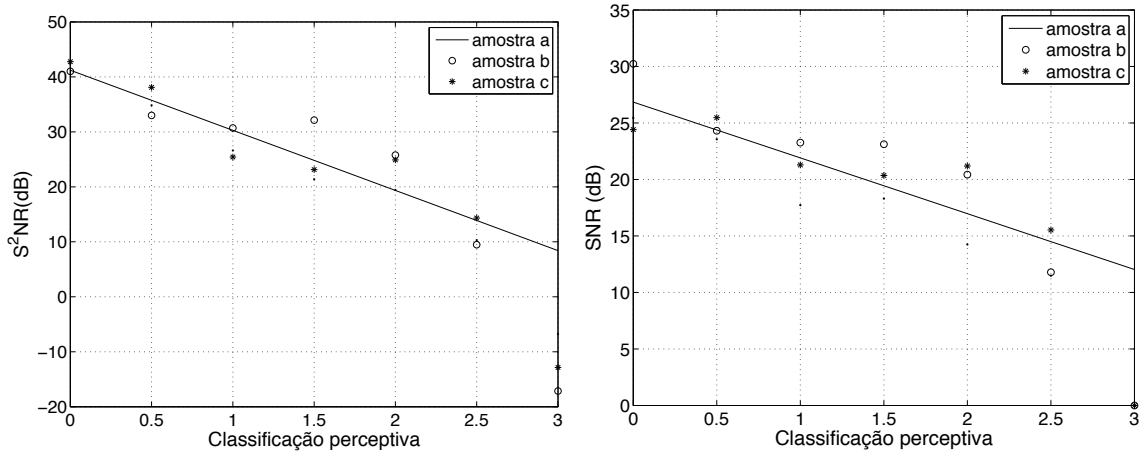

Figura 4.34: **Estimativas de SNR com voz real.** Comparativo das medidas de S^2NR e $SNR(t)$, com voz real, vogal /a/. Linha cheia representa a linearização dos dados obtidos, em função da classificação perceptiva de soprosidade. As escalas não são as mesmas nos dois casos.

A faixa dinâmica da S^2NR também foi maior, conseguindo abranger de aproximadamente $43\,dB$ a $-17\,dB$. Os casos com grau três são mostrados, mas não foram computados na linearização. Para o algoritmo baseado na análise temporal, a medição neste grau de soprosidade não foi possível, no entanto, atribuiu-se valor nulo à medição.

4.3.3 Demonstração de medidas selecionadas

Nesta seção serão descritas algumas características interessantes do método, quando aplicado à voz real. Na Figura 4.35, onde tem-se uma voz classificada como soprosidade grau 0, notam-se linhas de espectro bem definidas em quase toda banda passante. Na S^2NR a medida encontra-se estável, ligeiramente acima de $40\,dB$ e sem interrupções. A medida de $SNR(t)$ não foi capaz de medir em todo instante, mas apenas onde tem a capacidade de segmentar os períodos. Por isso, ficaram grandes períodos de tempo sem efetuar medição, como no intervalo 1,0 s até 2,5 s. Uma inspeção do espectrograma mostra que nos segmentos onde $SNR(t)$ falhou, houve variações no sinal que levaram à não detecção do vozeamento.

No caso da Figura 4.36, o método $SNR(t)$ foi capaz de acompanhar toda a duração da amostra, mas como no item anterior, a magnitude foi bem inferior à aferida pela S^2NR. Nos casos reais, sempre existe alguma flutuação simultânea em amplitude e freqüência, o que pode explicar a diferença entre as medidas, que se aproximavam nos casos de voz sintética.

Tomando um caso com maior grau de soprosidade (1,5), mostrado na figura 4.37, nota-se já pelo espectrograma regiões onde a estrutura harmônica perde a resolução. Tais flutuações podem ser observadas no gráfico da S^2NR. Por exemplo, em instantes logo após $t = 1\,s$, onde o valor medido cai de quase $37\,dB$ para próximo de $20\,dB$, nota-se o surgimento de ruído entre os harmônicos, além do desaparecimento das linhas espectrais em freqüências

próximas a $2\,kHz$. Esta flutuação ocorre até aproximadamente $t = 4\,s$. Os sub-harmônicos são detectados em $t \approx 6s$. A soprosidade neste caso decorre de um escape de ar, perceptível acusticamente, provável causa da flutuação citada anteriormente.

A Figura 4.38 é um exemplo marcante das limitações da $SNR(t)$, uma voz com muito escape de ar. No exemplo, o algoritmo que analisa a forma de onda consegue demarcar apenas dois pequenos trechos, enquanto S^2NR mede em toda extensão da amostra. As linhas harmônicas são detectáveis, mesmo quando não são contínuas por longos períodos, mostrando a utilidade do método em casos onde a detecção da freqüência fundamental é difícil.

Em uma amostra com nível mais severo de soprosidade, tem-se a Figura 4.39, onde novamente vemos a capacidade de medição contínua. Têm-se, neste exemplo, mais segmentos onde ocorre a degradação das linhas espectrais e conseqüente diminuição da S^2NR. Isto ocorre, por exemplo, próximo de $t = 1,5\,s$ e $t = 2,5\,s$.

O mesmo fenômeno ocorre na Figura 4.40, nas proximidades de $t = 0,8\,s$, $t = 1,7\,s$ e $t = 2,4\,s$. Nestes casos, a perda de energia harmônica é ainda mais acentuada, e a medida foi afetada pelo filtro utilizado para suavizar o sinal de S^2NR no tempo.

A Figura 4.41 mostra um caso com soprosidade grau 3,0. A $SNR(t)$ não apresentou medição em nenhum instante de tempo. O método da S^2NR apresenta medição contínua e percebe-se o aumento da medição nos instantes em que surgem linhas harmônicas definidas, como próximo de $t = 0,5\,s$.

Mostra-se assim, com estes exemplos a capacidade do método e sua aplicação em voz real com grande grau de perturbação.

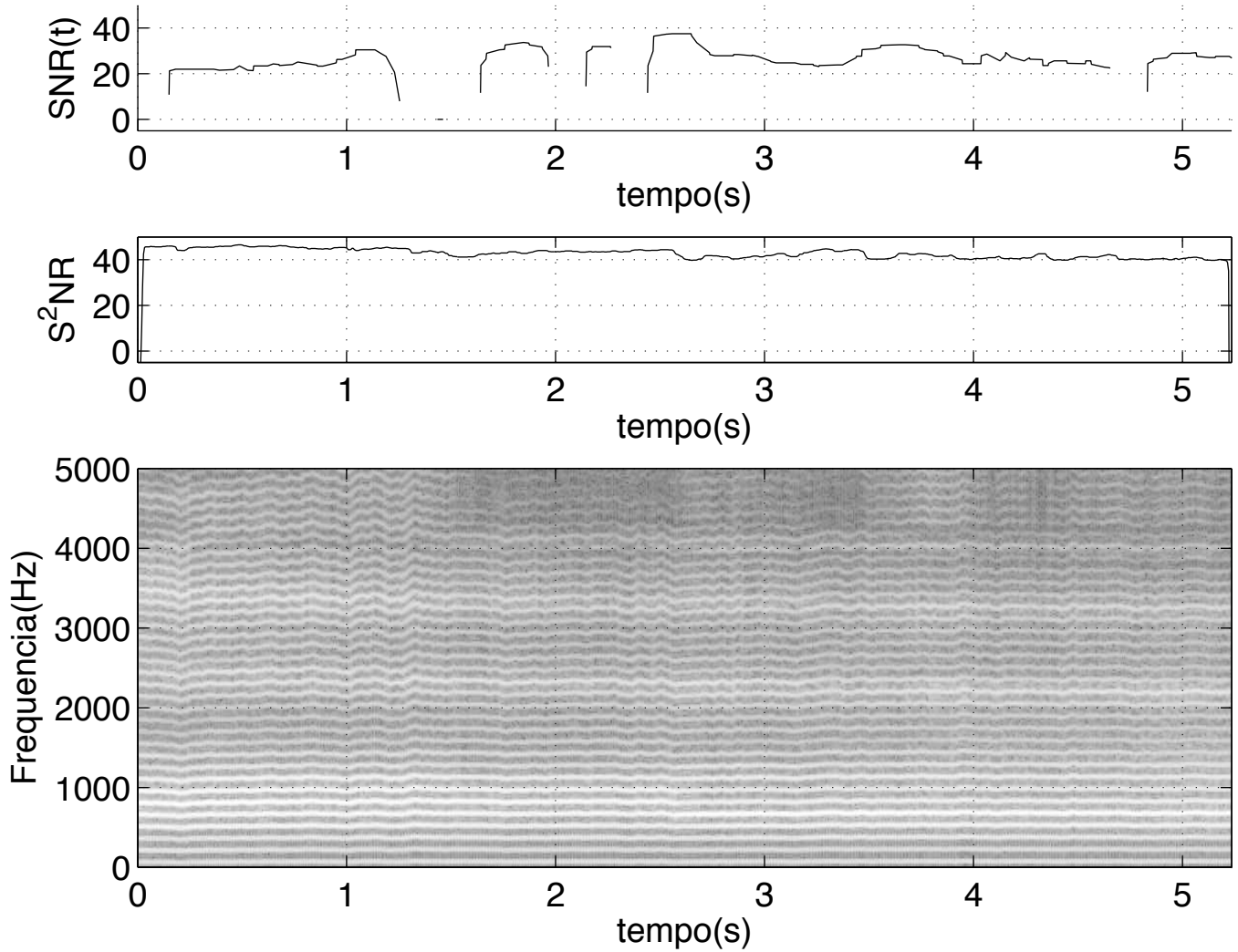

Figura 4.35: **Estimativas de SNR com voz real.** Comparativo das medidas de S^2NR e $SNR(t)$, com voz real, vogal /a/. Voz classificada subjetivamente como soprosidade grau 0 (ausente).

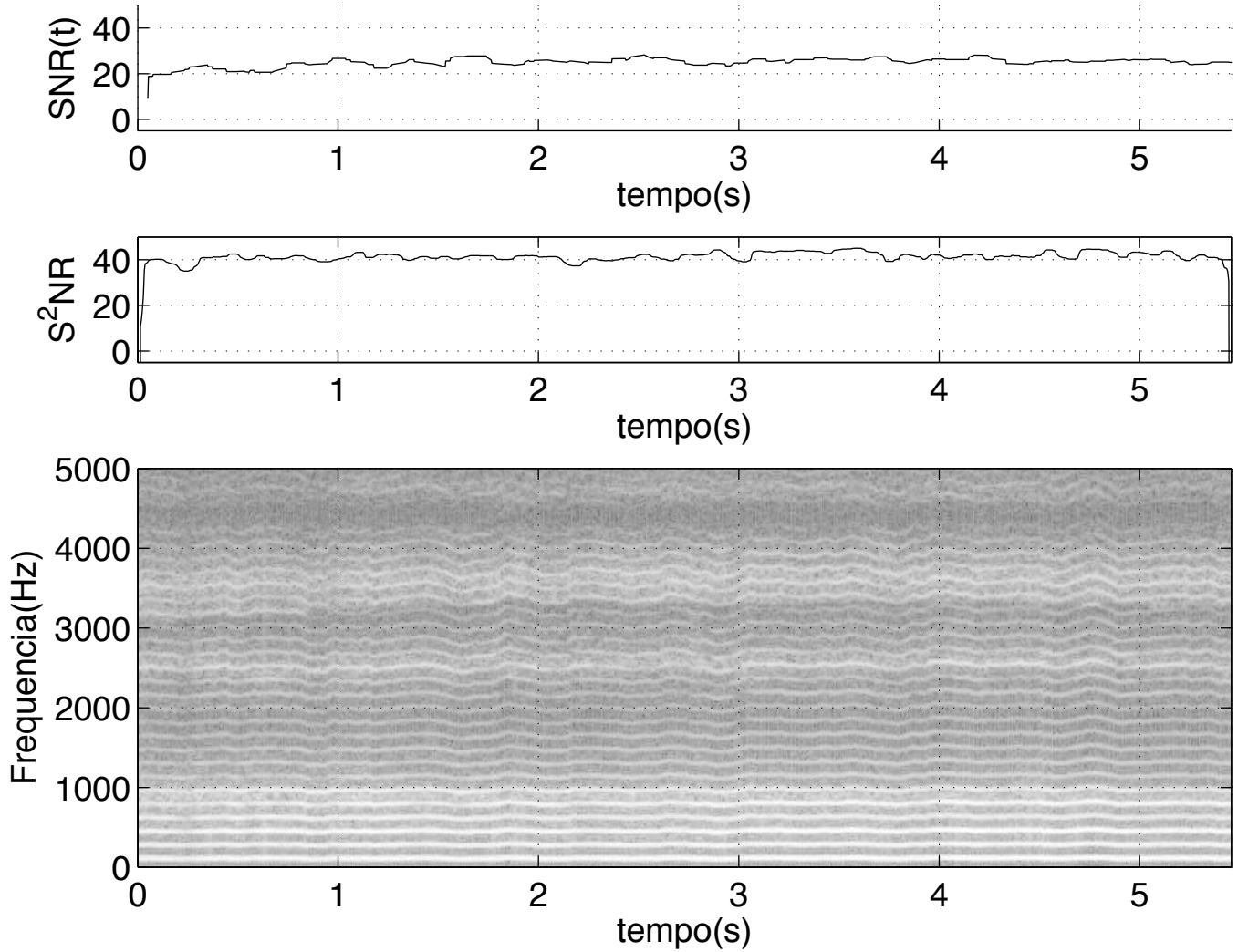

Figura 4.36: **Estimativas de SNR com voz real.** Comparativo das medidas de S^2NR e $SNR(t)$, com voz real, vogal /a/. Voz classificada perceptivamente como soprosidade grau 0,5.

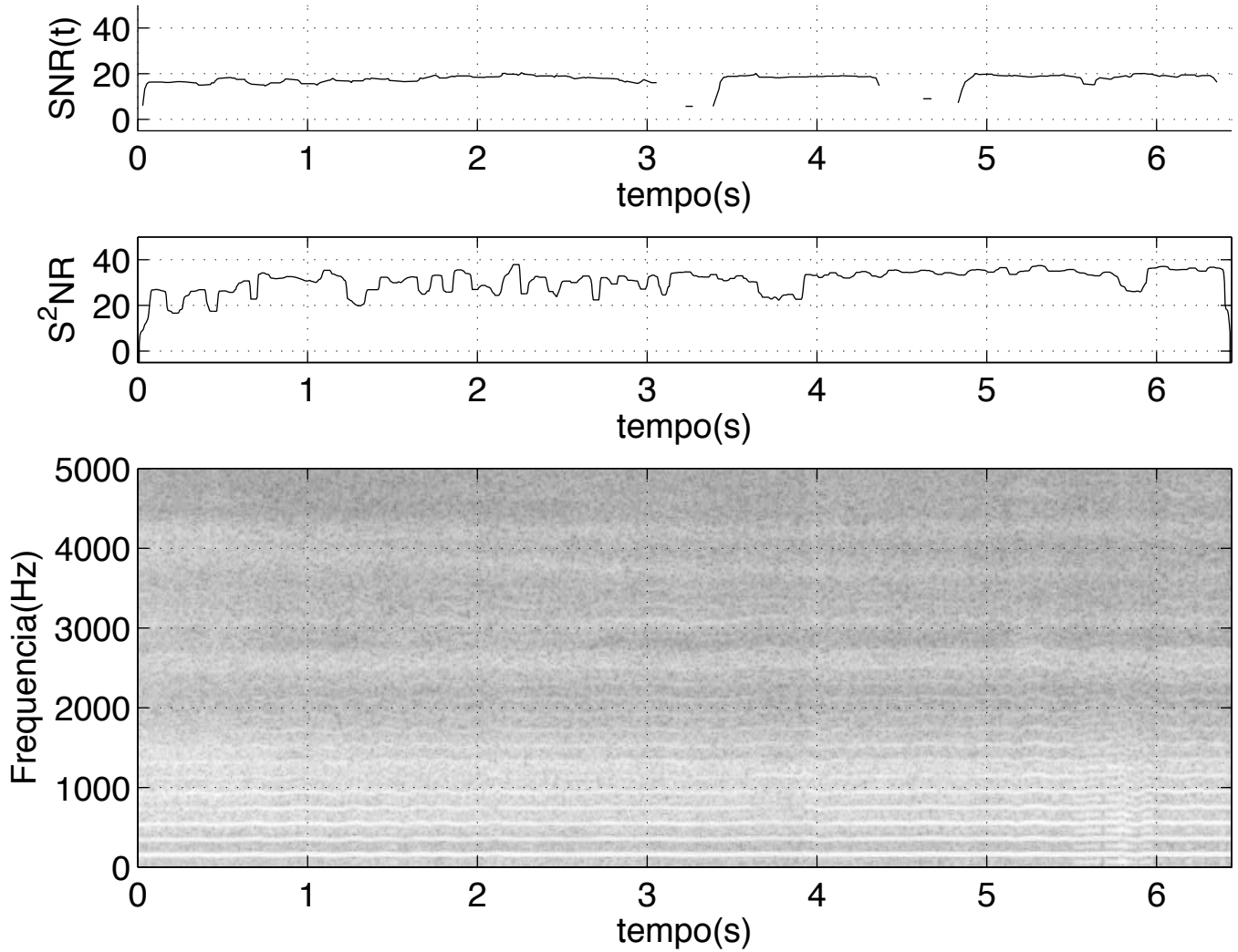

Figura 4.37: **Estimativas de SNR com voz real.** Comparativo das medidas de S^2NR e $SNR(t)$, com voz real, vogal /a/. Voz classificada perceptivamente como soprosidade grau 1,5.

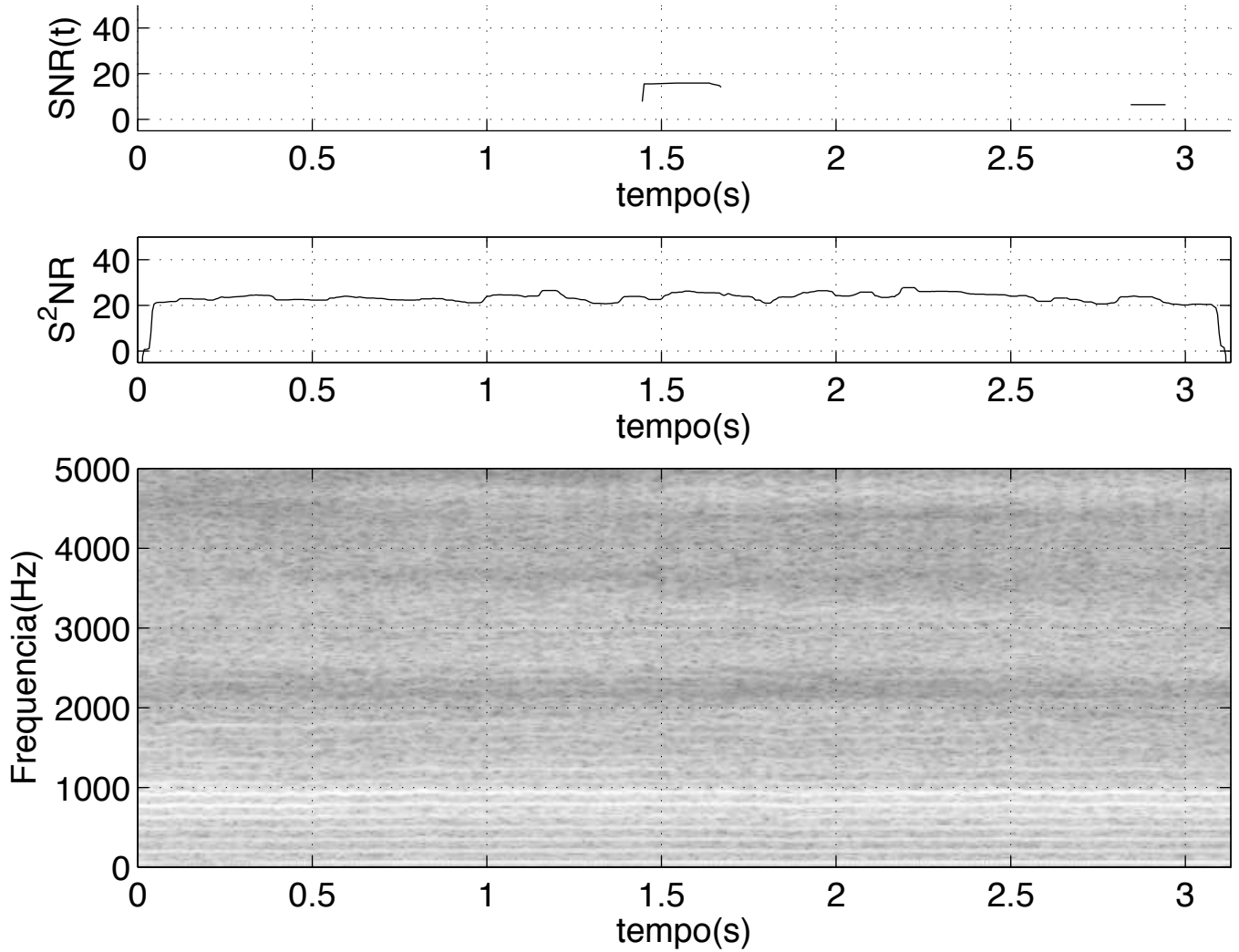

Figura 4.38: **Estimativas de SNR com voz real.** Comparativo das medidas de S^2NR e $SNR(t)$, com voz real, vogal /a/. Voz classificada subjetivamente como soprosidade grau 2,0.

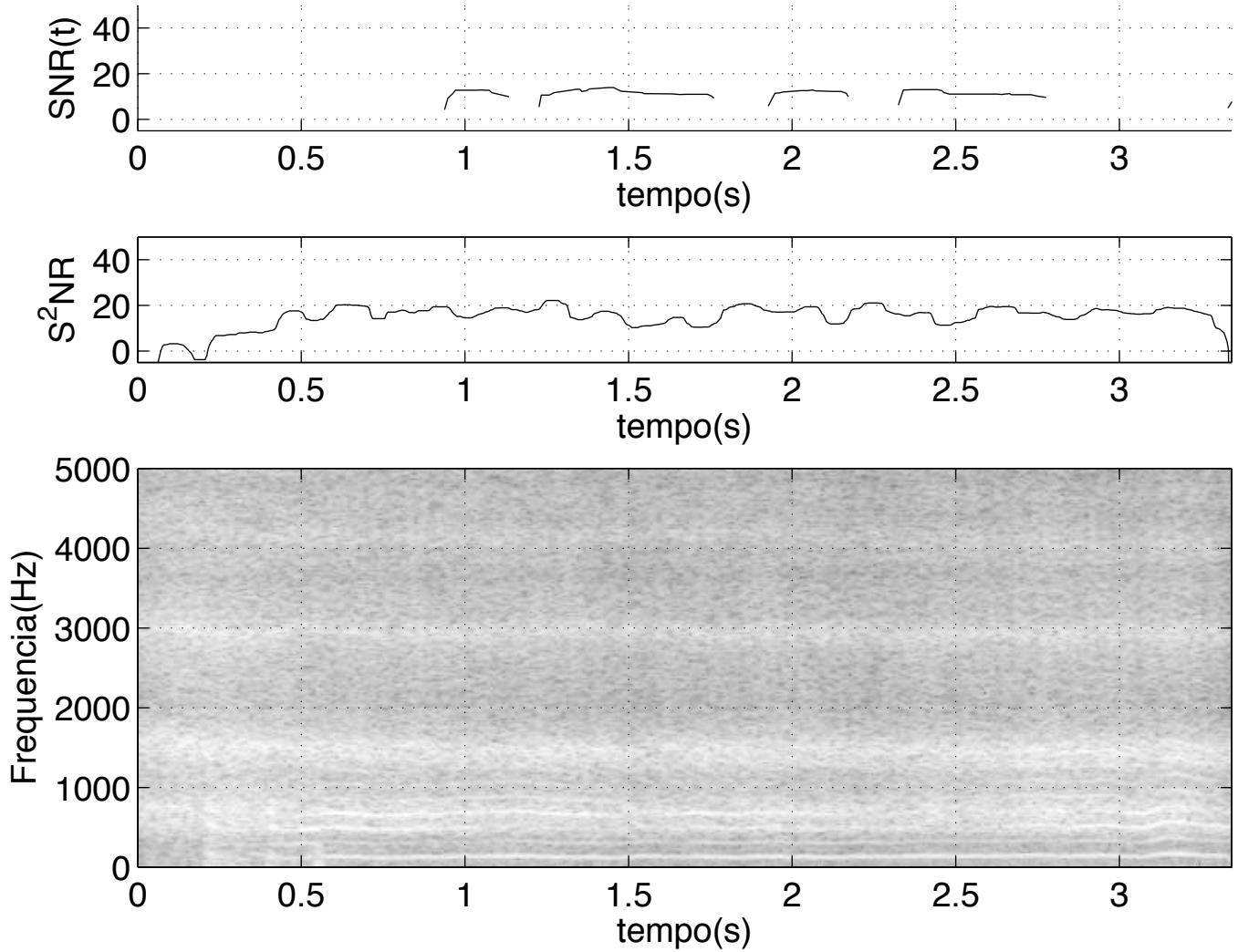

Figura 4.39: **Estimativas de SNR com voz real.** Comparativo das medidas de S^2NR e $SNR(t)$, com voz real, vogal /a/. Voz classificada subjetivamente como soprosidade grau 2,5.

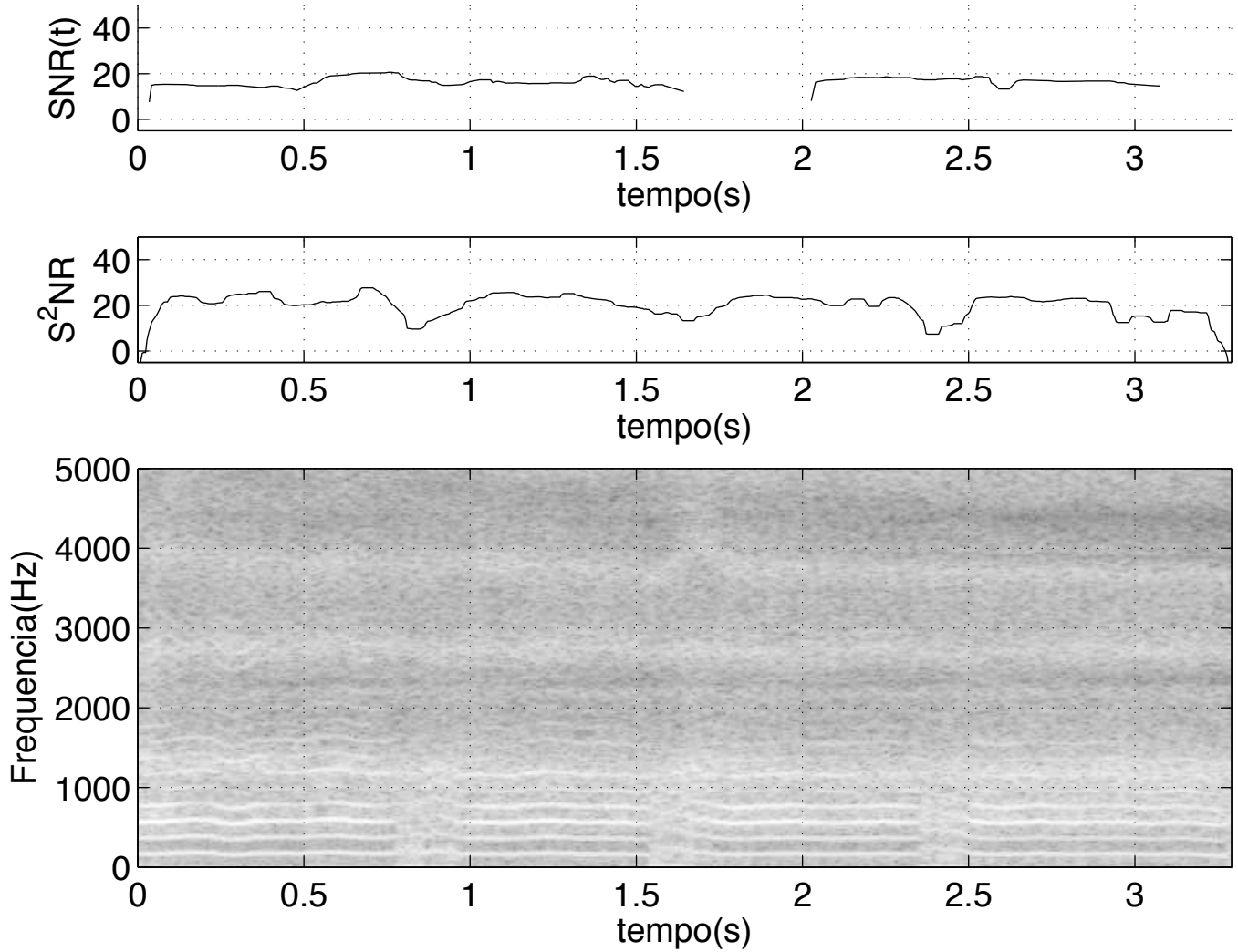

Figura 4.40: **Estimativas de SNR com voz real.** Comparativo das medidas de S^2NR e $SNR(t)$, com voz real, vogal /a/. Voz classificada subjetivamente como soprosidade grau 2,5.

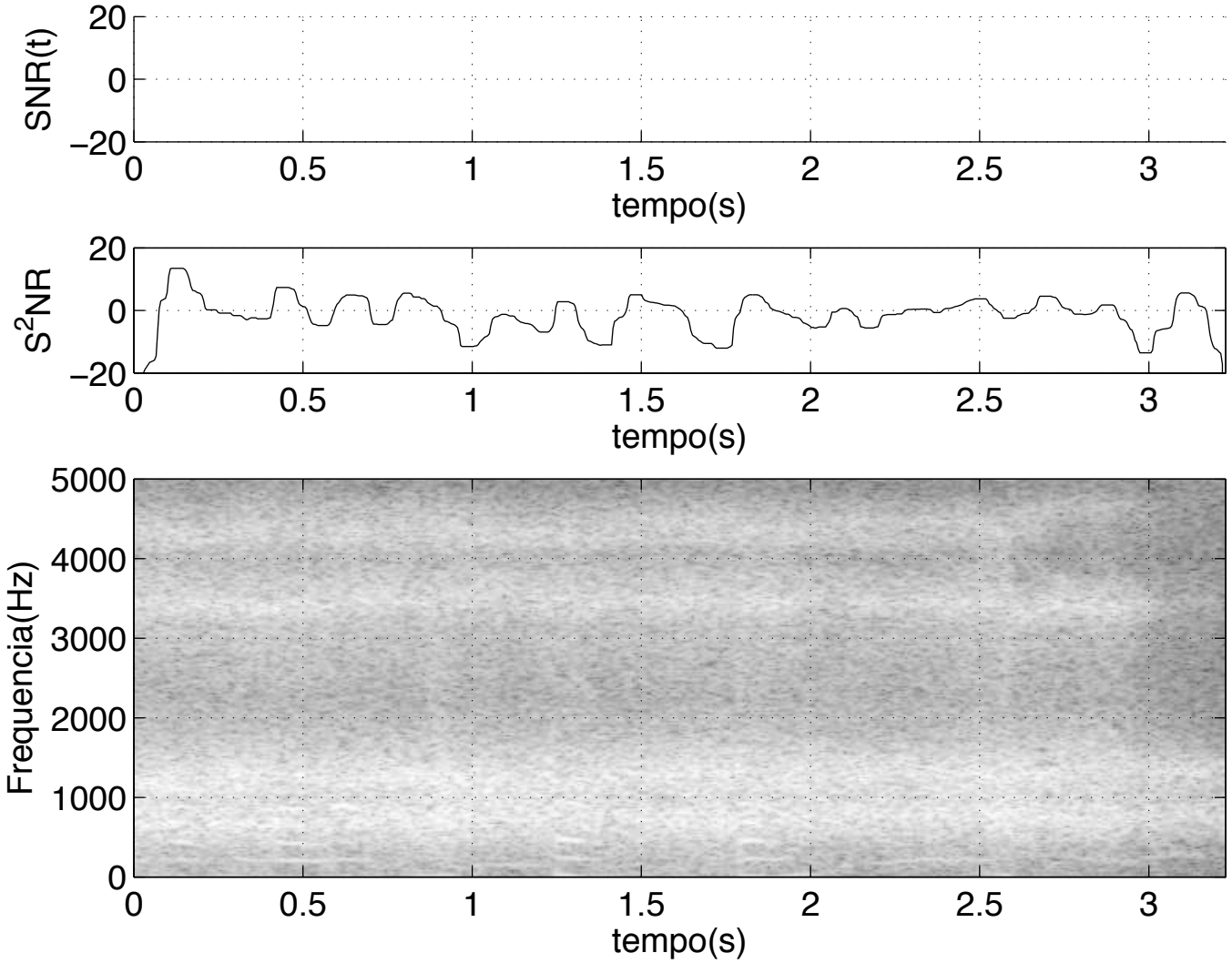

Figura 4.41: **Estimativas de SNR com voz real.** Comparativo das medidas de S^2NR e $SNR(t)$, com voz real, vogal /a/. Voz classificada subjetivamente como soprosidade grau 3,0.

4.3.4 Aplicação em fala encadeada

Aplicando o algoritmo em fala encadeada, mesmo sem um detector de vozeamento, é possível estimar a relação sinal-ruído ao longo do tempo. Nestes casos, os períodos sem fonação apresentam relação sinal ruído baixa e a resposta é rápida quando ocorrem períodos com vozeamento. Saliente-se que a calibração só foi aferida com sucesso para a vogal /*a*/.

O texto lido é *"Shall we come in? Yes, open the door, where have you been? We were away a year ago. And Arthur ? Arthur went out every afternoon with Amy, Oliver and Ian"* [Vieira (1997)], com duração aproximada de 10 a 15 segundos, dependendo do locutor.

O algoritmo $SNR(t)$ mede enquanto é possível detecção do vozeamento. Já o algoritmo da S^2NR mede continuamente e retorna valores mesmo nos momentos sem vozeamento. No entanto, nota-se uma grande redução da relação sinal-ruído nestes instantes, estabelencendo um limiar de confiança, pode-se propor um algoritmo de detecção de vozeamento. No caso da Figura 4.42, o valor do limiar é de aproximadamente $10\,dB$, fato que pode ser confirmado pelo espectrograma sincronizado e pela medição de $SNR(t)$.

Tomando a figura citada anteriormente, em seus três segundos iniciais, mostra-se a transcrição do trecho na Figura 4.43. Nota-se que nas regiões onde há maior relação sinal ruído, há melhor definição das linhas espectrais harmônicas. No primeiro segmento, tem-se a pronunciação de *Sh-*, e nota-se um início ruidoso, característico da fricativa /ʃ/, seguido pela vogal æ, onde a S^2NR aumenta significativamente. O segmento corresponde a *we*, onde a S^2NR cresce até $32,0\,dB$, por se tratar de um trecho vocálico. Os trechos *co-* e *-me*, possui instantes com vozeamento, que podem ser constatados quando se supera o limiar definido, quando atinge-se $15,3\,dB$. No trecho *in*, SNR atinge $\approx 28,2\,dB$ na parte da vogal.

A etapa seguinte é um período longo de silêncio, devido à baixa potência do sinal, o algoritmo detecta ruído como sinal, dando indicação incorreta na S^2NR, o que pode ser facilmente contornado com um detector de nível. O segmento seguinte é um artefato de gravação, devido à respiração do locutor, e o algoritmo detecta corretamente como ruído. Após esta etapa, tem-se novamente silêncio, seguido de nova etapa vozeada, com uma semi-vogal e outra vogal, *ye-* incluindo a transição entre elas. Temos então o trecho *-s*, correspondente à fricativa ʃ, onde ocorre um severa queda na SNR. Entrando a vogal *o-*, de *open*, a S^2NR aumenta para $35\,dB$. A etapa *-p-* corresponde à consoante plosiva p, que apresenta S^2NR (em torno de $7,4\,dB$) maior que a da fricativa ʃ, mas inferior a uma vogal. A etapa *-en*, a relação aumenta novamente para $28\,dB$.

Tomando um locutor com grau mais pronunciado de disfonia, como mostra a Figura 4.44, o algoritmo temporal tem grandes dificuldades em encontrar os períodos de vozeamentos. A relação sinal ruído nesta amostra é inferior à da amostra citada anteriormente, e o limiar de vozeamento deve ser reduzido, para aproximadamente $5\,dB$.

Marcaram-se os mesmos pontos do exemplo anterior, que são mostrados na Figura 4.45. Comparativamente, pode-se dizer que o algoritmo do $SNR(t)$ mediu nos trechos: *-all,we,-me,co-,o-,-en*, onde ocorrem as vogais. No entanto, esta medida foi intermitente, mesmo

nos trechos onde no espectro visivelmente ocorre vozeamento. Em termos quantitativos, as medidas ficaram em valores próximos nos trechos *-all*, *we* e *-en*. Em regiões como em *-en*, pico do S^2NR, a medição da $SNR(t)$ foi mais baixa (em torno de $20,0\,dB$, contra $35,0\,dB$ da S^2NR). Tal disparidade deve-se à variação de freqüência na região, característica que prejudica a medição no tempo.

Novamente, foi possível a medição contínua de S^2NR com medições em todos os sons vocálicos da amostra. No entanto, pelas razões citadas no item anterior (artefatos de gravação, respiração), nos intervalos de silêncio, surgem medidas espúrias como no trecho próximo a $t = 1,5\,s$. A $SNR(t)$ não mediu nos trechos *-me*, *in*, *ye-* e *o-*. O valor dos picos da S^2NR variou de $15\,dB$ até $28\,dB$ no trecho *-en*, ao passo que a medição no tempo não ultrapassou $20\,dB$.

Pelos exemplos anteriores, devido à grande sensibilidade do algoritmo S^2NR, regiões não vozeadas ou em silêncio (apenas com ruído) de fundo podem ser erroneamente medidas. No algoritmo temporal o contrário ocorre, apenas regiões com vogais bem pronunciadas ou com sinal ruído mais alto são detectadas.

O algoritmo S^2NR apresentou grande sensibilidade e resposta rápida quando existe vozeamento. Pode-se aplicá-lo, portanto, à fala articulada. Associado a um detector de silêncio, definido um limiar para cada amostra *a posteriori* automaticamente, pode-se eliminar as regiões espúrias, aumentando a confiabilidade da medição.

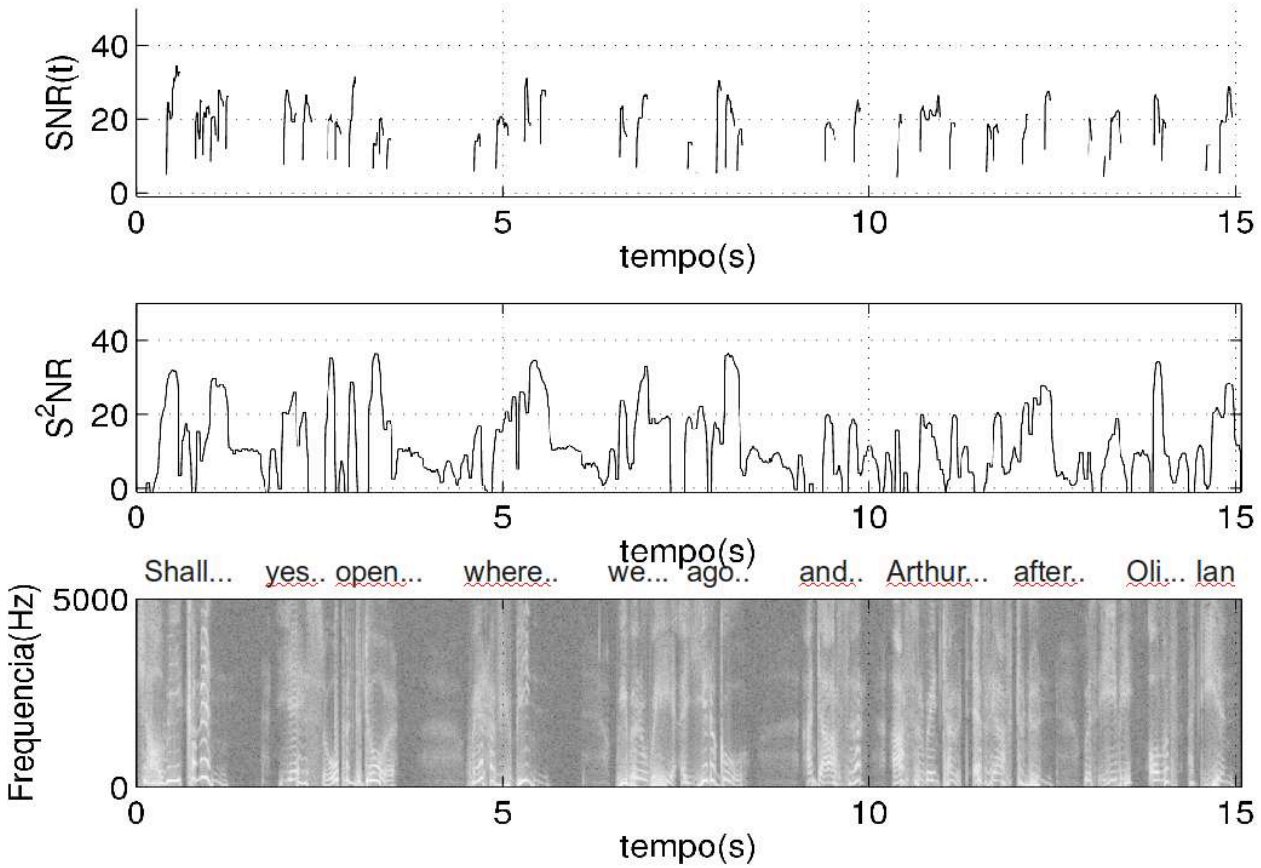

Figura 4.42: **Estimativas de SNR em fala encadeada.** Comparativo das medidas de S^2NR e $SNR(t)$, fala encadeada. A passagem lida é *"Shall we come in? Yes, open the door, where have you been? We were away a year ago. And Arthur ? Arthur went out every afternoon with Amy, Oliver and Ian"*.

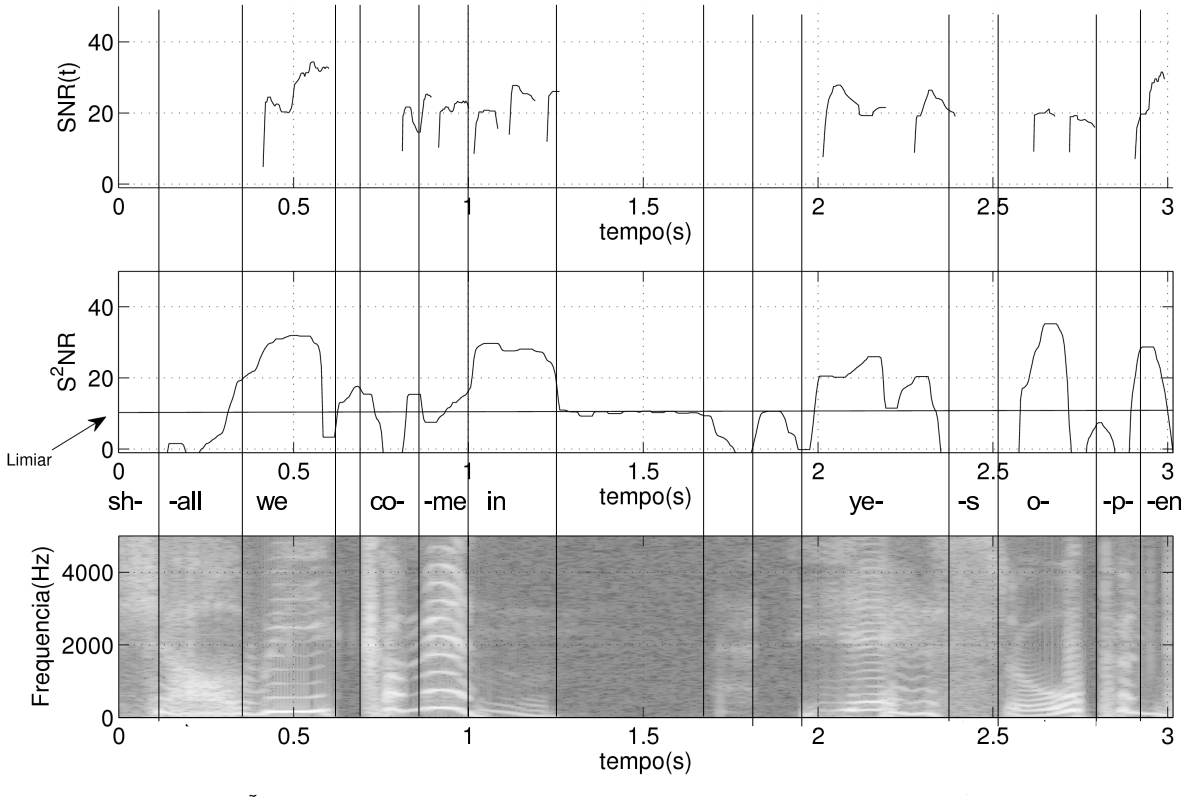

Figura 4.43: **Estimativas de SNR em fala encadeada.** Comparativo das medidas de S^2NR e $SNR(t)$, fala encadeada, 3 segundos iniciais da amostra mostrada na Figura 4.42.

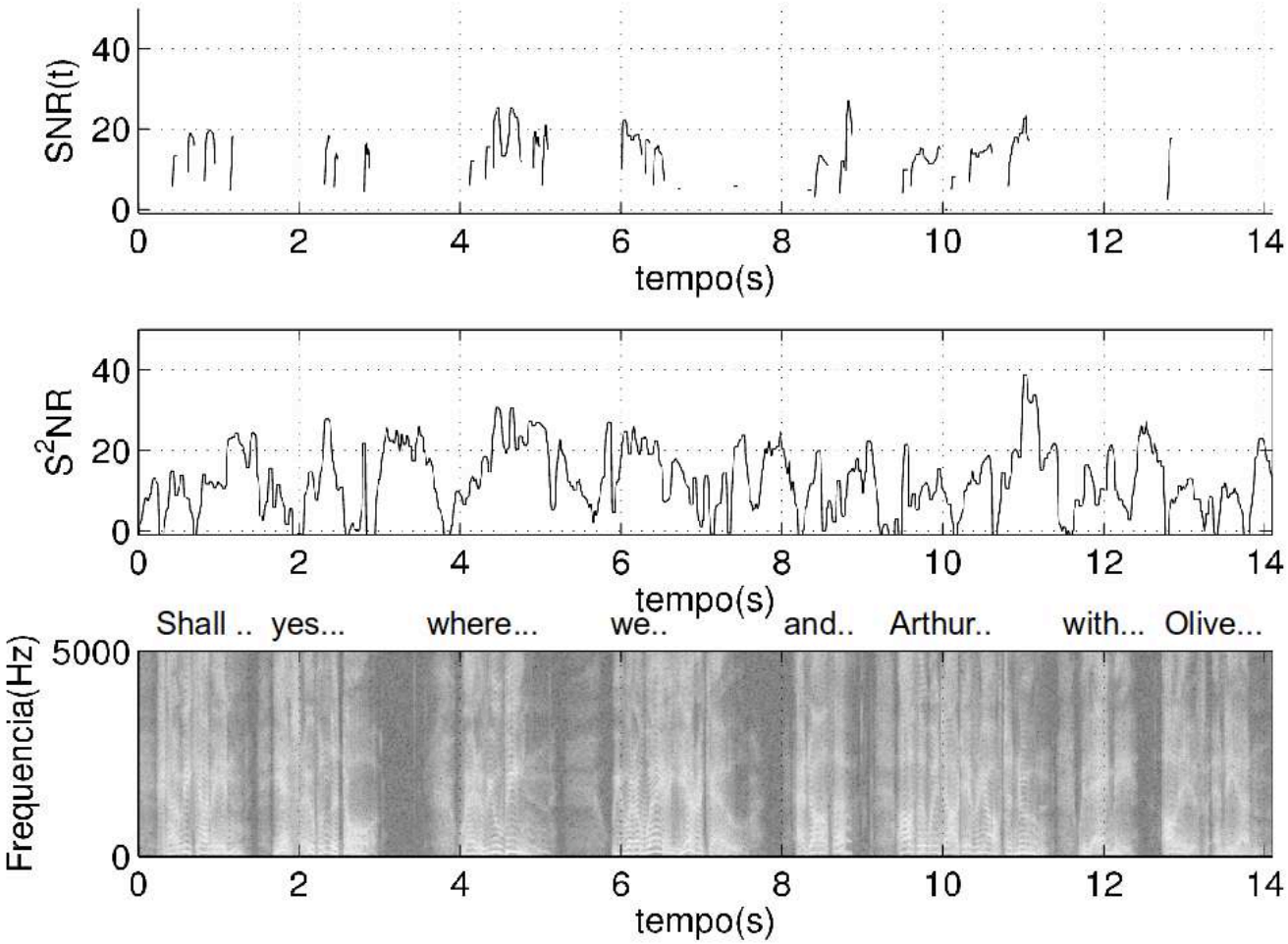

Figura 4.44: **Estimativas de SNR em fala encadeada.** Comparativo das medidas de S^2NR e $SNR(t)$, fala encadeada. A passagem lida é *"Shall we come in? Yes, open the door, where have you been? We were away a year ago. And Arthur ? Arthur went out every afternoon with Amy, Oliver and Ian"*.

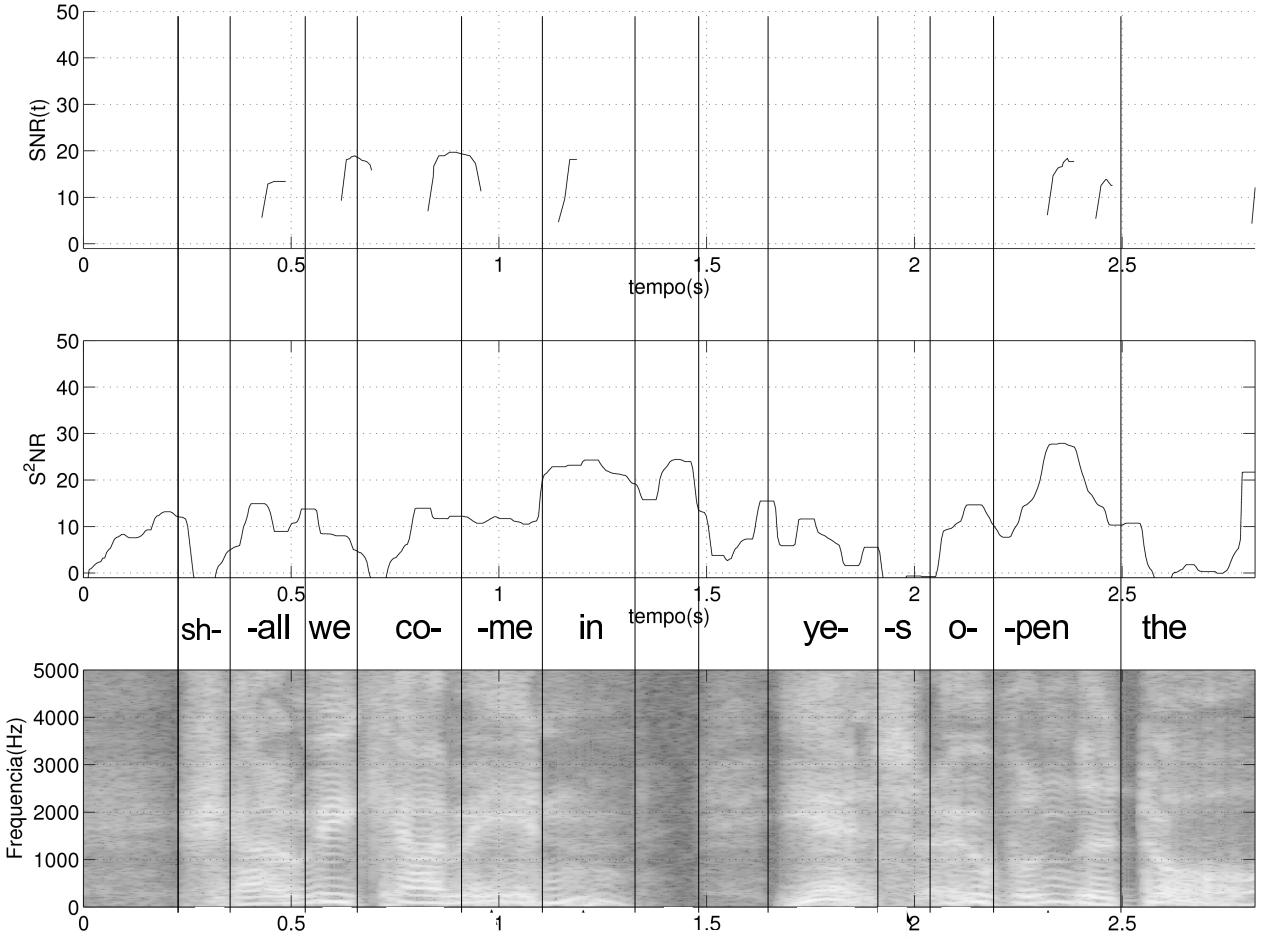

Figura 4.45: **Estimativas de SNR em fala encadeada.** Comparativo das medidas de S^2NR e $SNR(t)$, fala encadeada, 3 segundos iniciais da amostra mostrada na Figura 4.44.

Capítulo 5

Conclusão

5.1 Sumário

Neste trabalho, estabeleceu-se inicialmente uma plataforma para geração de voz sintética para se testar com parâmetros controlados de freqüência fundamental, *jitter*, *shimmer*, SNR teórico a medição dos algoritmos de estimação relação sinal ruído.

Em seguida, discutiram-se os métodos tradicionais de medição de sinal ruído, suas limitações e proposição de um novo método, baseado na análise da imagem do espectrograma de vogais utilizando técnicas de identificação de impressão digital.

De posse da plataforma, gerou-se uma massa de dados para o teste dos algoritmos, de onde pode-se avaliar o desempenho em condições controladas de perturbação vocal, estrutura harmônica da vogal e freqüência fundamental.

5.2 Considerações finais

Uma medida ideal de relação sinal-ruído deve ser robusta a variações na estrutura harmônica da vogal, em perturbações de freqüência e amplitude, assim como ao valor médio da freqüência fundamental. A S^2NR apresentou características interessantes, como grande insensibilidade às perturbações vocais (*jitter/shimmer*). Nas vogais sintéticas (calibração), apresentou-se desempenho inferior na vogal /i/. Sob *jitter*, com $f_o = 120\,Hz$, os valores de desvio máximo em relação a referência foram de $2,1\,dB$, $11,5\,dB$ e $2,9\,dB$ para as vogais /a/, /i/ e /u/, respectivamente. Já sob *shimmer*, estes valores de desvio foram de $2,5\,dB$, $4,4\,dB$ e $3,6\,dB$.

Destes objetivos, pode-se perceber para vogal masculina:

- Maior robustez nas vogais /a/, /u/ em casos de *jitter*
- Maior robustez nas vogais /a/, /i/, /u/ em casos de *shimmer*

Para a vogal feminina, notou-se que com a janela padrão de $N = 1024$, o comportamento do algoritmo não foi compatível com o masculino. Alterando para $N = 512$, obteve-se resultados semelhantes ao caso masculino.

Ao aplicar perturbações simultâneas, o algoritmo continuou a responder conforme os casos individuais.

Aplicando o algoritmo em voz real, com variados graus de soprosidade, a medida da relação sinal ruído mostrou-se coerente com a classificação subjetiva.

5.3 Trabalhos futuros

Neste trabalho, o ajuste de grande parte dos parâmetros da S^2NR foi realizado empiricamente. Neste caso, um estudo sobre o efeito da variação paramétrica no cálculo da S^2NR seria útil para melhorar o ajuste. Neste aspecto, provalvelmente é impossível um ajuste global para todas as condições de voz. Um alternativa possível seria o ajuste adaptivo de parâmetros em condições limite de detecção, tais como relação sinal ruído muito baixa, ou o oposto, relação sinal ruído muito alta. Um ajuste automático do tamanho da janela de FFT relativo a freqüência fundamental também poderia melhorar o algoritmo.

No caso de vogais sustentadas, comum em fonoaudiologia, pode ser feito um ajuste pelo usuário (por sexo ou estimativa de f_0).

Visando uma implementação mais ampla do algoritmo, a otimização de determinadas etapas seria importante, como no cálculo do filtros de Gabor, etapa que ocupa grande parte do processamento. Uma alternativa possível é restringir as faixas de busca do ângulo na etapa de filtragem.

Uma investigação das causas das limitações em estruturas harmônicas como na vogal /i/ faz-se necessária, já que a avaliação apresentou desempenho inferior nesta vogal, embora pareça não ter ocorrido na análise de fala corrente.

Referências Bibliográficas

Awan, S. N. e Frenkel, M. L. (1994). Improvements in estimating the harmonics-to-noise ratio of the voice. *Journal of Voice*, 8(3):255–262.

Baken, R. e Orlikoff, R. F. (2000). *Clinical Measurement of Speech and Voice*. Singular.

Bazen, A. M. (2002). Systematic methods for the computation of the directional fields and singular points of fingerprints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24.

Beech, J.; Harding, L. e Hilton-Jones, D. (1993). *Assessment in Speech and Language Therapy*. Rooutledge.

Burnett, T. A.; Senner, J. E. e Larson, C. R. (1997). Voice f0 responses to pitch-shifted auditory feedback: a preliminary study. *Journal of Voice*, 11(2):202 – 211.

Cox, N. B.; Ito, M. R. e Morrison, M. D. (1989). Data labeling and sampling effects in harmonics-to-noise ratios. *The Journal of the Acoustical Society of America*, 85(5):2165–2178.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160.

de Krom, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech and Hearing Research*, 36:254–266.

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton, Hague.

Fant, G. (1979). Glottal source and excitation analysis. *Speech Transmission Laboratory - Quaterly Progress and Status Report*, 1:85–107.

Fant, G.; Liljencrant, J. e guang Lin, Q. (1985). A four-parameter model of glottal flow. *Speech Transmission Laboratory - Quaterly Progress and Status Report*, 4:1–13.

Flanagan, J. L. (1972). *Analysis, Synthesis, and Perception of Speech*. Springer Verlag, Berlin.

Hemler, R.; Wieneke, G. H. e Dejonckere, H. (1997). The effect of relative humidity of inhaled air on acoustic parameters of voice in normal subjects. *Journal of Voice*, 11:295–300.

Hirano, M. (1981). *Clinical Examination of Voice*. Springer Verlag.

Hong, L.; Wan, Y. e Jain, A. (1998). Fingerprint image enhancement: algorithm and performance evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):777–789.

Ishizaka, K. e Flanagan, J. (1972). Synthesis of voiced sounds from a two-mass model of the vocal cords. *The Bell System Technical Journal*, 51(6):1233–68.

Jain, A. e Farrokhnia, F. (1990). Unsupervised texture segmentation using gabor filters. *IEEE International Conference on Systems, Man and Cybernetics, 1990. Conference Proceedings.*, pp. 14–19.

Kasuya, H.; Ogawa, S. e Kikuchi, Y. (1986a). An adaptive comb filtering method as applied to acoustic analyses of pathological voice. *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '86.*, 11:669–672.

Kasuya, H.; Ogawa, S.; Mashima, K. e Ebihara, S. (1986b). Normalized noise energy as an acoustic measure to evaluate pathologic voice. *The Journal of the Acoustical Society of America*, 80(5):1329–1334.

Kovesi, P. D. (2005). MATLAB and Octave functions for computer vision and image processing. School of Computer Science & Software Engineering, The University of Western Australia. Disponível em: <http://www.csse.uwa.edu.au/∼pk/research/matlabfns/>, acessado em Dezembro de 2006.

Kreiman, J. e Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *The Journal of the Acoustical Society of America*, 104(3):1598–1608.

Kreiman, J.; Gerratt, B. R.; Kempster, G. B.; Erman, A. e Berke, G. S. (1993). Perceptual Evaluation of Voice Quality: Review, Tutorial, and a Framework for Future Research. *Journal of Speech, Language, and Hearing Research*, 36(1):21–40.

Laver, J.; Wirz, S.; Mackenzie, J. e Hiller, S. (1981). A perceptual protocol for the analysis of vocal profiles. *Work in Progress*, 14:139–155.

Markel, J. e Gray Jr, A. H. (1976). *Linear Prediction of Speech*. Springer-Verlag.

Marple, S. (1987). *Digital Spectral Analysis with Applications*. Prentice-Hall, Englewood Cliffs.

Murray, J. D. e vanRyper, W. (1996). *Encyclopedia of graphics file formats (2nd ed.)*. O'Reilly & Associates, Inc., Sebastopol, CA, USA.

Nixon, M. K. e Aguado, A. S. (2002). *Feature Extraction and Image Processing*. Newnes.

Orlikoff, R. F. e Kahane, J. C. (1991). Influence of mean sound pressure level on jitter and shimmer measures. *Journal of voice*, 5:113–119.

Rothenberg, M. (1973). A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *Journal of the Acoustical Society of America*, 53(6):1632–1645.

Schäfer-Vincent, K. (1983). Pitch period detection and chaining: method and evaluation. *Phonetica*, 40:177–202.

Snider, J. e Osgood, C. (1969). *Semantic Differential Technique: A Sourcebook*. Aldine Pub. Co.

Stevens, K. N. (1998). *Acoustic Phonetics*. The MIT Press, Cambridge, Massachusetts.

Story, B. H.; Titze, I. R. e Hoffman, E. A. (1996). Vocal tract area functions from magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 100(1):537–554.

Titze, I. R. (1973). The human vocal cords: a mathematical model, i. *Phonetica*, 28:129–170.

Titze, I. R. (1995). *Workshop on Acoustic Voice Analysis*. National Center for Voice and Speech.

Titze, I. R. e Strong, W. J. (1975). Normal modes in vocal cord tissues. *The Journal of the Acoustical Society of America*, 57(3):736–744.

Vieira, M. N. (1997). *Automated Measures of Dysphonias and the Phonatory Effects of Asymmetries in the Posterior Larynx*. PhD thesis, University of Edinburgh.

Wong, D.; Ito, M. R.; Cox, N. B. e Titze, I. R. (1991). Observation of perturbations in a lumped-element model of the vocal folds with application to some pathological cases. *The Journal of the Acoustical Society of America*, 89(1):383–394.

Wong, D.; Lange, R.; Titze, I. R. e Guo, C. G. (1995). A qualitative study of mechanisms of jitter-induced shimmer in the voice. *The Journal of the Acoustical Society of America*, 97(5):3421–3421.

Wong, D. Y.; Markel, J. D. e Jr, A. H. G. (1979). Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27:350–355.

Yanagihara, N. (1967). Significance of harmonic changes and noise components in hoarseness. *Journal of Speech, Language, and Hearing Research*, 10(3):531–541.

Yumoto, E. (1983). The quantitative evaluation of hoarseness. *Archives of Otolaryngology*, 109(1):48–52.

Yumoto, E.; Gould, W. J. e Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of Acoustical Society of America*, 71(6):1544–1549.

Yumoto, E.; Sasaki, Y. e Okamura, H. (1984). Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness. *Journal of Speech, Language, and Hearing Research*, 27(1):2–6.