TESE DE DOUTORADO Nº 283

ANÁLISE DE MEDIDAS OBJETIVAS E DA PERCEPÇÃO AUDITIVA DA SOPROSIDADE VOCAL

João Pedro Hallack Sansão

DATA DA DEFESA: 21/09/2018

João Pedro Hallack Sansão

Análise de Medidas Objetivas e da Percepção Auditiva da Soprosidade Vocal

Belo Horizonte, Minas Gerais, Brasil

2018

Universidade Federal de Minas Gerais

Escola de Engenharia

Programa de Pós-Graduação em Engenharia Elétrica

ANÁLISE DE MEDIDAS OBJETIVAS E DA PERCEPÇÃO AUDITIVA DA SOPROSIDADE VOCAL

João Pedro Hallack Sansão

Tese de Doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do Título de Doutor em Engenharia Elétrica.

Orientador: Prof. Maurílio Nunes Vieira

Belo Horizonte - MG

Setembro de 2018

"Análise de Medidas Objetivas e da Percepção Auditiva da Soprosidade Vocal"

João Pedro Hallack Sansão

Tese de Doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do grau de Doutor em Engenharia Elétrica.

Aprovada em 21 de setembro de 2018.

Por:

Maurílio Nunes Vieira
Prof. Dr. Maurílio Nunes Vieira
DELT (UFMG) - Orientador

Prof. Dr. Hani Camille Yehia
DELT (UFMG)

Prof. Dr. Adriano Vilela Barbosa
DELT (UFMG)

Profa. Dra. Ana Cristina Côrtes Gama
Fonoaudiologia (UFMG)

Prof. Dr. Rafael Michelin Laboissière
LPNC (Université Grenoble Alpes)

Prof. Dr. Miguel Arjona Ramírez
PSI (USP)

Agradecimentos

Aos meus pais, Raphael e Suraia e à minha irmã Júlia, que sempre me apoiaram durante o tortuoso caminho que foi o desenvolvimento deste trabalho.

Ao professor Maurílio Nunes Vieira, que me acompanhou durante toda minha formação acadêmica e me introduziu ao tema desenvolvido neste trabalho.

Ao professor Hani Camille Yehia pelo incansável apoio e por novamente ter operado milagres que permitiram a conclusão desta tese.

Aos colegas do CEFALA e CELTA que me apoiaram no desenvolvimento deste trabalho, em especial os professores Armando, Leonardo Carneiro, Leonardo Mozelli e Marcos Vinícius.

Ao professor Adriano Vilela Barbosa, à professora Ana Cristina Côrtes Gama, ao professor Miguel Arjona Ramírez, ao professor Rafael Laboissière por atuarem na banca deste trabalho e contribuírem com a melhoria deste texto.

Reforço os agradecimentos à Professora Ana Cristina e ao Professor Rafael, pelas diversas sugestões que auxiliaram no desenvolvimento deste trabalho.

Aos pacientes da *Royal Infirmary of Edinburgh* que cederam seus registros de voz para o desenvolvimento de diversos trabalhos ao longo de todos estes anos. Em especial, agradeço às fonoaudiólogas Ana Paula da Penha e Mariana Borges, que executaram o trabalho inicial de avalição perceptiva destas vozes.

Aos voluntários que tiveram a disponibilidade e paciência para participar dos testes perceptivos executados neste trabalho.

Aos professores Kenneth Knoblauch, Lawrence Maloney, Frederick Kingdom, Nicolass Prins, Peter Kovesi e outros que contribuíram com softwares de código aberto utilizados neste trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal em Nível Superior - Brasil (CAPES) - Código de Financiamento 001 e da Universidade Federal de São João del-Rei.

Resumo

A análise da qualidade de voz é um campo vasto, que abrange áreas como ciência e tecnologia da fala, telecomunicações, fonética e fonoaudiologia. Ela visa complementar os exames de visualização das pregas vocais (como video-laringoscopia) na investigação da voz disfônica. Em geral, este tipo de distúrbio decorre de alterações que afetam 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. A análise da voz disfônica é feita em duas modalidades: através do uso de medidas instrumentais do sinal acústico da voz como, por exemplo, a freqüência fundamental ciclo-a-ciclo, e usando a percepção auditiva, por meio de escalas perceptivas. Estas escalas surgem da rotulação (i.e. atribuição de adjetivos) de impressões observadas pelos ouvintes de uma determinada voz. A escala GRBAS (e variações) é a mais comumente utilizada no Brasil. Nela, os atributos estudados são a rugosidade (R), soprosidade (B), astenia (A), tensão (S) e a impressão geral ou grau (G). Atributos como os citados acima são subjetivos e, apesar de serem empregados clinicamente na avaliação de problemas na voz, ainda há dificuldade no estabelecimento de escalas perceptivas consistentes entre ouvintes.

Neste trabalho, escolheu-se o atributo perceptivo de soprosidade para estudo, uma característica percebida devido ao fluxo turbulento gerado na glote ou devido ao escape de ar excessivo através de uma fenda. Diferentes graus de severidade são encontrados nos pacientes, varrendo de níveis leves de perturbação até casos extremos. O estudo da soprosidade vocal foi executado em duas frentes: das medidas instrumentais e da percepção auditiva.

Em relação às medidas instrumentais, desenvolveu-se uma plataforma para teste de variados métodos. Esta comparação de métodos consistiu em (i) avaliar amostras sintéticas de voz com relação sinal-ruído, jitter e shimmer controlados; e (ii) avaliar amostras de voz real predominantemente soprosas, avaliadas em trabalhos anteriores em uma escala perceptiva de 7 pontos. Foram testados métodos conhecidos da literatura, como a CPPS (*cepstral peak prominence smoothed*), SFR (*spectral flatness residue signal*), HNR (*harmonic-to-noise ratio*) e S^2NR (*spectrographic signal-to-noise ratio*), tendo sido este último método desenvolvido neste trabalho. Através deste estudo, determinou-se que os métodos que apresentam maior correlação entre medidas instrumentais e perceptiva são a CPPS e a S^2NR. Notou-se também que a S^2NR apresenta maior robustez a perturbações em freqüência e amplitude que a CPPS, sendo, neste critério, uma medida melhor de soprosidade.

Em relação à percepção, o passo inicial foi a caracterização da psicofísica da soprosidade, relacionando as variações na dimensão física (nível de ruído glótico) e a escala percep-

tiva. Outro resultado foi a obtenção das mínimas diferenças perceptíveis na variação de intensidade de ruído.

O objetivo seguinte deste trabalho foi estabelecer um método de classificação comparativo para a qualidade da voz soprosa, o qual exija mínimo treinamento e conhecimento prévio do julgador e que tenha a maior concordância intra-julgadores e inter-julgadores. Este método foi baseado na busca de elementos em uma árvore binária. A seqüência das comparações segue a estrutura da árvore, seguindo até não ser possível a distinção de diferenças ou atingir o último nível. Para tanto, são escolhidas amostras de referência (âncoras), que são comparadas às amostras em análise. As escolhas das âncoras e da profundidade da árvore de busca são feitas com base na psicofísica da soprosidade. Explora-se, deste modo, a natureza relativa do ouvido humano, em contraste com os métodos atuais de avaliação perceptiva, nos quais o ouvido é tratado como absoluto. No método desenvolvido, são escolhidas 3 ou 7 amostras-âncora, com relação sinal-ruído crescente, necessitando de 2 ou 3 comparações distintas para a avaliação de uma amostra de teste. Nos experimentos, utilizando vogais sintéticas e voz humana soprosa, a avaliação comparativa apresentou alta confiabilidade inter-julgadores, mesmo com avaliadores inexperientes.

Palavras-chaves: voz disfônica, correlatos acústicos, soprosidade vocal, percepção, psicofísica, relação sinal-ruído

Abstract

Voice quality is a vast field, which covers the areas of speech science and technology, telecommunications, phonetics and speech therapy. It aims to complement vocal fold visualization exams (such as video-laryngoscopy) in the investigation of dysphonic voices. In general, this type of anomaly results from changes affecting vocal folds anatomy due to lesions (nodules, polyps, etc.) and phonatory behavior due to abnormal tension in the laryngeal muscles. Clinical voice analysis is usually done in two ways: through the use of instrumental measures of the acoustic signal, for example, fundamental frequency cycle-to-cycle perturbations, or by auditory perceptive scales. In the GRBAS scale, which is the most commonly used, the perceptual attributes are general (holistic) impression (G), roughness (R), breathiness (B), asthenia (A) and strain (S). Attributes such as those are subjective and, although widely used clinically in the evaluation of problems in the voice, there are still difficulties in using them consistently.

In this work, we chose the breathiness perceptual attribute, a characteristic perceived due to turbulentflow generated in the glottis or due to excessive air escape through a slit. Different degrees of severity are found in patients, varying from mild disturbance levels to extreme cases. The study of vocal breathiness was performed on two fronts: instrumental measurements and auditory perception.

Related to instrumental measures, a platform has been developed to test various methods. This method comparison consists of (i) evaluating synthetic samples of voice with known signal-to-noise ratios, jitter and shimmer; and (ii) to evaluate predominantly breathy real voice samples on a 7-point perceptual scale. Methods known from the literature, such as CPPS (cepstral peak prominence smoothed), SFR (spectral flatness residue signal), HNR (harmonic-to-noise ratio) and S^2NR (spectrographic signal-to-noise ratio) were tested, the latter method being developed in this work. In this study, it was determined that the methods that present the greatest correlation between acoustic (objective) and perceptive (subjective) measurements are CPPS and S^2NR. It was also observed that the S^2NR presents greater robustness to frequency and amplitude perturbations than CPPS, being, in this criterion, a better measure of breathiness.

Regarding perception, the initial step was breathiness psychophysics characterization relating the variations in the physical dimension (glottal noise level) and the perceptive scale. Another result was to obtain the just noticeable differences related to noise intensity variation.

The next objective of this work was establishing a comparative classification method for voice quality, which would require raters' minimal training and prior knowledge, and has high intra-rater and inter-rater agreement. This method was based on the search for elements in a binary tree. The sequence of comparisons follows the structure of the tree,

following until it is not possible to distinguish differences or reach the last level. To do so, reference samples (anchors) are chosen, which are compared to the samples under analysis. Choices of Anchors and search tree depth are made based in breathiness psychophysics. The relative nature of the human ear is thus explored in contrast to current methods of perceptual evaluation in which the ear is treated as absolute. In the developed method, 3 or 7 anchors were chosen, with increasing signal-to-noise ratio, requiring 2 or 3 different comparisons for a single voice sample evaluation. In the experiments, using synthetic vowels and human breathy voice, the comparative evaluation presented high inter-rater reliability, even with inexperienced evaluators.

Key-words: dysphonic voice, acoustic correlates, vocal breathiness, perception, signal-to-noise ratio

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AWGN	*Additive white gaussian noise* - Ruído branco gaussiano aditivo
CPPS	*Cepstral Peak Proeminence smoothed* - Proeminência de pico Cepstral suavizada
FFT	*Fast Fourier Transform* - Transformada rápida de Fourier
GRBAS	*Grade, Roughness, Breathiness, Asthenia, Strain* - Geral, Rugosidade, Soprosidade, Astenia, Tensão
HNR	*harmonic-to-noise ratio* - Relação harmônico-ruído
ICC	*Intraclass correlation coefficients* - coeficientes de correlação intra-classe
jnd	*just noticiable differences* - Mínimas diferenças perceptíveis
LPC	*linear predictive coding* - Codificação por predição linear
MLCM	*Maximum Likelihood Conjoint Measurement* - Medidas conjuntas por máxima verossimilhança
MLE	*Maximum likelihood estimation* - Estimação por máxima verossimilhança
MLDS	*Maximum likelihood difference scaling* - Escalonamento de diferenças por máxima verossimilhança
PA	*Pitch Amplitude*
RPK	*Pearson r at autocorrelation peak* - Correlação de Pearson r no pico de autocorrelação
SFR	*Spectral flatness residue signal* - Planicidade espectral do resíduo
SNR	*Signal-to-noise ratio* - Relação sinal-ruído
S^2NR	*spectrographic signal-to-noise ratio* - Relação sinal-ruído espectrográfica
VPAS	*Vocal profile analysis scheme* - Esquema de análise do perfil vocal

Sumário

1 Introdução

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 requer a criação de instrumentos que traduzam características perceptivas em valores numéricos de atributos como rouquidão, aspereza e soprosidade (KREIMAN; 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 (ROBINSON, 1993). Em geral, este tipo de anomalia decorre de alterações que afetam 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 queixa comum que leva pacientes aos consultórios de otorrinolaringologistas é a sensação de fadiga e desconforto vocal. Alterações deste tipo podem ser investigadas através de exames de vídeo-laríngeo-estroboscopia e a avaliação do especialista indica o tratamento a ser adotado, como por exemplo intervenção cirúrgica ou fonoterapia. Outro tipo de queixa comum é a mudança na voz, percebida pelo próprio paciente (propriocepção) ou por terceiros (ORLIKOFF, 1999). Mesmo sem a sensação de desconforto, pode-se indicar a presença de alterações orgânicas e funcionais, apenas pela percepção da voz do paciente. A alteração de parâmetros acústicos da voz, como a variação de freqüência fundamental ciclo-a-ciclo (referida como *jitter*), variação de amplitude ciclo-a-ciclo (*shimmer*), instabilidades na fonação e ruído em níveis anormais podem ser percebidas e quantificadas, através de medidas instrumentais a partir do sinal acústico da voz e podem também ser indícios de anormalidades orgânicas/funcionais. Esta discussão está esquematizada no diagrama mostrado na Figura 1.1.

Uma forma de se verificar 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 comunidade (ROBINSON, 1993), destacando-se o *Voice Profile Analysis Scheme*(LAVER et al., 1981) e a escala GRBAS (HIRANO, 1981). Na escala GRBAS, por exemplo, os atributos estudados são a rouquidão (R), soprosidade (B), astenia (A) e

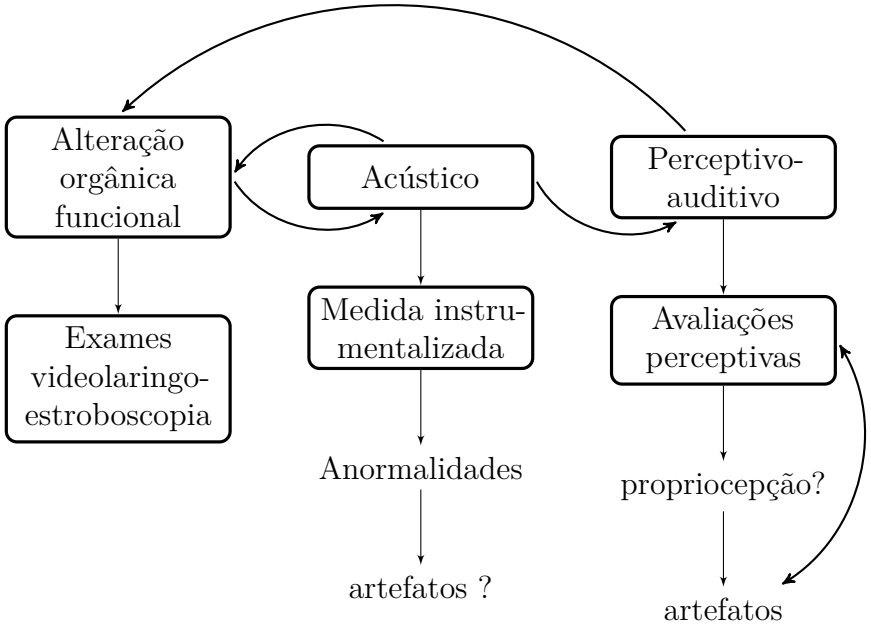


Figura 1.1 – Avaliação da voz

tensão (S).

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; GERRATT, 1998).

As medidas instrumentais 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.

De modo geral, as investigações sobre o uso de medidas acústicas descrevem parâmetros objetivos que (i) discriminam falantes disfônicos dos não disfônicos (LIEBERMAN, 1963; SCHOENTGEN, 1982; HIRANO et al., 1988; ZHANG et al., 2005), (ii) são correlacionadas à avaliação perceptiva da qualidade de voz (KANE; WELLEN, 1985; ESKENAZI; CHILDERS; HICKS, 1990; FEIJOO; HERNANDEZ, 1990; MARTIN; FITCH; WOLFE, 1995), ou (iii) permitem o monitoramento longitudinal dos pacientes (MUTA et al., 1988; DEJONCKERE; WIENEKE, 1994).

A dificuldade com medidas instrumentais tem sido a robustez na análise de voz disfônica (KREIMAN et al., 1993; KREIMAN; GERRATT, 1998). Em geral, as medidas automáticas têm sua confiabilidade e precisão reduzida justamente quando as perturbações a serem medidas aumentam.

A correlação entre as medidas instrumentais e a escalas perceptiva é um tema

controverso entre os especialistas em voz (ORLIKOFF, 1999). Um atributo perceptivo que apresenta correlação com medidas instrumentais é a soprosidade, característica percebida devida ao fluxo turbulento gerado na glote ou devido ao escape de ar excessivo devido a uma fenda. Neste caso, o grau de classificação está correlacionado à relação sinal-ruído da voz (VIEIRA; SANSÃO; YEHIA, 2014; KREIMAN; GERRATT, 1999). Este atributo será o foco do presente trabalho.

No estágio atual, a classificação perceptiva da voz é um componente essencial da avaliação clínica da voz, já que ela é usada para avaliar o sucesso do tratamento. No entanto, ela não pode ser substituída completamente pelas medidas instrumentais pela ausência de ligações teóricas e empíricas destas medidas com a percepção (ORLIKOFF, 1999).

1.1 Objetivos

Os objetivos deste trabalho são:

1. Propor e avaliar medida instrumental de soprosidade robusta às aperiodicidades;
2. Caracterização da psicofísca da soprosidade, isto é, entender a percepção auditiva da soprosidade em relação às variações na dimensão física (nível de ruído).
3. Estabelecer um método de classificação comparativo para a qualidade de voz disfônica, o qual exija mínimo treinamento/conhecimento do julgador e que tenha a maior concordância intra-julgadores e inter-julgadores.

Em relação ao primeiro objetivo, um método baseado na análise de imagens espectrográficas é proposto e seu desempenho relativo às perturbações é avaliado. Em seqüência, o comportamento do método é comparado a outros descritos na literatura.

Em relação ao segundo objetivo, da caracterização da psicofísica da soprosidade, Almeja-se: (i) obter relações entre a dimensão física (nível de ruído glótico) e a escala perceptiva; (ii) Obter relações entre medidas instrumentais e perceptivas de soprosidade; (iii) determinar as *jnd* [1] em relação à variação dos parâmetros acústicos na voz, isto é, estabelecer a mínima variação na perturbação suficiente para ser percebida, indicando níveis distintos na percepção do ruído glótico.

De posse das informações determinadas na caracterização da psicofísica da soprosidade, parte-se para o terceiro objetivo, que é a proposição de um método de avaliação perceptivo-auditiva comparativo. Este método é baseado na busca de elementos em uma

[1] *just noticeable differences* - mínimas diferenças perceptíveis (ZWICKER; FASTL, 1999)

árvore binária. Para tanto, são escolhidas amostras de referência (âncoras), que são comparadas às amostras em análise. As escolhas das âncoras e da profundidade da árvore de busca são feitas com base na psicofísica da soprosidade.

O processo de classificação passa a ser realizado de forma semelhante à afinação de um instrumento musical, onde é ouvida uma referência a qual é ajustada à freqüência do instrumento a ser afinado. As avaliações passam a ser feitas de forma relativa. Explora-se, deste modo, a natureza relativa do ouvido humano (ZWICKER; FASTL, 1999), em contraste com os métodos tradicionais de avaliação perceptiva nos quais o ouvido é tratado como absoluto.

1.2 Relevância da contribuição

- As medidas instrumentais tradicionais da soprosidade são síncronas, isto é, baseadas na demarcação precisa dos ciclos glóticos. Essa demarcação é dificultada na presença de ruído típico da voz soprosa. Em conseqüência, as medidas tradicionais de soprosidade não são confiáveis. Alternativamente, um método baseado na análise de imagens espectrográficas (analogamente ao que é feito pelos profissionais da área (YANAGIHARA, 1967; WOLFE; STEINFATT, 1987)) é potencialmente mais robusto e confiável;

- Este trabalho desenvolve uma plataforma para avaliação de desempenho e limitações de correlatos acústicos da soprosidade, submetendo os métodos a variadas condições de deterioração no sinal de voz (*jitter, shimmer*, ruído), além de diferentes tratos vocais;

- Caracterização da relação de uma dimensão física com a dimensão perceptiva da soprosidade, através da obtenção de escalas (KNOBLAUCH; MALONEY et al., 2008) e da estimativa das funções psicométricas, inferindo as mínimas diferenças perceptíveis (KINGDOM; PRINS, 2010; ZWICKER; FASTL, 1999);

- Um método de classificação comparativa perceptiva irá potencialmente fornecer resultados mais confiáveis, precisos e consistentes entre os profissionais do que os métodos atuais que dependem da referência absoluta (memória interna do avaliador), em semelhança ao chamado ouvido absoluto de um músico (habilidade raramente encontrada) (KREIMAN et al., 1993; KREIMAN; GERRATT, 1998).

1.3 Estrutura do trabalho

O trabalho é desenvolvido em duas frentes: melhores medidas automáticas no sinal acústico e melhores avaliações perceptivo-auditiva do grau de soprosidade da voz.

No primeiro momento, as medidas instrumentais serão tratadas, em especial os correlatos acústicos da soprosidade. A relação sinal-ruído espectrográfica (S^2NR) será descrita e seu desempenho avaliado. Na seqüência, será comparada com outros métodos descritos na literatura. Os variados métodos são baseados na análise de séries temporais, análise cepstral e espectral (HILLENBRAND; CLEVELAND; ERICKSON, 1994; HIRANO et al., 1988; MARYN et al., 2009). O tema será desenvolvido no Capítulo 2.

Em relação à avaliação perceptiva, desenvolve-se inicialmente a caracterização da relação da dimensão física com a dimensão perceptiva, obtendo escala de diferenças de percepção para diversos tipos de estímulo (tons puros, voz sintética, voz natural). Também busca-se caracterizar a função psicométrica da percepção de soprosidade e a obtenção dos limiares de percepção. Os referidos experimentos estão no Capítulo 3.

Na seqüência, um método baseado em avaliações comparativas (relativas) é desenvolvido, que consiste em efetuar uma busca em árvore binária, onde o avaliador é guiado em comparações para se classificar o grau de soprosidade de um determinado estímulo em relação às referências. O método é tratado no Capítulo 4.

2 Medidas instrumentais da soprosidade

Este capítulo é baseado no trabalho publicado na revista Speech Communication (VIEIRA; SANSÃO; YEHIA, 2014).

2.1 Introdução

Uma vasta literatura científica tratando da análise acústica da voz disfônica está disponível. De modo geral, as investigações sobre o uso de medidas acústicas descrevem parâmetros objetivos que (i) discriminam falantes disfônicos dos não-disfônicos (LIEBERMAN, 1963; SCHOENTGEN, 1982; HIRANO et al., 1988; ZHANG et al., 2005), (ii) são correlacionados à avaliação perceptiva da qualidade de voz (KANE; WELLEN, 1985; ESKENAZI; CHILDERS; HICKS, 1990; FEIJOO; HERNANDEZ, 1990; MARTIN; FITCH; WOLFE, 1995), ou (iii) permitem o monitoramento longitudinal dos pacientes (MUTA et al., 1988; DEJONCKERE; WIENEKE, 1994). Estes estudos, em sua maioria, usaram vogais sustentadas, efetuando medidas de *jitter* (perturbações ciclo-a-ciclo da freqüência fundamental), *shimmer* (perturbações ciclo-a-ciclo da amplitude) e relação sinal-ruído (ou harmônico-ruído).

Jitter e *shimmer* contribuem para a sensação perceptiva (na audição) comumente denominada aspereza. O ruído glótico, causado pela passagem do fluxo turbulento de ar devido ao fechamento incompleto da glote, está correlacionado à percepção de soprosidade (LAVER; HILLER; BECK, 1992; KLATT; KLATT, 1990; CHEN et al., 2013). Apesar do amplo uso, as medidas acústicas da disfunção fonatória apresentam comportamento paradoxal (BIELAMOWICZ et al., 1996), isto é, sua confiabilidade é reduzida com níveis crescentes da quantidade a ser medida. Enquanto esta característica não é crítica quando se faz a discriminação de voz disfônica para não-disfônica, na avaliação automática da qualidade de voz ou no monitoramento longitudinal dos pacientes, é necessário o uso de parâmetros acústicos confiáveis que se relacionem com o grau de disfunção vocal. Adicionalmente, a medida de um determinado parâmetro não deve ser afetada pela co-existência de outras perturbações do sinal. Por exemplo, *jitter* e *shimmer* contaminam a maior parte das medidas de relação sinal-ruído (SNR), enquanto ruído glótico infla medidas de perturbações ciclo-a-ciclo (HILLENBRAND, 1987; MUTA et al., 1988; QI, 1992; KROM, 1993; QI et al., 1995; MURPHY, 1999; VIEIRA; MCINNES; JACK, 2002).

A literatura sobre a quantificação de perturbações vocais (ver Murphy (1999) para uma revisão completa) apresenta métodos de estimação da SNR baseados na análise do sinal nos domínios do tempo (e.g. Yumoto, Gould e Baer (1982), Kasuya et al. (1986)), da freqüência (e.g. Hiraoka et al. (1984), Kasuya, Ogawa e Kikuchi (1986)) e do domínio

cepstral (e.g. Krom (1993), Murphy (1999), Murphy e Akande (2007)).

Yumoto, Gould e Baer (1982) foram precursores das medidas da SNR na voz disfônica. No método por eles apresentado, 50 ciclos glóticos sucessivos são sincronizados pela freqüência fundamental e medianizados, de forma a gerar um componente periódico, enquanto o ruído glótico é estimado através da subtração da média dos ciclos de cada ciclo. Variações na duração do ciclos são tratadas através do truncamento da série temporal ou por preenchimento por zeros (*zero-padding*), mas estas estratégias inflam as medidas.

Kasuya et al. (1986) descreveram um algoritmo de sincronização baseado em filtros pente (*comb-filter*) adaptativos. Estes filtros, sintonizados pela medida do período fundamental, casando os "dentes" do filtro pente aos harmônicos do sinal original. Desta forma, a parcela "periódica" do sinal é extraída, enquanto o ruído é obtido através da subtração do sinal original pelo sinal "limpo".

A demarcação dos ciclos glóticos, que na voz disfônica não é confiável, é uma grande limitação dos métodos baseados no domínio do tempo.

Kasuya, Ogawa e Kikuchi (1986) também descreveram um método baseado no domínio da freqüência, onde a cada 7 ciclos glóticos sucessivos, um espectro para o trecho é calculado. Nesta abordagem, a componente de ruído é estimada da energia inter-harmônica enquanto a componente periódica é a energia ao redor dos picos espectrais (sendo necessária a detecção dos picos).

A estimativa correta do ruído nos vales espectrais é o problema central nos métodos baseados na análise no domínio espectral.

No métodos baseados no domínio cepstral, originalmente propostos por Krom (1993), o ruído glótico é estimado após a remoção dos picos cepstrais (componente periódica do sinal).

Também é fato conhecido que os métodos de estimativa da SNR têm bom comportamento com variado nível de ruído aditivo, mas são sensíveis a *jitter* e *shimmer* (YUMOTO; GOULD; BAER, 1982; HILLENBRAND, 1987; MUTA et al., 1988; QI, 1992; KROM, 1993; QI et al., 1995; MURPHY, 1999; MURPHY, 2000), aparato de gravação (TITZE; WINHOLTZ, 1993) e estrutura de formantes das vogais (COX; ITO; MORRISON, 1989).

Embora estratégias para redução destes artefatos já tenham sido descritas (e.g. Hillenbrand (1987), Qi (1992), Qi et al. (1995), Murphy (1999)), valores estimados de SNR não necessariamente são evidência corroborativa de ruído glótico (mas decorrentes de limitações nos métodos na presença de artefatos).

Por outro lado, em clínica, especialistas têm utilizado a inspeção visual de espectrogramas de faixa-estreita para estimar o grau de alteração da disfonia (YANAGIHARA, 1967; WOLFE; STEINFATT, 1987) pois características relevantes, tais como estrutura

harmônica da voz, ruído inter-harmônico, quebras fonatórias e sub-harmônicos podem ser prontamente vistos nestas imagens.

A confiabilidade da informação na imagem espectrográfica da fala é uma das maiores motivações deste estudo, que é descrever um método de medida da SNR baseado no processamento de imagens espectrográficas.

Processamento de imagens da fala tem sido usado para modificação e re-síntese (HORN, 1998), melhoria de inteligibilidade (SOON; KOH, 2003; DING et al., 2009), estimativa de formantes e freqüência fundamental (EZZAT; BOUVRIE; POGGIO, 2007) e codificação da fala (JELLYMAN et al., 2009). A aplicação aqui descrita está baseada em um algoritmo originalmente desenvolvido para melhoramento de imagens de impressão digital (HONG; WAN; JAIN, 1998; BAZEN, 2002).

De forma similar aos espectrogramas, impressões digitais apresentam estruturas de linhas aproximadamente paralelas (formadas por cristas e vales da pele), ruído (causadas por artefatos na coleta, aberrações na formação ou marcas ocupacionais) e minúcias peculiares (sobretudo finalização das cristas e bifurcações). O rastreamento robusto das estruturas das linhas de impressão digital pelo algoritmo motivou seu uso na identificação de estruturas harmônicas nos espectrogramas.

Neste estudo, a implementação fornecida por Kovesi (2005) do algoritmo de Hong, Wan e Jain (1998) foi aplicada para estimativa da SNR em espectrogramas de banda estreita, tanto em voz sintética quanto em voz humana disfônica. O objetivo é obter um parâmetro de voz confiável relacionado ao nível de ruído glótico e independente dos padrões de formantes das vogais e de aperiodicidades vibratórias das pregas vocais.

2.2 Relação sinal-ruído espectrográfica (S^2NR)

A Figura 2.1 mostra uma visão esquemática do método, algoritmo da relação sinal-ruído espectrográfica (*Spectrographic signal-to-noise ratio* - S^2NR). Os vários blocos e variavéis mencionados no diagrama são discutidos nesta seção. De forma breve, a máscara binária M_{signal} (preto se 1, branco se 0) e seu complemento M_{noise}, são obtidos e então superpostos ao espectrograma (multiplicando pixel a pixel da imagem pela máscara) para separar as regiões com harmônicos intensos das regiões ruidosas. O código computacional[1] é baseado no algoritmo original (HONG; WAN; JAIN, 1998), como implementado por (KOVESI, 2005), com as modificações necessárias para o processamento de espectrogramas. O cálculo da relação sinal-ruído, que será detalhado nesta seção, também foi adicionado ao código.

[1] Disponível em <https://github.com/jsansao/s2nr/>

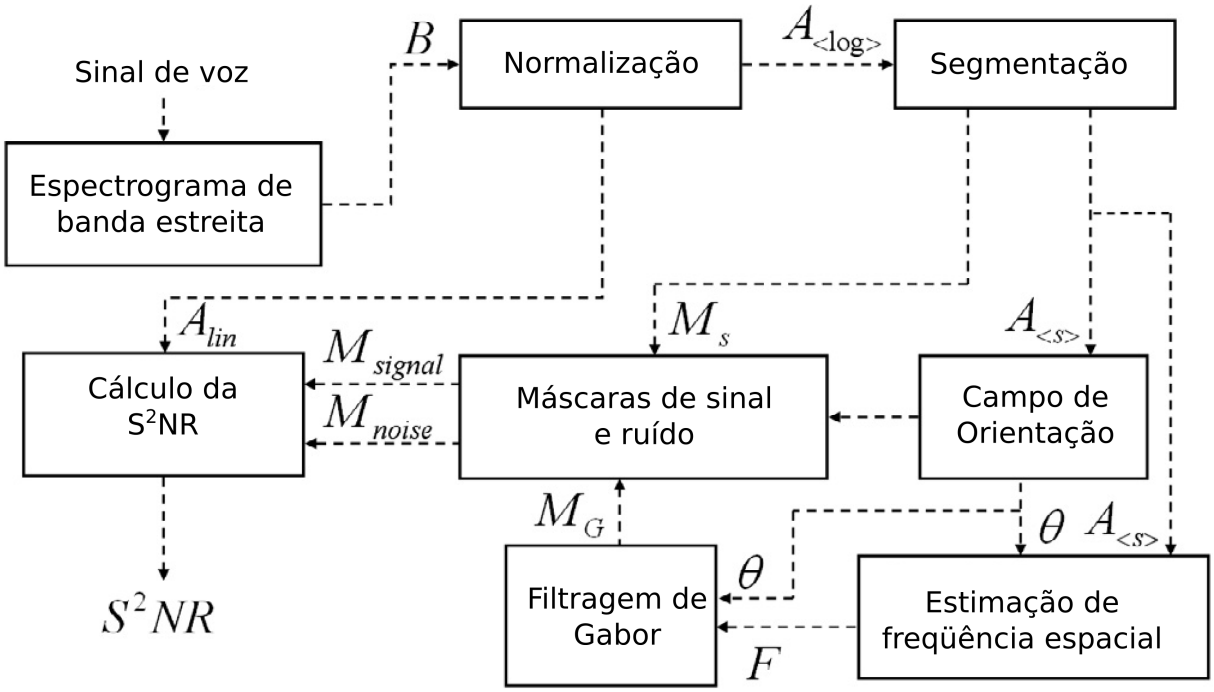


Figura 2.1 – Diagrama esquemático do algoritmo da S^2NR. A entrada é um sinal de voz digitalizado e a saída um valor medianizado no tempo. Os blocos funcionais e variavéis são descritos no texto. Adaptado de Vieira, Sansão e Yehia (2014).

2.2.1 Cálculo da imagem espectrográfica

O sinal de voz é amostrado a 22050 Hz, 16 bits por amostra. O respectivo espectrograma de magnitude, A, é obtido usando a técnica da Transformada Rápida de Fourier (FFT, N pontos, janela de Hamming, 10% de deslocamento temporal). Com $N = 512$, por exemplo, cada pixel em A corresponde horizontalmente a uma resolução temporal de $T_r = 2{,}3\,\text{ms}$ e verticalmente ao intervalo de amostragem espectral $F_r = 43{,}1\,\text{Hz}$. A determinação empírica de N e outros parâmetros relevantes será descrita nas próximas seções.

2.2.2 Normalização

Versões normalizadas da imagem espectrográfica A padronizam a faixa dinâmica e simplificam as rotinas de processamento de imagens. O espectrograma de magnitude linear A_{lin} é dado por

$$A_{lin} = \frac{A - \min|A|}{\max|A|} \tag{2.1}$$

e, similarmente, a versão normalizada logarítmica é

$$A_{log} = \frac{\log|A| - \min(\log|A|)}{\max(\log|A|)}, \tag{2.2}$$

onde o valor mínimo é deslocado para zero. Em seguida, normalizando a amplitude do espectro para ter desvio padrão unitário e média nula, temos A^0_{log}, obtido por

$$A^0_{log} = \frac{A_{log} - \overline{A}_{log}}{\sigma_{Alog}}, \tag{2.3}$$

onde $\overline{A}_{log}$ e σ_{Alog} são respectivamente a média e o desvio padrão do espectrograma em amplitude logarítmica com M quadros:

$$\overline{A}_{log} = \frac{2}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N/2-1} A_{log}(i,j), \tag{2.4}$$

$$\sigma_{Alog} = \sqrt{\frac{2}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N/2-1} [A_{log}(i,j) - \overline{A}_{log}]^2}. \tag{2.5}$$

Nas equações 2.4 e 2.5, o somatório em j é limitado a $N/2-1$ devido a simetria par, pois o sinal é real. O espectro log-normalizado de uma vogal /a/ humana está mostrado na Figura 2.2 (abaixo, direita). Os outros painéis desta figura vão ilustrar o algoritmo da S^2NR nesta seção.

2.2.3 Segmentação

A segmentação mostrada na Figura 2.1 busca por traços de harmônicos no espectrograma. Nesta etapa, a imagem A^0_{log} é particionada em blocos (5 x 5 pixels). Se fragmentos de harmônicos são detectados nos blocos, estes são designados como sinal. Caso contrário, são tratados como ruído. *Grosso modo*, pode-se interpretar esta etapa como uma detecção vozeamento/não-vozeamento executada em cada bloco. No algoritmo original, utilizado em impressões digitais, a classificação é baseada na observação heurística que o desvio padrão (σ_{block}) dos blocos tendo linhas paralelas é maior que o desvio padrão dos blocos com ruído. Para o k-ésimo bloco de A^0_{log}, comparação entre σ_{block} e o limiar empiricamente determinado σ_{th} gera a máscara de segmentação M_s, isto é,

$$M_s(i,j) = \begin{cases} 1, & \text{se} \quad \sigma_{block}(k) \geq \sigma_{th} \\ 0, & \text{c.c.} \end{cases} \tag{2.6}$$

onde i e j varrem as dimensões horizontais e verticais de cada bloco, respectivamente. A máscara está exemplificada na Figura 2.2 (topo, esquerda), onde as regiões em preto indicam porções do espectrograma onde as linhas harmônicas foram identificadas de forma confiável (onde os elementos da máscara têm valor 1).

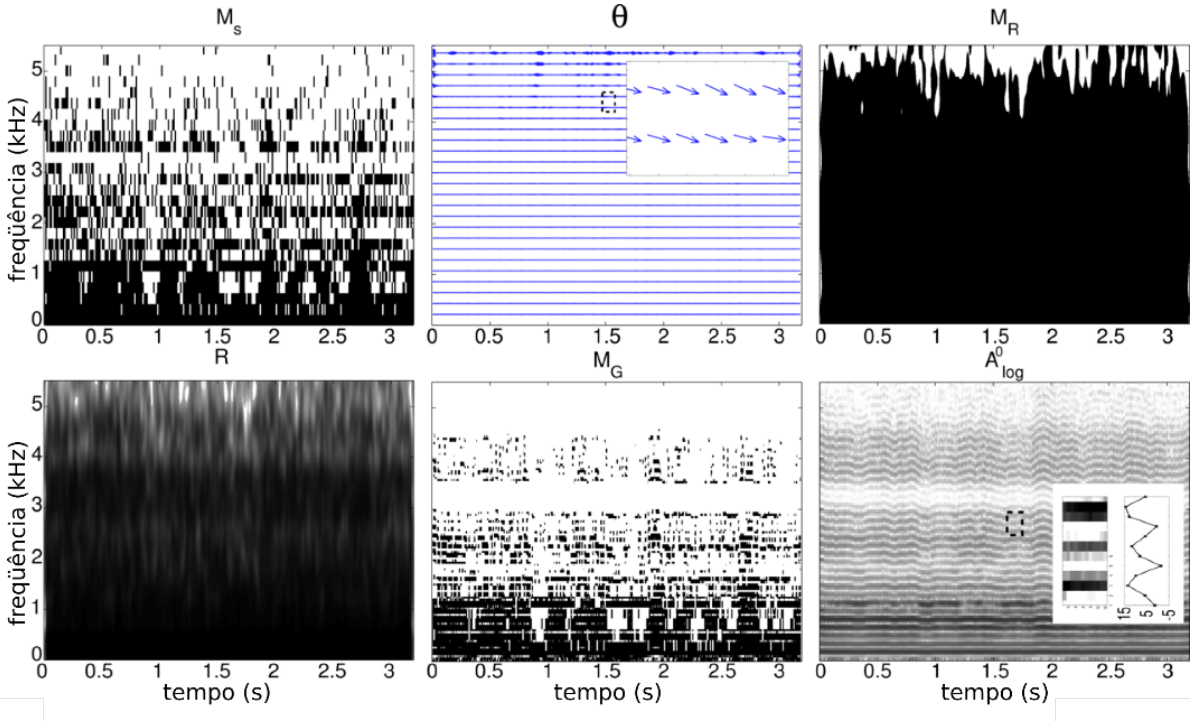


Figura 2.2 – Imagens ilustrativas do algoritmo da S^2NR. Os eixo de freqüência é mostrado até 5,5 kHz para melhorar visualização. (Topo, esquerda) Máscara de segmentação (M_s) com $\sigma_{th} = 0{,}50$. (Topo, centro) Campo direcional das linhas harmônicas baseadas em θ; a figura inserida mostra detalhes das variações de inclinação. (Topo, direita) Máscaras de confiabilidade M_R com $R_{th} = 0{,}60$. (abaixo, esquerda) Índices de confiabilidade R correspondentes. (abaixo, centro) Máscara de Gabor, M_G, sendo neste exemplo $M_G \cong M_{signal}$. (Abaixo, direita) Espectrograma da vogal /a/, em fonação natural. A figura inserida mostra a projeção em tons de cinza das parciais. Adaptado de Vieira, Sansão e Yehia (2014).

A combinação de M_s e A_{log} leva à imagem segmentada A_s, uma imagem com estrutura harmônica aprimorada, definida por

$$A_s = \frac{A_{log} - \overline{A}^1_{log}}{\sigma^1_{Alog}}, \tag{2.7}$$

onde a normalização é similar à Equação 2.3 mas os cálculos de $\overline{A}^1_{log}$ e σ^1_{Alog} incluem apenas as regiões de A_{log} com fragmentos harmônicos, isto é, onde $M_s(i,j) = 1$:

$$\overline{A}^1_{log} = \frac{2}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N/2-1} A_{log}(i,j) M_s(i,j), \tag{2.8}$$

$$\sigma^1_{Alog} = \sqrt{\frac{2}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N/2-1} [A_{log}(i,j) M_s(i,j) - \overline{A}^1_{log}]^2}. \tag{2.9}$$

Comparado com A^0_{log}, A_s terá um valor médio aumentado, desvio padrão aumentado e estrutura harmônica evidenciada. Futuros melhoramentos serão atingidos utilizando os filtros de Gabor sintonizados com a orientação local e a freqüência espacial dos harmônicos. A determinação dos parâmetros dos filtros é descrita a seguir.

2.2.4 Campo de orientação

Nesta etapa, a imagem segmentada A_s é analisada de modo a obter θ, uma imagem com ângulos locais das linhas harmônicas. As orientações dos pixels são dadas pelo vetor gradiente (BAZEN, 2002)

$$\begin{bmatrix} G_i(i,j) \\ G_j(i,j) \end{bmatrix} = \text{sign}(G_i)\nabla A_s(i,j) \tag{2.10}$$

onde i e j são coordenadas verticais e horizontais respectivamente, e sign(x), a função sinal, que retorna 1, se $x > 0$, 0 se $x = 0$ ou -1, se $x < 0$. Numericamente (NIXON; AGUADO, 2002), o gradiente pode ser aproximado por $\nabla A_s \cong A_s \otimes N_{3\sigma}$, onde $\otimes$ é operador de convolução bi-dimensional e $N_{3\sigma}$ é o grandiente da distribuição normal com desvio padrão unitário truncado em 3σ.

A imagem θ na Figura 2.2 (topo, centro) representa a orientação das linhas harmônicas em uma vizinhança e é estimada pelo gradiente quadrático médio (BAZEN, 2002). Os gradientes quadráticos $G_{s,i}$ e $G_{s,j}$ podem ser escritos por

$$G_{s,i} + \jmath G_{s,j} = (G_i + \jmath G_j)^2 = (G_i^2 - G_j^2) + \jmath(2G_iG_j), \tag{2.11}$$

onde $\jmath = \sqrt{-1}$. De forma equivalente, esta equação pode ser expressa na forma matricial como

$$\begin{bmatrix} G_{s,i} \\ G_{s,j} \end{bmatrix} = \begin{bmatrix} G_i^2 - G_j^2 \\ 2G_iG_j \end{bmatrix}, \tag{2.12}$$

e a soma dos gradientes quadráticos em uma vizinhaça é dada por

$$\begin{aligned} \begin{bmatrix} \overline{G_{s,i}} \\ \overline{G_{s,j}} \end{bmatrix} &= \begin{bmatrix} \sum^W G_{s,i} \\ \sum^W G_{s,j} \end{bmatrix} \\ &= \begin{bmatrix} \sum^W (G_i^2 - G_j^2) \\ \sum^W 2G_iG_j \end{bmatrix} = \begin{bmatrix} G_{ii} - G_{jj} \\ 2G_{ij} \end{bmatrix}, \end{aligned} \tag{2.13}$$

onde $G_{ii} = \sum^W G_i^2$, $G_{jj} = \sum^W G_j^2$, $G_{ij} = \sum^W G_iG_j$ e $W = 5 \times 5$ pixels.

Ao se fazer os quadrados, vetores opostos irão apontar para mesma direção, mas os ângulos dos gradientes serão dobrados. Desta forma, após a média, direções dos gradientes deverão ser divididas por dois para ser ter Φ, formando uma imagem de direção do gradiente médio, definida por:

$$\Phi(i,j) = \frac{1}{2}\angle(G_{ii} - G_{jj}, 2G_{ij}) \tag{2.14}$$

onde $\angle(u,v)$ é

$$\angle(u,v) = \begin{cases} \tan^{-1}(v/u), & u \geq 0 \\ \tan^{-1}(v/u) + \pi, & u < 0 \wedge v \geq 0 \\ \tan^{-1}(v/u) - \pi, & u < 0 \wedge v < 0 \end{cases} . \tag{2.15}$$

Finalmente, a imagem θ é calculada através da rotação dos ângulos:

$$\theta(i,j) = \begin{cases} \Phi(i,j) + \pi/2, & \Phi(i,j) \leq 0 \\ \Phi(i,j) - \pi/2, & \Phi(i,j) > 0 \end{cases} . \tag{2.16}$$

onde $-\pi/2 < \theta(i,j) < \pi/2$. Um exemplo está mostrado na Figura 2.2 (topo, centro). Como θ é estimado em cada pixel, esta figura foi desenhada com espaçamento arbitrário para propósitos de visualização. Linhas praticamente paralelas são vistas em toda a figura. Acima de 4,5 kHz, as linhas têm aparência ruidosa devido ao campo de orientação não confiável. A figura inserida mostra as direções locais em uma pequena região, aproximadamente 1,5 s e 4,4 kHz.

A confiabilidade (força ou coerência) na estimativa dos gradientes quadráticos pode ser expressa por (BAZEN, 2002)

$$R(i,j) = \frac{\sqrt{(G_{ii} - G_{jj})^2 + 4G_{ij}^2}}{G_{ii} + G_{jj}}, \tag{2.17}$$

onde $0 \leq R(i,j) \leq 1$. Figura 2.2 (abaixo, esquerda) mostra os valores de R para a imagem θ (topo, centro). Em R, regiões com tons de cinza claro implicam em estimativas não confiáveis de $\theta(i,j)$. Comparações de R com R_{th}, outro limiar empírico, levam a máscara binária de confiabilidade M_R definida por

$$M_R(i,j) = \begin{cases} 1, & \text{se} \quad R(i,j) \geq R_{th} \\ 0, & \text{c.c.} \end{cases} \tag{2.18}$$

Figura 2.2 (topo, direita) mostra a máscara de confiabilidade para o espectrograma ilustrativo. Áreas sem confiabilidade (manchas brancas) são vistas na parte superior de M_R devido aos ângulos ruidosos em θ.

2.2.5 Estimação espacial de freqüência

A distância λ entre os harmônicos é associada com a freqüência fundamental. Para determinar λ, A_s é particionada em grande blocos (16×16 pixels) e os valores dos pixels são projetados ortogonalmente em θ, como mostrado na figura inserida no espectrograma da Figura 2.2 (abaixo, direita). A inserção corresponde a um pequeno fragmento de 1,6 s e 2,5 kHz. O comprimento de onda espacial é estimado pela média das distâncias entre os picos projetados. Se λ está fora da extensão $1 \leq \lambda < 15$ (em termos de freqüência $43.1 \leq f_0 < 646{,}5$ Hz) ou não pode ser medido, por apenas um pico ser detectado, um valor nulo é atribuído. A freqüência espacial, $f = 1/\lambda$, é armazenada na imagem F e utilizada para sintonizar os filtros de Gabor.

2.2.6 Filtragem de Gabor

Filtros de Gabor bi-dimensionais têm seletividade em freqüência e orientação similares aos da visão humana, sendo empregados no processamento de imagens para detecção de bordas (Daugman, 1985). Como detalhado em Daugman (1985), filtros de Gabor atingem o limite téorico mínimo de incerteza em duas dimensões, conseqüentemente tendo resolução ótima espacial e em orientação. Estes filtros têm sido usados em análise bi-dimensional (tempo-freqüência) da fala para rastreamento de f_0 e formantes (EZZAT; BOUVRIE; POGGIO, 2007). No algoritmo aqui investigado, o filtro é aplicado à imagem A_s para gerar E, uma imagem com estrutura harmônica evidenciada e com ruído inter-harmônico atenuado.

A resposta ao impulso do filtro de Gabor é uma função Gaussiana modulada por um onda plana harmônica. A parte real da função de Gabor é um filtro simétrico par, $G(u,v;\theta,f)$, definido como (JAIN; FARROKHNIA, 1991)

$$G(u,v:\theta,f) = \cos(\alpha) \times \exp\left\{-\frac{1}{2}\left[\frac{u_\theta^2}{\sigma_u^2} + \frac{v_\theta^2}{\sigma_v^2}\right]\right\}, \tag{2.19}$$

onde u e v são coordenadas horizontal e vertical cartesianas, respectivamente, $\alpha = 2\pi f u_\theta$, θ é a orientação, f é a freqüência da onda plana, $u_\theta = u\cos\theta + v\,\mathrm{sen}\,\theta$, $v_\theta = -u\,\mathrm{sen}\,\theta + v\cos\theta$ e $\sigma_u = \sigma_v = 0{,}15$ (empírico) são os desvios padrão do envelope ao longo dos eixos horizontal e vertical, respectivamente. De acordo com Hong, Wan e Jain (1998), grandes valores de σ_u e σ_v aumentam a capacidade de redução de ruído mas induzem a incidência de cristas e vales espúrios. No sentido contrário, baixos valores reduzem a formação de vales e cristas, mas a rejeição de ruído é insuficiente.

No algoritmo, a orientação local $\theta(i,j)$ e freqüência espacial local $F(i,j)$ são utilizados para sintonizar o filtro de Gabor em cada pixel (i,j). A convolução da resposta

impulsiva do filtro $\gamma = G[u,v;\theta(i,j),F(i,j)]$ com a imagem segmentada A_s leva a imagem realçada E,

$$E(i,j) = \sum_{u=-w_i/2}^{w_i/2} \sum_{v=-w_j/2}^{w_j/2} \gamma A_s(i-u,j-v), \tag{2.20}$$

onde $w_i = w_j = 6\max(\sigma_i,\sigma_j)$, truncamento é usado pelo fato de maior parte da energia estar confinada no raio de 3 desvios padrão. A imagem E tem média nula e pode ser convertida em uma máscara binária M_G aplicando um limiar nulo, isto e,

$$M_G(i,j) = \begin{cases} 1, & \text{se} \quad E(i,j) \geq 0 \\ 0, & \text{c.c.} \end{cases} \tag{2.21}$$

A máscara de Gabor resultante tem como foco linhas harmônicas confiavelmente detectadas. Na Figura 2.2 (abaixo, centro), M_G pode ser vista como um refinamento da máscara de segmentação M_s. Listras espectrais com harmônicos atenuados próximos à freqûencia de 3,5 kHz e acima de 4,5 kHz (ver também a Figura 2.2, abaixo, direita), foram excluidas. A robustez do método está relacionada à habilidade do operador gradiente rastrear as alterações dinâmicas na estrutura harmônica devido a fatores como deslizamento de f_0 e tremor vocal.

2.2.7 Máscaras de sinal e ruído

A máscara de sinal é obtida pela interseção das máscaras de segmentação, confiabilidade, máscara de Gabor, isto é, $M_{signal} = M_S \cap M_R \cap M_G$. Seu complemento, $M_{noise} = \overline{M_{signal}}$ é a máscara de ruído. A máscara de sinal não é mostrada na Figura 2.2 pois neste caso, $M_R = 1$ em praticamente toda a região onde $M_G = 1$ e desta forma $M_{signal} \approx M_G$. Neste caso também ocorre de $M_s = 1$ na maior parte dos pixels onde $M_G = 1$.

2.2.8 Cálculo da S^2NR

As componentes de sinal e ruído são extraídas aplicando as respectivas máscaras ao espectrograma linearmente escalonado, isto é, $A_{signal} = M_{signal} \cap A_{lin}$ e $A_{noise} = M_{noise} \cap A_{lin}$. Na seqüência, as energias instantâneas de sinal e ruído e os valores da S^2NR foram calculados por

$$S_{signal}(i) = \sum_{j=0}^{N/2-1} [A_{signal(i,j)}]^2, \tag{2.22}$$

$$S_{noise}(i) = \sum_{j=0}^{N/2-1} [A_{noise(i,j)}]^2, \tag{2.23}$$

$$S^2NR(i) = 10 \log \frac{S_{signal}(i)}{S_{noise}(i)}, \tag{2.24}$$

onde i e j varrem os eixos horizontal e vertical respectivamente. Fazendo $t = iT_R$, onde $T_R = 2{,}3\,ms$ é resolução temporal do espectrograma, $S^2NR(t)$ é a série temporal obtida. Para cada estímulo, S^2NR médio foi computado a partir da respectiva série temporal.

2.3 Estímulos e experimentos

2.3.1 Sinais de voz sintética

Sinais de voz sintetizados com relação sinal-ruído controlada foram utilizados para a sintonia do algoritmo e avaliar de forma sistemática sua sensibilidade às variações na freqüência fundamental, *jitter*, *shimmer* e estrutura de formantes. Como detalhado abaixo, vogais sintetizadas /a/, /u/ e /i/ (amostragem de 22050 Hz, 16 bits por amostra) foram criadas convoluindo pulsos glóticos com a resposta ao impulso do trato vocal, a radiação labial sendo modelada por um diferenciador de primeira ordem no fluxo bucal. O modelo de fluxo glótico (FANT, 1979) utilizado nos experimentos tem fator de inclinação (*skew*) de $K = 0{,}65$ ($0{,}5 \leq K \leq 1$). As respostas ao impulso do trato vocal foram estimadas utilizando análise com codificação por predição linear (LPC, *linear predictive coding*), com os parâmetros de 95% de pré-enfase, 22^a ordem, método da covariância (MARKEL; JR, 1976), das vogais /a/, /u/ e /i/ emitidas pelo autor desta tese.

Jitter instantâneo $J(k)$ e similarmente *shimmer* $S(k)$ foram definidos para o k-ésimo ciclo do sinal glótico de acordo com a função de perturbação de primeira ordem,

$$J(k) = \frac{[P + \Delta(k)] - [P + \Delta(k-1)]}{\frac{1}{2}\{[P + \Delta(k)] + [P + \Delta(k-1)]\}}, \tag{2.25}$$

onde P é período médio estipulado e $\Delta()$ são as perturbações instantâneas. Na sintese, $J(k)$ é uma variável aleatória uniformemente distribuída entre $\pm\overline{J}$, onde $\overline{J}$ é a média absoluta do *jitter*. A distribuição do *jitter* tende a ser Gaussiana (HORII, 1979), mas distribuição uniforme foi escolhida por simplicidade em implementação na linguagem C. A seguinte expressão, obtida da equação 2.25, foi utilizada para o cálculo recursivo de $\Delta(k)$ durante a síntese:

$$\Delta(k) = \Delta(k-1)\frac{2 + J(k)}{2 - J(k)} + 2P\frac{J(k)}{2 - J(k)}. \tag{2.26}$$

Perturbação $\Delta(k)$ foi adicionada ao período nominal P para obter o período efetivo $T(k) = P + \Delta(k)$. Períodos instantâneos foram confinado à região $0{,}8P \leq T(k) \leq 1{,}2P$ para evitar crescimento descontrolado de $T(k)$. A série de pulsos glóticos criada, $p(n)$, foi processada para obter a pressão irradiada $s(n)$, isto é, $s(n) = \Delta_1[p(n) * h(n)]$, onde $h(n)$ é resposta ao impulso do trato vocal, $*$ representa convolução e $\Delta_1[]$ é a diferença discreta de primeira ordem.

No experimento principal, o ruído foi adicionado ao sinal $s(n)$ para produzir sinais com relação sinal-ruído conhecidos a priori, independente da filtragem do trato vocal. Um sinal ruidoso $x(n)$ foi criado adicionando uma seqüência aleatória $e(n)$ com média nula, uniformemente distribuída ao ciclos de $s(n)$. A largura W_d da função distribuição de probabilidade foi ajustada em cada ciclo glótico de acordo com a relação sinal-ruído desejada, $SNR_{ref}(k)$ e da potência do sinal livre de ruído, $s(n)$:

$$SNR_{ref}(k) = 10 \log \left[\frac{\frac{1}{T} \sum_{n=0}^{T-1} s^2(n)}{\frac{1}{T} \sum_{n=0}^{T-1} e^2(n)} \right] = \tag{2.27}$$

$$= 10 \log \left[\frac{\frac{1}{T} \sum_{n=0}^{T-1} s^2(n)}{W_d^2/12} \right], \tag{2.28}$$

uma vez W_d determinada. Esta simplificação admite a ergodicidade da seqüência de ruído. O ajuste ciclo-a-ciclo em W_d permite o controle preciso da relação sinal-ruído do sinal sintetizado pois compensa as perturbações da forma de onda causadas pelo *jitter* e *shimmer*. Este tipo de síntese falha em termos de qualidade perceptiva mas permite controle fino das propriedades de interesse nos estímulos. Estudos recentes em avaliação perceptiva da voz soprosa têm usado estímulos que falham neste sentido, de modo a ter controle dos parâmetros relevantes (PATEL; SHRIVASTAV; EDDINS, 2012).

Embora a síntese aqui descrita tenha sido utilizada sistematicamente para avaliar o algoritmo, é oportuno examinar o comportamento do valor da S^2NR em estímulos com ruído aplicados à fonte glótica. Considere fluxo glótico ruidoso $g(n) = p(n) + kr(n)$, onde $p(n)$ é a seqüência de pulsos glóticos como descrito anteriormente, $r(n)$ é ruído branco gaussiano, com média nula e variância unitária representando a fonte de fricção (KLATT, 1980) e k é fator de controle da amplitude do ruído. Fazendo a filtragem do trato vocal em separado de $p(n)$ e $kr(n)$, de acordo com o teorema da superposição, os valores de SNR_{ref} podem ser obtidos a posteriori das componentes quasi-periódicas e de ruído do sinal radiado. Isto está ilustrado na Figura 2.3 (esquerda), com dados das vogais /a/, $f_0 = 220$ Hz, níveis variados de *jitter*, *shimmer* e ruído, o último controlado por k. As linhas verticais nesta figura têm magnitudes aproximadas de 12 dB causadas pelas perturbações vocais, notadamente a falta de ajuste ciclo-a-ciclo do ruído, como pode ser inferido no resultados subseqüentes (Figuras 2.5-2.6). As análises feitas neste estudo requerem um controle mais fino dos valores SNR_{ref} do que o permitido quando adiciona-se o ruído ao

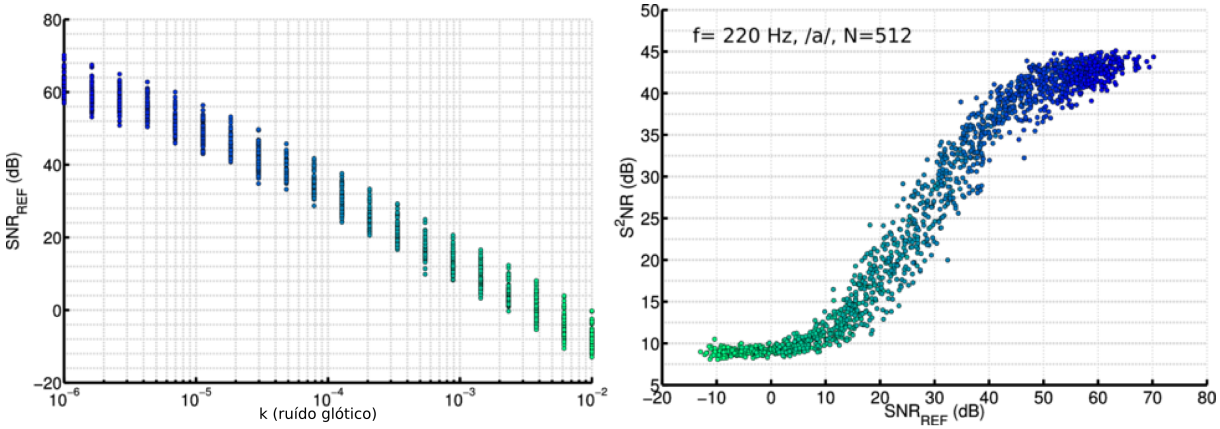


Figura 2.3 – Fluxo glótico com ruído e perturbações. Dados de vogal sintética /a/, $f_o = 120$ Hz, *jitter* (0,00-3,00%, passos de 0,25%) e *shimmer* (0,0-30,0 %, passos de 2,5 %). Parâmetros da S^2NR: N=512, $R_{th} = 0{,}60$ e $\sigma_{th} = 0{,}10$). Nas máscaras M_R, M_G e M_S, as regiões escuras indicam os locais onde o valor é 1. Figura da esquerda mostra a dependência da SNR_{ref} com nível de ruído k e perturbações vocais. Figura da direita mostra a relação da S^2NR e SNR_{ref}. Adaptado de Vieira, Sansão e Yehia (2014).

.

fluxo glótico.

A dispersão dos valores de S^2NR em relação aos valores de SNR_{ref}, baseados no mesmo estímulos mostrados na Figura 2.3 (esquerda), está mostrada na Figura 2.3 (direita). Linearidade é vista na faixa de 10-45 dB de SNR_{ref}, saturação ocorrendo nos extremos superior e inferior. O limite de saturação inferior está relacionado ao ajuste de parâmetros efetuado para que a faixa útil do método coincidisse com a faixa efetiva da voz humana. A saturação superior é causada pela limitação de resolução da transformada de Fourier com uma janela curta ($N = 512$). A Figura 2.3 (direita) dá uma visão geral do algoritmo da S^2NR que será avaliado em maior detalhe nas próximas seções, onde os efeitos individuais de *jitter*, *shimmer* e ruído serão investigados.

2.3.2 Sintonia da S^2NR

Testes preliminares revelaram certa dependência à vogal, perda de precisão para SNR_{ref} acima de 30 dB e sensibilidade a três parâmetros: N (número de pontos da FFT), σ_{th} (limiar da máscara de segmentação) e R_{th} (limiar da máscara de confiabilidade). A influência de N não foi relevante para $\overline{J} < 2\%$. Sinais de voz sinteticamente gerados com relação sinal-ruído e níveis de *jitter* conhecidos foram utilizados para sintonizar o algoritmo. Neste processo, os parâmetros foram empiricamente ajustados de modo a minimizar o desvios entre as medidas de S^2NR e SNR_{ref}. Os sinais utilizados na fase de sintonia foram diferentes dos sinais utilizados na fase de avaliação do algoritmo.

2.3.3 Influência da vogal, f_o e perturbações na S^2NR

Os efeitos do tipo de vogal, freqüência fundamental e perturbações fonatórias na S^2NR foram sistematicamente investigados com vogais sintéticas /a/, /u/ e /i/. Dois conjuntos de sinais foram criados por vogal, com freqüência fundamental média masculina (120 Hz) e feminina (220 Hz). Para cada vogal e f_0, os sinais tinham combinações de *jitter*, *shimmer* e relação sinal-ruído. *Jitter* variou de 0,00% a 3,00% com passos de 0,25%, enquanto o *shimmer* variou de 0,0% a 30,0% com incrementos de 2,5%, estas faixas correspondendo aos valores medidos em voz disfônica (VIEIRA, 1997). Valores conhecidos de ruído foram adicionados ao sinal irradiado em degraus de 5 dB, cobrindo a faixa de SNR de 5-50 dB observados em sinais de voz (MURPHY, 1999).

2.3.4 Influência do falante e do tipo de vogal

Para investigar a influência do falante na sintonia da S^2NR, o método foi testado com dados de outros falantes. Estímulos sintéticos foram criados a partir de vogais /a/, /i/ e /u/ de seis falantes adultos (3 homens, 3 mulheres). Como antes, ruído e perturbações vibratórias foram introduzidas e a S^2NR medida foi comparada com SNR_{ref} conhecida. Os estímulos foram criados com combinações de ruído (5-50 dB, passo de 5 dB), *jitter* (0-3%, passo de 0,25%) e *shimmer* (0-30%, passo de 5%). Adicionalmente, para investigar a influência de outras vogais, o algoritmo da S^2NR foi similarmente testado com um conjunto estendido de vogais do português, vogais produzidas pelo Prof. Maurílio Nunes Vieira.

2.3.5 Efeitos das perturbações nos outros métodos

Em busca de uma metodologia para comparar as medidas acústicas com a percepção geral da qualidade de voz, Maryn et al. (2009) fizeram um estudo sistemático. Como os dados de diferentes estudos envolviam diferentes metodologias e critérios perceptivos diferentes (diferentes protocolos, tamanho e tipo de escala perceptiva, qualidade avaliada, etc), recorreram à meta-análise.

A primeira distinção feita é entre as medidas acústicas realizadas em vogais sustentadas ou em fala encadeada. Medidas em vogal sustentada são bem mais freqüentes (PARSA; JAMIESON, 2001), por alguns fatores, como menor variação da configuração fonatória e dos ajustes glóticos e supra-glóticos durante a elocução, menor presença de efeitos decorrentes da prosódia, articulação e sons não-vozeados. Tais características são interessantes para a implementação computacional; No entanto, perde-se pistas perceptivas que aparecem mais freqüentemente na fala encadeada.

No estudo de Maryn et al. (2009), os critérios estatísticos utilizados foram o número de sujeitos, o número de avaliações feitas com cada medida (se comparada com apenas

um atributo perceptivo ou mais), o coeficiente de correlação (ou coeficiente de correlação ponderado, se mais de uma avaliação foi feita para a mesma medida). O último parâmetro, utilizado para determinar se a medida é homogênea ou heterogênea é o desvio padrão do resíduo. Para determinar as medidas acústicas melhor classificadas, um limiar de 0,6 foi escolhido para a correlação e de 1/4 da correlação para o desvio padrão do resíduo, para a determinação das medidas homogêneas, além de um número de avaliações maior ou igual a dois.

Assim, o estudo aponta que para a vogal sustentada, as melhores medidas acústicas são: *Smoothed cepstral peak prominence* (CPPS), *spectral flatness of the residue signal* (SFRS), *Pearson r at autocorrelation peak* (RPK) e *pitch amplitude* (PA).

As sensibilidades dos métodos CPPS, RPK, PA e SFRS a mudanças de vogal, f_0, *jitter*, *shimmer* foram testados com o mesmo conjunto de dados usados para avaliar o algoritmo da S^2NR. Os métodos citados ofereceram significativa contribuição à análise da voz disfônica, mas limitações, sobretudo devido à existência simultânea de diferentes tipos de perturbação vocal, requerem investigações adicionais (HILLENBRAND; HOUDE, 1996).

Smoothed cepstral peak prominence (CPPS)

No estudo meta-analítico (MARYN et al., 2009), 36 medidas para vogal sustentada e 3 para fala contínua foram comparadas. A medida *Smoothed cepstral peak prominence* (CPPS, Hillenbrand, Cleveland e Erickson (1994)) emergiu como a medida mais robusta do estudo. CPPS não depende da detecção precisa de picos, é adequada para vogais sustentadas e fala contínua e é facilmente implementada. É calculada como a diferença suavizada na amplitude entre o logaritmo do pico do cepstrum e "ruído de fundo" representado por reta da regressão linear.

Pearson r at autocorrelation peak (RPK)

Esta medida, descrita por Hillenbrand, Cleveland e Erickson (1994), é baseada em um rastreador de freqüência fundamental padrão. Este rastreador calcula a correlação do sinal e de sua versão com atraso, para uma faixa de valores de f_0, variando de um valor mínimo ao máximo esperado (60-330 Hz). A implementação utilizada[2], baseada na descrição do trabalho citado, utiliza uma faixa de atraso entre 3,3 ms e 16,7 ms. Em um sinal perfeitamente períodico, o pico da função de autocorrelação ocorre no atraso correspondente à freqüência fundamental. Em um sinal corrompido por ruído, estes picos são menos proeminentes.

[2] Disponível em <https://github.com/jsansao/insmes_tools/blob/master/rpk.m>

O processamento é feito em janelas de 30 ms. Para cada janela, detecta-se o pico da autocorrelação, e após a detecção do valor de atraso correspondente, calcula-se a correlação de Pearson da janela em questão, e sua versão atrasada com o valor detectado.

A limitação deste método é a necessidade de detecção do pico de autocorrelação, dependendo de uma busca de máximo dentro da janela especificada.

Para um sinal perfeitamente periódico, a RPK vale 1,0.

Pitch amplitude (PA)

Neste método (PROSEK et al., 1987), calculam-se os coeficientes de predição lineares do sinal. Com este coeficientes, efetua-se a filtragem inversa do sinal de voz, obtendo o resíduo do sinal.

Calcula-se então a função de autocorrelação normalizada do resíduo (operando em janelas de 100 ms). Buscam-se então os picos da autocorrelação. O valor da função de autocorrelação neste pico é valor da medida na janela, designada *Pitch amplitude*.

Para um sinal perfeitamente periódico, a PA vale 10,0 (na normalização empregada por Prosek et al. (1987), também utilizada na implementação utilizada neste trabalho[3]).

Spectral flatness of the residue signal (SFRS)

A SFRS (MARKEL; JR, 1976) por sua vez é baseada no espectro do resíduo da predição linear. Sua definição, dada em decibéis, é de $SFRS = 10\log_{10}(\epsilon)$, onde $0 \leq \epsilon \leq 1$, e ϵ é a razão da média geométrica com a média da energia espectral. Idealmente, $\epsilon = 1$, para o caso de um espectro perfeitamente plano de ruído não correlacionado. Desta forma, a medida tende a 0 dB com perturbação crescente e a $-\infty$ com periodicidade crescente.

2.3.6 Correlação com a soprosidade

Para estudar a relação entre os valores da S^2NR e das outras medidas citadas com a avaliação perceptiva da soprosidade, serviu-se de um banco de dados formado pela vogal sustentada /a/ de 21 indivíduos adultos. As gravações, com três amostras por nível de alteração foram selecionadas do banco de dados elaborado por Vieira (1997) e foram perceptualmente avaliados em uma escala de 7 pontos (ausente, pequeno, moderado, extremo, com níveis intermediários). A classificação perceptiva foi feita por decisão consensual do Prof. Maurílio Nunes Vieira e de duas estudantes do Curso de Fonoaudiologia da Universidade Federal de Minas Gerais (Ana Paula da Penha e Mariana Borges). As gravações selecionadas (22050 amostras por segundo, 16 bits por amostra; Vieira, McInnes e Jack (2002)) tinham duração superior a 3 segundos, predominantemente soprosas e

[3] Disponível em <https://github.com/jsansao/insmes_tools/blob/master/pitchamp.m>

tinham o nível de soprosidade praticamente estável ao longo da elocução. Em falantes disfônicos, é comum a ocorrência de instabilidades fonatórias, principalmente nas amostras de soprosidade extrema.

Para cada gravação, foram calculados os valores médios de S^2NR, CPPS, RPK, PA, SFRS e adicionalmente o valor da relação harmônico-ruído (HNR, Krom (1993)) e comparados com os respectivas avaliações de soprosidade. A estimação do método HNR, conforme descrita por Krom (1993), embora freqüentemente citada na literatura, não foi incluída na meta-análise (MARYN et al., 2009) pois em seu estudo usou sinais sintéticos. O valores da HNR são estimados pelo uso de um filtro pente (comb-lifter) no componente periódico do cepstrum (energia harmônica) e comparando com energia do piso do ruído. Os dados aqui mostrados foram obtidos usando o código computacional disponibilizado por Shue (2010).

2.3.7 Aplicação em fala corrente

O uso do método da S^2NR é viável para uso em fala corrente, pois não requer rastreamento preciso da f_0. Uma demonstração desta possibilidade será dada e suas atuais limitações brevemente abordadas.

2.4 Resultados e discussão

2.4.1 Sintonia do algoritmo

O algoritmo da S^2NR foi empiricamente sintonizado. As influências de N (intervalo da amostragem espectral da FFT), σ_{th} (limiar da máscara de segmentação) e R_{th} (limiar da máscara de confiabilidade) são ilustrados na Figura 2.4, com dados de uma vogal /a/ sintética com $f_0 = 220$Hz, $\overline{S} = 0\%$ (média do *shimmer*), $\overline{J} = 2\%$ (valor médio absoluto de *jitter*) e $SNR_{ref} = 30$ dB. Figura 2.4 (direita) é um histograma com a distribuição do desvio padrão dos blocos 5 x 5 no espectrograma do sinal de teste. Os valores de σ_{block} são concentrados no intervalo de 0,10-0,6 aproximadamente e pico simples na distribuição não fornece uma clara indicação para o valor adequado de σ_{th}. Análises futuras são necessárias.

Na Figura 2.4 (esquerda), os efeitos de R_{th} e N são mostrados no par superior de curvas, onde $\sigma_{th} = 0{,}10$ em ambas linhas e o eixo horizontal representa R_{th}. Com um σ_{th} tão pequeno, a máscara de segmentação M_s terá valores unitários na maior parte das regiões e consequentemente, quase todo o espectrograma será processado pelas rotinas de determinação de campo de orientação e da filtragem de Gabor. A influência de R_{th}, o limiar para confiabilidade do cálculo do campo de orientação, pode assim ser provada. Como visto nas curvas superiores, medidas de S^2NR superestimaram o valor de referência

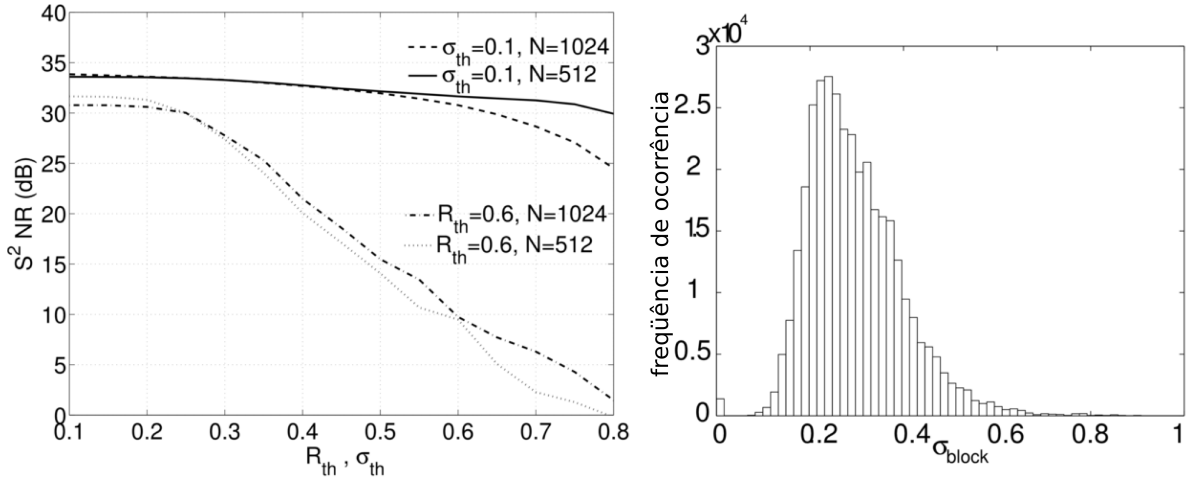


Figura 2.4 – Calibração. Dados obtidos com vogal /a/ sintetizada, com $f_0 = 220$ Hz, $SNR_{ref} = 30$ dB, *shimmer* $= 0\%$, *jitter* $= 2\%$. Figura da esquerda: σ_{th} e R_{th}, e N. No par superior de curvas, o eixo horizontal mostra R_{th}. Para o par inferior, o eixo horizontal é σ_{th}. Figura da direita mostra a distribuição de σ_{th} no sinal. Adaptado de Vieira, Sansão e Yehia (2014).

($SNR_{ref} = 30$ dB) para $R_{th} < 0{,}80$ (com $N = 512$) ou $R_{th} < 0{,}63$ (com $N = 1024$). As curvas superiores também mostram que a influência de N foi irrelevante para $R_{th} < 0{,}50$.

As influências de σ_{th} e N estão ilustradas no par inferior de curvas da Figura 2.4 (esquerda), onde $R_{th} = 0{,}60$ e o eixo horizontal representa variações em σ_{th}. Independentemente de N, ocorreu uma visível queda nos valores das medidas para $\sigma_{th} > 0{,}25$. Para $\sigma_{th} < 0{,}25$, os valores de S^2NR foram próximos de SNR_{ref} com viés inferior a 0,8 dB ($N = 1024$) ou menor que 1,6 dB ($N = 512$). Como será justificado adiante, perturbações no sinal, principalmente o *shimmer*, tenderam a reduzir a medida da S^2NR com valores elevados de SNR_{ref}. Para propósitos de sintonia, esta redução pode ser parcialmente compensada pelo viés observado por $\sigma_{th} < 0{,}25$.

Levando em conta as considerações anteriores, o algoritmo foi sintonizado com $N = 512$, $R_{th} = 0{,}60$ e $\sigma_{th} = 0{,}10$ para os experimentos restantes. Com tais valores, a influência da máscara de segmentação foi minimizada e maior importância foi dada à máscara de Gabor M_G, onde residem os estágios mais sofisticados do algoritmo.

2.4.2 S^2NR: avaliação com voz sintética

As medidas da S^2NR das vogais sintetizadas /a/, /u/, e /i/ são mostradas nas Figuras 2.5, 2.6 e 2.7, respectivamente. Os gráficos são similares aos usados nos estudos relacionados (e.g. Cox, Ito e Morrison (1989), Qi et al. (1995), Murphy (1999)). Em cada figura, um ponto representa um estímulo dos 1690 sinais de teste (13 níveis de *jitter* por 13 níveis de *shimmer* por 10 níveis de SNR_{ref}). Medidas ao longo de uma linha quase horizontal representam sinais sintetizados com uma mesma relação sinal-ruído de referência

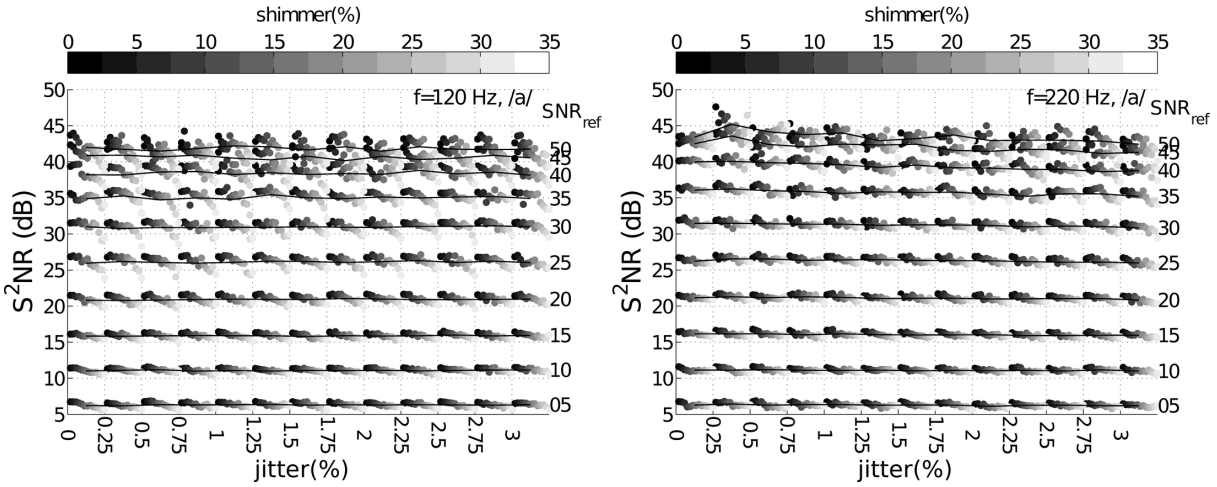


Figura 2.5 – Medidas usando S^2NR em vogais /a/ sintéticas. Figura da esquerda: voz sintética masculina ($f_0 = 120$ Hz). Figura da direita: Voz sintética feminina ($f_0 = 220$ Hz). Cada ponto, em tom de cinza, representa um sinal sintético. *Shimmer* é indicado repetidamente na horizontal superior e como os diferentes tons de cinza. O valores de referência SNR_{ref} são indicados na auxiliar vertical a direita. Sinais com mesmo SNR_{ref} são conectados por linhas aproximadamente horizontais. Adaptado de Vieira, Sansão e Yehia (2014).

(SNR_{ref}, no eixo da direita). Variações nos níveis de *shimmer* e *jitter* são combinados no mesmo gráfico. Valores de *jitter* são indicados na escala horizontal, enquanto os níveis de *shimmer* são representados pelo tom de cinza dos pontos, como mostrado na escala horizontal superior. Os pontos foram também deslocados para evitar superposição. Para cada nível de *jitter* (0%, 0,25%, etc.), os pontos varrem o intervalo de *shimmer* de 0%-30%.

Idealmente, as linhas (médias) através das medidas nas Figuras 2.5, 2.6 e 2.7 deveriam ser horizontais e igualmente espaçadas, indicando independência da S^2NR em relação às perturbações. Uma queda nos valores de S^2NR com *shimmer* crescente é vista em todas as vogais e níveis de *jitter*. Na Figura 2.5, variações na f_0, *shimmer* e *jitter* não afetam a maior parte das medidas em /a/ e as linhas horizontais têm aproximadamente a mesma distância, até $SNR_{ref} = 40$ dB. Em termos de erros de estimação, a diferença $\Delta = S^2NR - SNR_{ref}$ pode ser estatisticamente descrita ao longo de cada linha SNR_{ref}, com $\Delta = \overline{\Delta} \pm \sigma_\Delta$, onde $\overline{\Delta}$ e σ_Δ são o viés e seu desvio padrão. Na Figura 2.5, $\overline{\Delta} < 1{,}3$ dB e $\sigma_\Delta < 0{,}5$ dB para $SNR_{ref} < 35$ dB. A medida máxima aceitável parece ser $SNR_{ref} = 40$ dB, onde $\overline{\Delta} = -1{,}6$ dB e $\sigma_\Delta = 0{,}6$ dB. Erros sistemáticos similares foram descritos na literatura (COX; ITO; MORRISON, 1989). A dispersão dos valores ao longo da linhas horizontais na Figura 2.5 é notadamente menor que magnitude dos segmentos verticais na Figura 2.3 (esquerda).

A Figura 2.6 mostra que os resultados para a vogal /u/ foram similares aos de /a/, exceto em $f_0 = 220$ Hz, baixo *jitter* ($\leq 0{,}25\%$) e alta SNR_{ref} (≥ 45 dB), onde concordância entre os valores da S^2NR e SNR_{ref} aumentou. O pior cenário ocorreu na

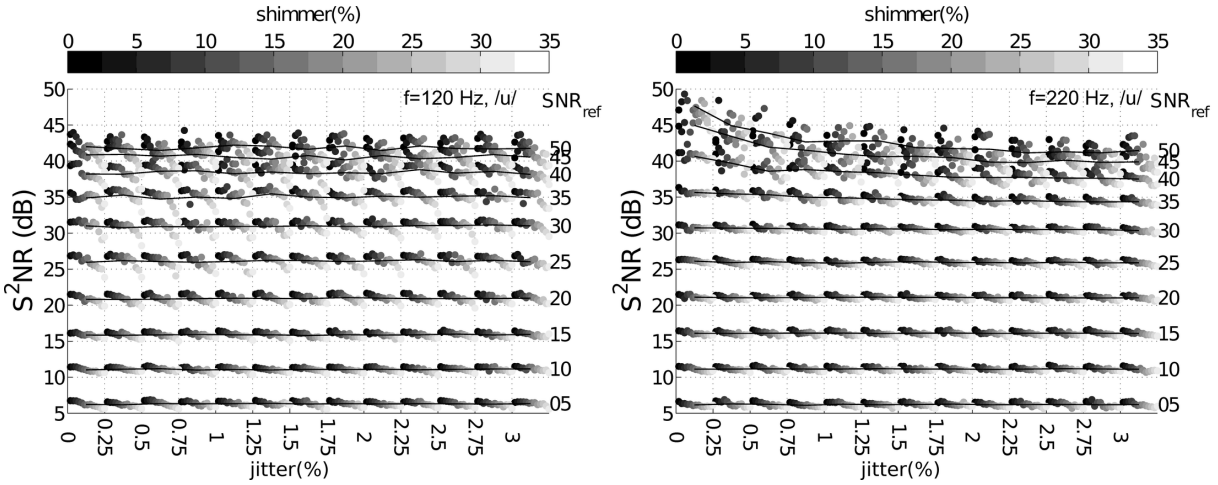


Figura 2.6 – Medidas da S^2NR com vogal /u/ sintética. Resultados similares aos da vogal /a/. Adaptado de Vieira, Sansão e Yehia (2014).

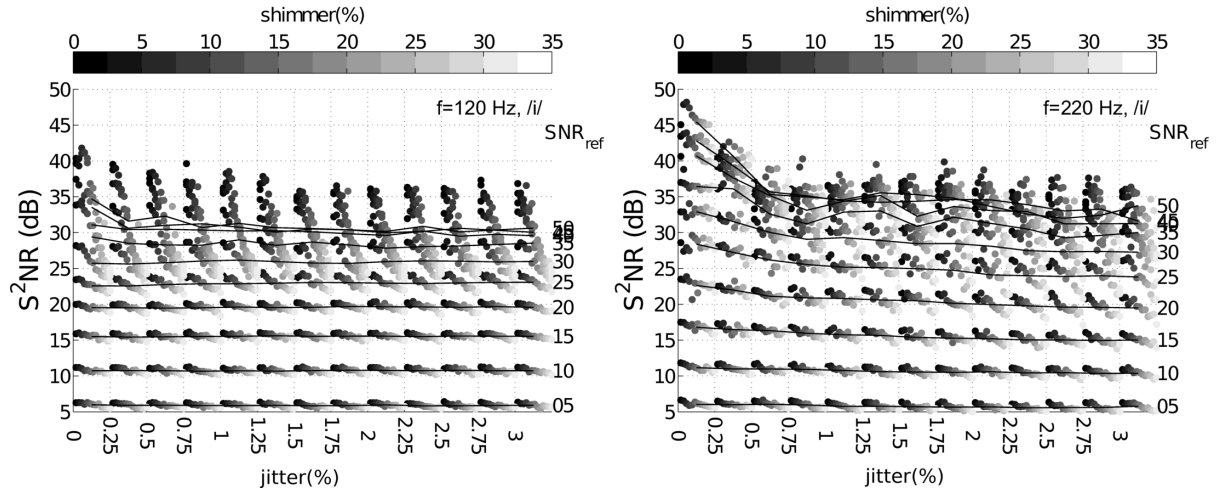


Figura 2.7 – Medidas da S^2NR com vogal /i/ sintética. Este conjunto apresenta pior resultado da avaliação. Adaptado de Vieira, Sansão e Yehia (2014).

vogal /i/ (Figura 2.7), onde os valores de S^2NR se tornam crescentemente comprimidos até atingir saturação em aproximadamente 30 dB (painel da esquerda) ou 35 dB (painel da direita). Nesta vogal, ocorreu sensibilidade crescente ao *shimmer*, notadamente para $SNR_{ref} > 25$ dB. Comparada com /a/ e /u/, a queda na precisão observada em /i/ é atribuída ao vale espectral relativamente largo e ruidoso entre os formantes F1 e F2. Conseqüentemente, os harmônicos entre F1 e F2 não foram detectados pela segmentação e máscaras de Gabor, e relação sinal-ruído foi sub-estimada. Isto sugere futuras melhorias do método, através de adaptação de σ_{th} e R_{th}. For exemplo, para cada instante de tempo, estes limiares podem seguir o envelope de uma fatia espectral da LPC.

Em resumo, a avaliação destas vogais sintetizadas sugere que para /a/ e /u/, o método foi robusto às perturbações para medidas da S^2NR na faixa de 5-40 dB. Como indicado pela dispersão dos valores ao longo do eixo de *jitter*, medidas deterioram para $SNR_{ref} > 40$ dB, em /a/ e /u/, e para $SNR_{ref} > 30$ dB, em /i/. O algoritmo foi

surpreedentemente melhor em baixos níveis de SNR_{ref} e degradação em altos valores de SNR_{ref} foi causada principalmente por *shimmer*. Desta forma, o método da S^2NR não apresentou comportamento paradoxal, principalmente quando observa-se as regiões com relação sinal-ruído mais baixa.

2.4.3 S^2NR: influência da vogal e do falante

O algoritmo da S^2NR foi avaliado com sinais sintéticos baseados em tratos vocais de seis outros falantes adultos. Resultados estão demonstrados nas Figuras 2.8 e 2.9, onde as estimativas de /i/ e /u/ são traçados em relação as medidas de /a/. Alguns detalhes, como a escala de graduação das perturbações foram perdidas nestas figuras, devido à grande quantidade de dados. Idealmente, as medidas deveriam formar aglomerados compactos ao longo da linha tracejada (reta 1:1). Cada aglomerado corresponde a nível de relação sinal-ruído e inclui 21 medidas (3 falantes por 7 níveis de perturbação).

Na Figura 2.8, onde a influência de *jitter* é investigada, os aglomerados são vistos abaixo de 35 dB em todos painéis. Acima deste valor, ocorre uma dispersão com diferentes comportamentos entre os falantes e as vogais. O agrupamento é melhor nos painéis inferiores da Figura 2.8 (falantes do sexo feminino) mas, como visto no painel inferior/esquerda, medidas do /i/ do falante f1 saturam em aproximadamente 35 dB, desviando da linha tracejada. Em geral, considerando as medidas na faixa de 5-40 dB, erros de estimação ($\Delta = S^2NR - SNR_{ref}$) na Figura 2.8 têm valor médio $\overline{\sigma} = 0{,}5$ dB e desvio padrão $\sigma_\Delta = 1{,}7$ dB.

A influência do *shimmer* é mostrada na Figura 2.9. Nos painéis inferiores (falantes do sexo feminino), aglomerados compactados são vistos em até aproximadamente 40 dB. Os painéis superiores mostram maior dispersão nas vozes de falantes do sexo masculino. Em particular, algumas medidas da vogal /a/ do falante m2 foram sub-estimadas em até 13,9 dB na faixa de 25-40 dB de SNR_{ref} e *shimmer* extremo (30%). Uma causa clara para a deterioração da medida em vogais com baixo f_0 (120 Hz) e altos níveis de *shimmer* não foi determinada. Em resumo, considerando os sinais com *shimmer* (Figura 2.9), medidas na faixa de 5-40 dB tiveram erro de estimação médio de $\overline{\Delta} = -0{,}6$ dB e desvio padrão de $\sigma_\Delta = 3{,}2$ dB.

A influência de outras vogais, com dados baseados no trato vocal do Prof. Maurílio Nunes Vieira, está mostrada na Figura 2.10. Nesta figura, as vogais estão indicadas por diferentes símbolos e os respectivos níveis de *shimmer* são codificados em tons de cinza. Os resultados tiveram menor dependência nas vogais e *shimmer* que nas análises anteriores (Figuras 2.6-2.9). Em particular, a degradação crescente na vogal (ou em outra vogal frontal) vista anteriormente não foi observada na Figura 2.10. Nesta avaliação, medidas foram precisas dentro de uma região de 2,5 dB, independemente da vogal, para SNR_{ref} de até 40 dB.

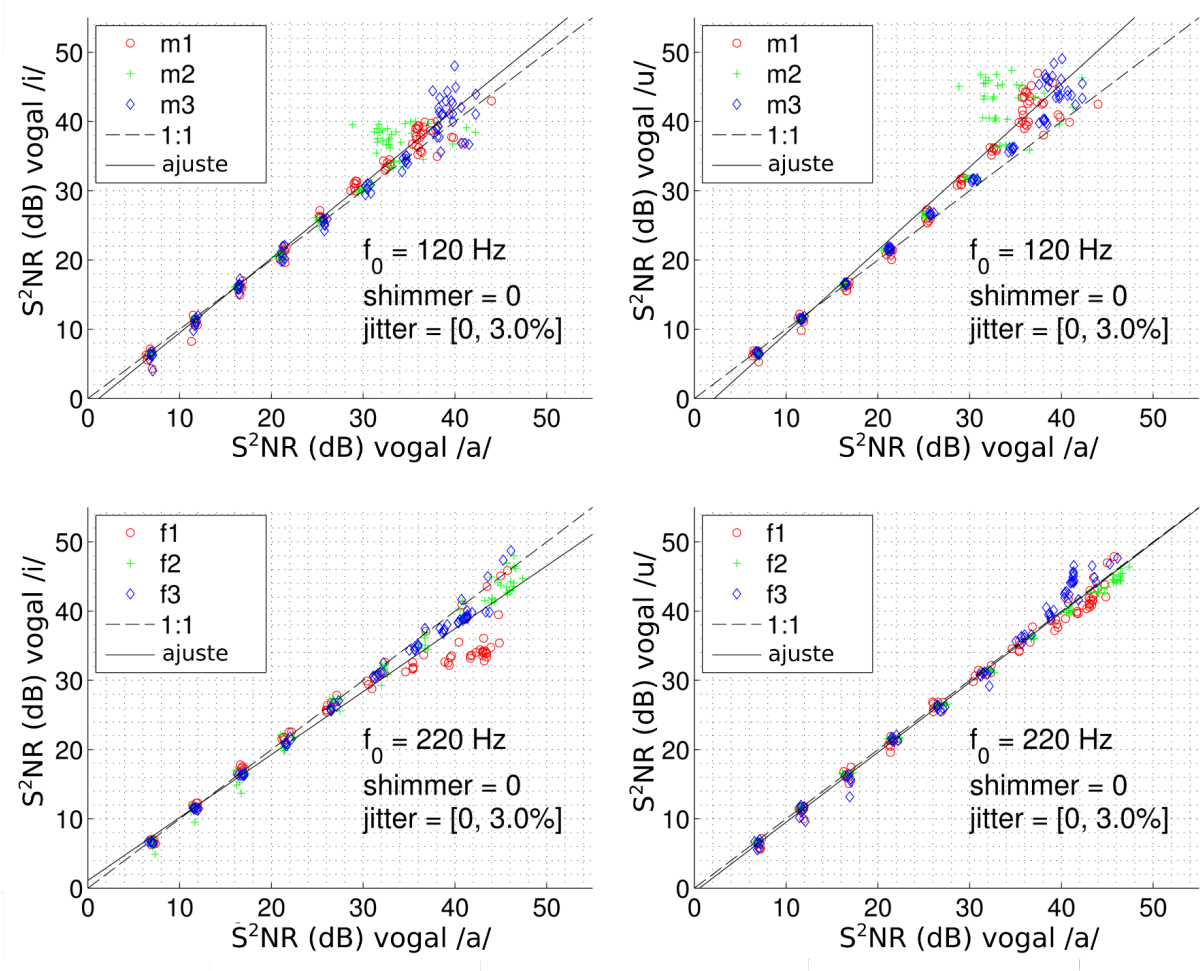


Figura 2.8 – Avaliação da dependência do falante. Dados de voz masculina (m1, m2, m3) e feminina (f1,f2,f3) para falantes sem *shimmer*. Adaptado de Vieira, Sansão e Yehia (2014).

2.4.4 Efeitos das perturbações nos outros métodos

Os algoritmos indicados pela meta-análise foram avaliados com o mesmo conjunto de dados que gerou as Figuras 2.5-2.7 (sensibilidade a vogal, f_0 e perturbações). Resultados representativos estão mostrados na Figura 2.11 para a vogal /a/ sintetizada em voz masculina ($f_0 = 120Hz$). Nos gráficos, *jitter* está representado nas ordenadas e o nível de *shimmer* está codificado em tons de cinza. O mesmo comportamento visto na Figura 2.11 ocorreu nas outras combinações de f_0 e vogais. Todos os métodos apresentaram forte dependência às perturbações no sinal, considerando que os valores traçados não estão distribuídos ao longo de linhas quase horizontais e em linhas igualmente espaçadas. A dependência em *jitter*, vista na Figura 2.11, não ocorreu na S^2NR (Figura 2.10). Dentre os métodos avaliados na Figura 2.11, CPPS apresentou o melhor resultado, notadamente para níveis de *jitter* acima de 1%.

Contrariamente aos outros métodos, S^2NR não depende da detecção de pico que também move com o *jitter* do sinal, sobretudo os picos cepstrais (CPPS) ou picos de

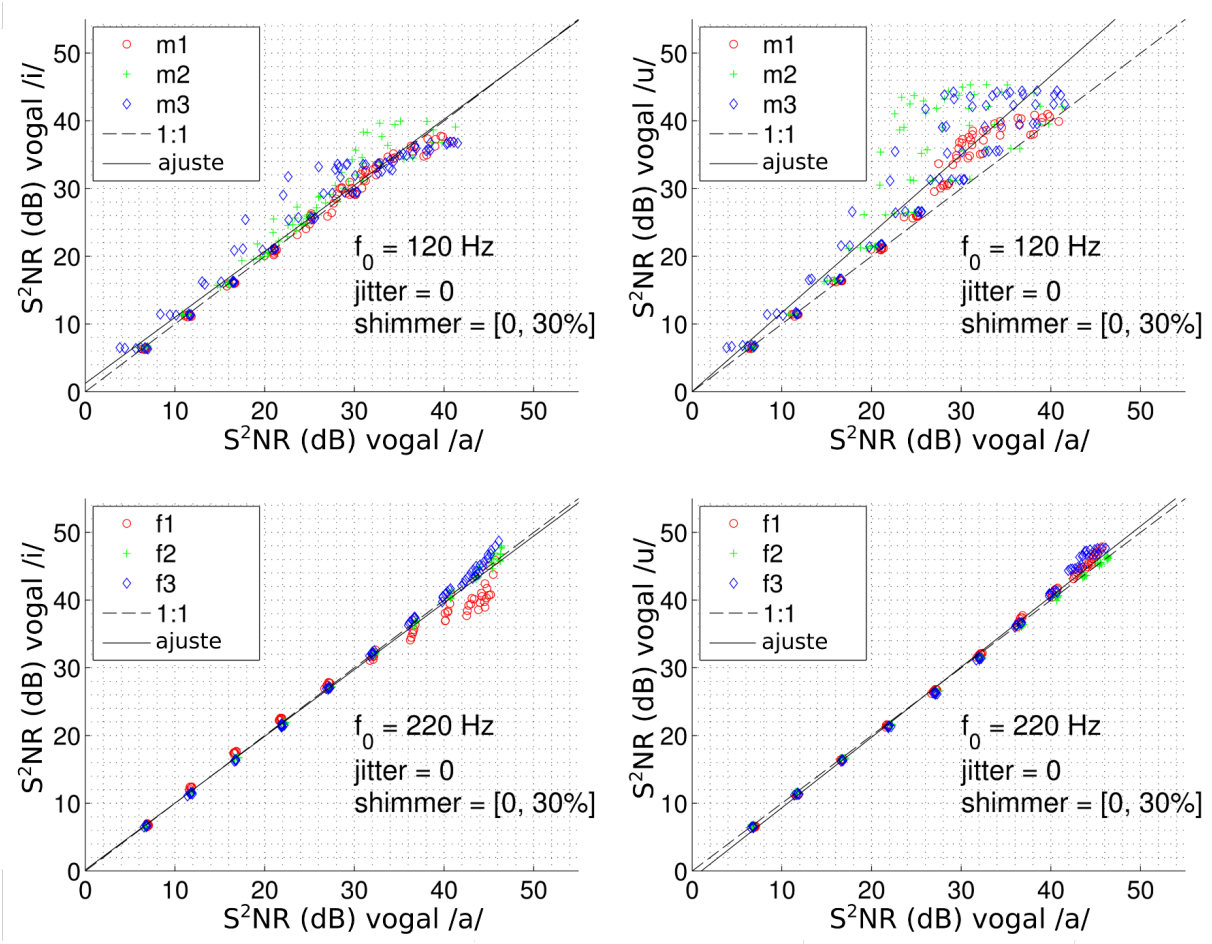


Figura 2.9 – Avaliação da dependência do falante. Dados de voz masculina (m1, m2, m3) e feminina (f1,f2,f3) para falantes sem *jitter*. Adaptado de Vieira, Sansão e Yehia (2014).

autocorrelação (RPK, PA). A sintonia dos filtros-pente cepstrais (comb-lifters) no método da HNR (KROM, 1993) também depende da identificação de picos cepstrais. Outra vantagem da S^2NR reside na identificação robusta das estruturas harmônicas pelos filtros de Gabor. O processamento é realizado em cada bloco de imagem, sem necessidade de rastreamento de f_0 ou método de busca de múltiplos da f_0 medida, como visto nos métodos espectrais de medida de relação sinal-ruído (KASUYA; OGAWA; KIKUCHI, 1986).

2.4.5 Predição da soprosidade: S^2NR e outros métodos

Medidas de vogal /a/ disfônica com diferentes níveis de soprosidade são mostradas na Figura 2.12, onde a série temporal da $S^2NR(t)$ é traçada sobre os respectivos espectrogramas. Esta figura permite a comparação entre as medidas e as características espectrais. Na Figura 2.12 (topo, esquerda), de um falante do sexo masculino com soprosidade grau 0,0 (ausente), $S^2NR(t)$ apresentou os maiores valores da medida no conjunto, flutuando entre 35-43 dB. Na Figura 2.12 (topo, direita), de uma voz feminina com soprosidade grau 2,5 (moderado-extremo), os valores da S^2NR foram próximos de 20 dB, mas quedas

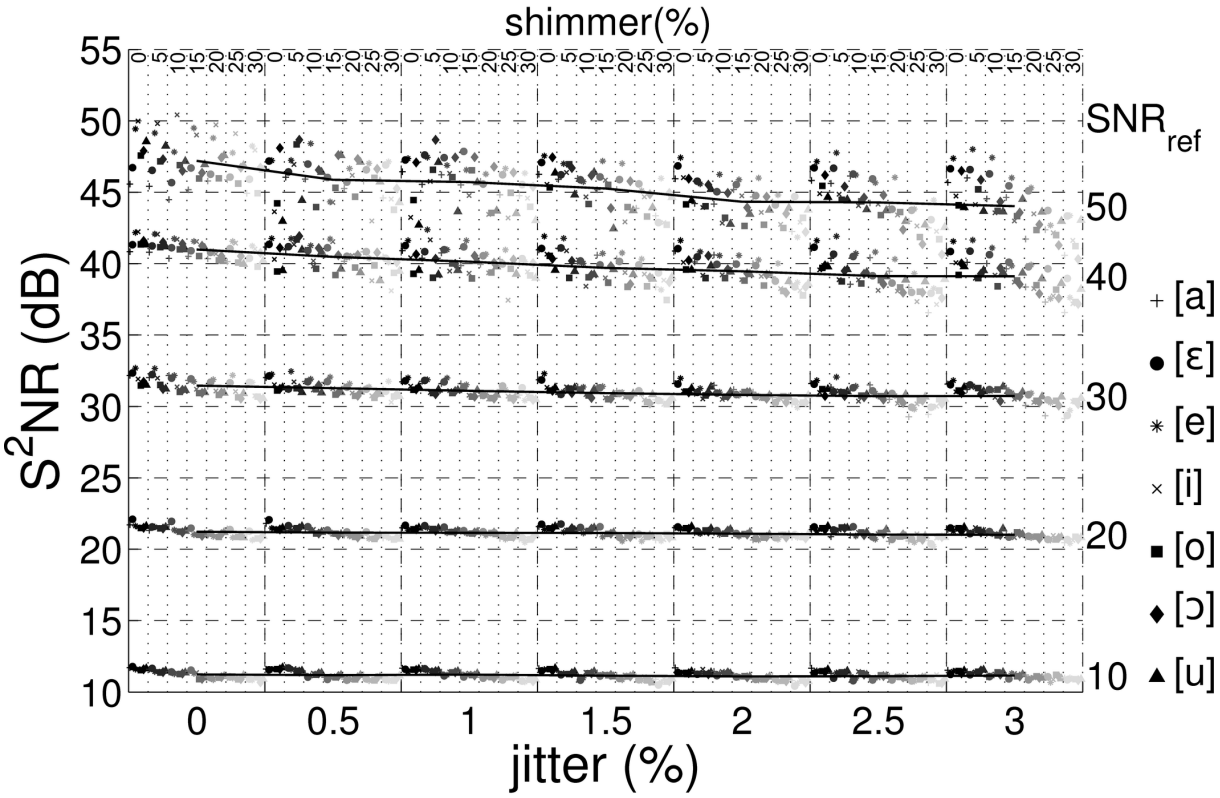


Figura 2.10 – Avaliação da dependência de vogal. Neste gráfico, cada símbolo representa uma vogal específica e o tom em cinza representa o *shimmer*. Estímulos sintetizados para vogais do português. Adaptado de Vieira, Sansão e Yehia (2014).

surgiram devido a quebras na voz, vistas no espectrograma, em 0,8 s, 1,6 s e 2,4 s. Neste exemplo, o espectrograma também mostrou baixa energia em harmônicos superiores a 1 kHz. Na Figura 2.12 (abaixo, esquerda), uma vogal masculina com soprosidade grau 2,0 (moderado), os valores da S^2NR oscilaram na vizinhança de 20 dB e o espectrograma mostra harmônicos atenuados acima de 2 kHz. Finalmente, Figura 2.12 (abaixo, direita), de uma vogal feminina com soprosidade grau 3,0 (extremo), os valores de S^2NR oscilaram entre 0 e 20 dB, e o espectrograma mostra principalmente ruído e estrutura harmônica esporádica abaixo de 1 kHz. Em resumo, os vários painéis mostram uma relação inversa entre a energia do ruído vista nos espectrogramas e o valor médio de $S^2NR(t)$.

A Figura 2.13 mostra a relação entre as medidas e a avaliação perceptiva da soprosidade em amostras de voz humana. Com exceção da RPK, que tem um *plateau* na faixa de soprosidade de 0,0-2,0, resultados indicam que a soprosidade pode ser predita a partir das medidas através do ajuste de um polinômio de 2^o grau. Na média, S^2NR (topo, esquerda) decai não linearmente, mas monotonicamente com soprosidade de grau B de acordo com o polinômio ajustado $S^2NR = -3{,}3500B^2 + 1{,}0114B + 40{,}9840$. Alternativamente, soprosidade pode ser predita das medidas usando o polinômio inverso, com $B = -0{,}0027(S^2NR)^2 + 0{,}0565(S^2NR) + 2{,}7130$. A medida na Figura 2.13, HNR35, é baseada na banda de 0-3500 Hz e apresentou os melhores resultados consideradas as

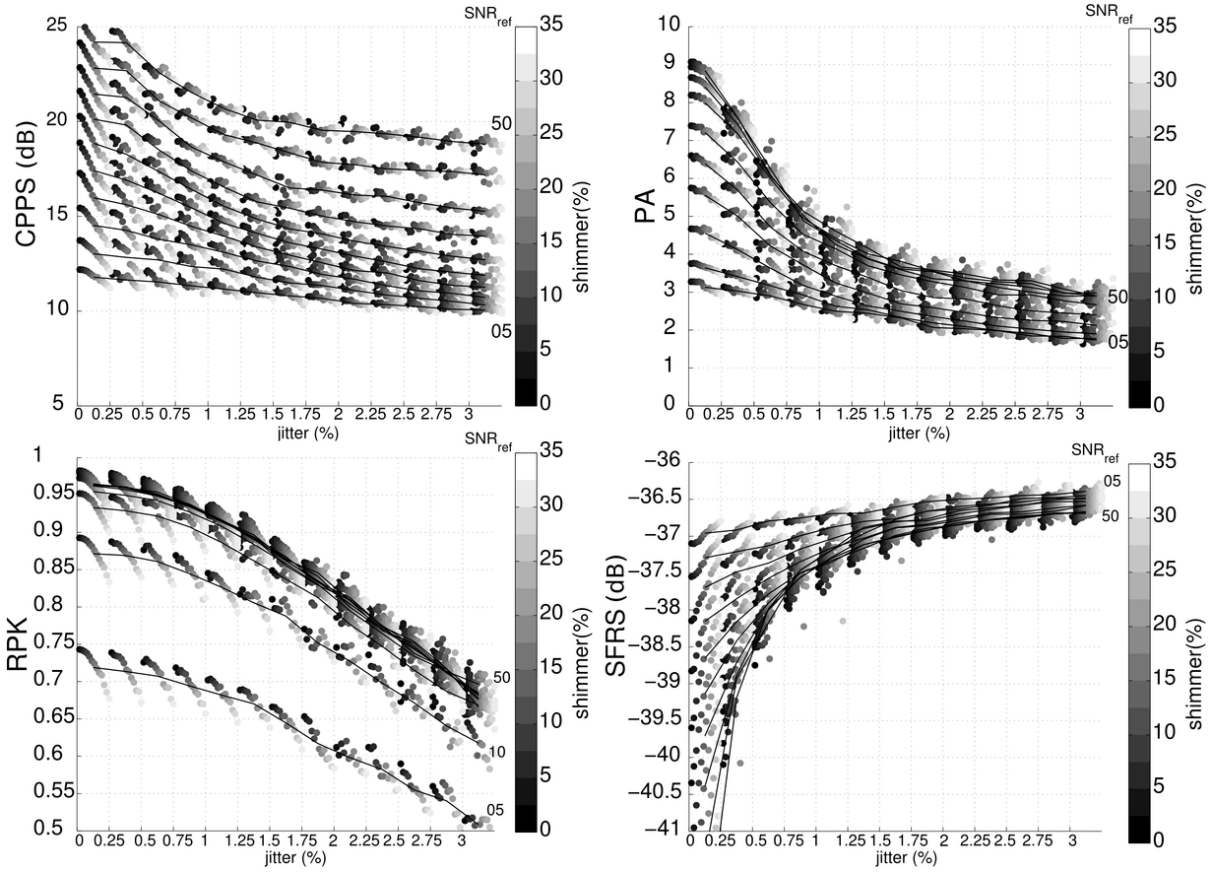


Figura 2.11 – Comparação com outros métodos. Efeitos das perturbações sintéticas nos métodos com /a/ masculino sintético. CPPS (cepstral peak prominence smoothed), PA (pitch amplitude, variando de 0 a 10), RPK (Pearson R at autocorrelation peak) e SFRS (spectral flatness of residue). Adaptado de Vieira, Sansão e Yehia (2014).

outras calculadas em Shue (2010), a saber, HNR05 (0-500 Hz), HNR15 (0-1500 Hz) e HNR25 (0-2500 Hz). Uma observação deve ser feita, as vozes humanas usadas na avaliação apresentada na Figura 2.13 são predominantemente soprosas, sem longos intervalos de aspereza. Estudos complementares são necessários usando um banco de dados com gravações afetadas simultaneamente por ruído glótico e perturbações ciclo-a-ciclo, perceptualmente avaliadas, de modo a avaliar de maneira mais completa a robustez destes métodos como preditores da soprosidade. A elaboração de tal banco de dados não é trivial.

2.4.6 Fala corrente

Finalmente, é possível usar a S^2NR em fala corrente, como exemplificado na Figura 2.14, que mostra o espectrograma (baixo) e a série temporal respectiva (topo) para uma sentença dita por uma falante do sexo feminino. A faixa dinâmica foi de aproximadamente 47 dB e as medidas cairam entre 0-10 dB durante intervalos de silêncio, paradas não-vozeadas ou fricativas não vozeadas. Um limiar estabelecido para $S^2NR(t) \approx 15$ dB parece adequado para decisão de vozeamento/não-vozeamento. A queda no valor da $S^2NR(t)$

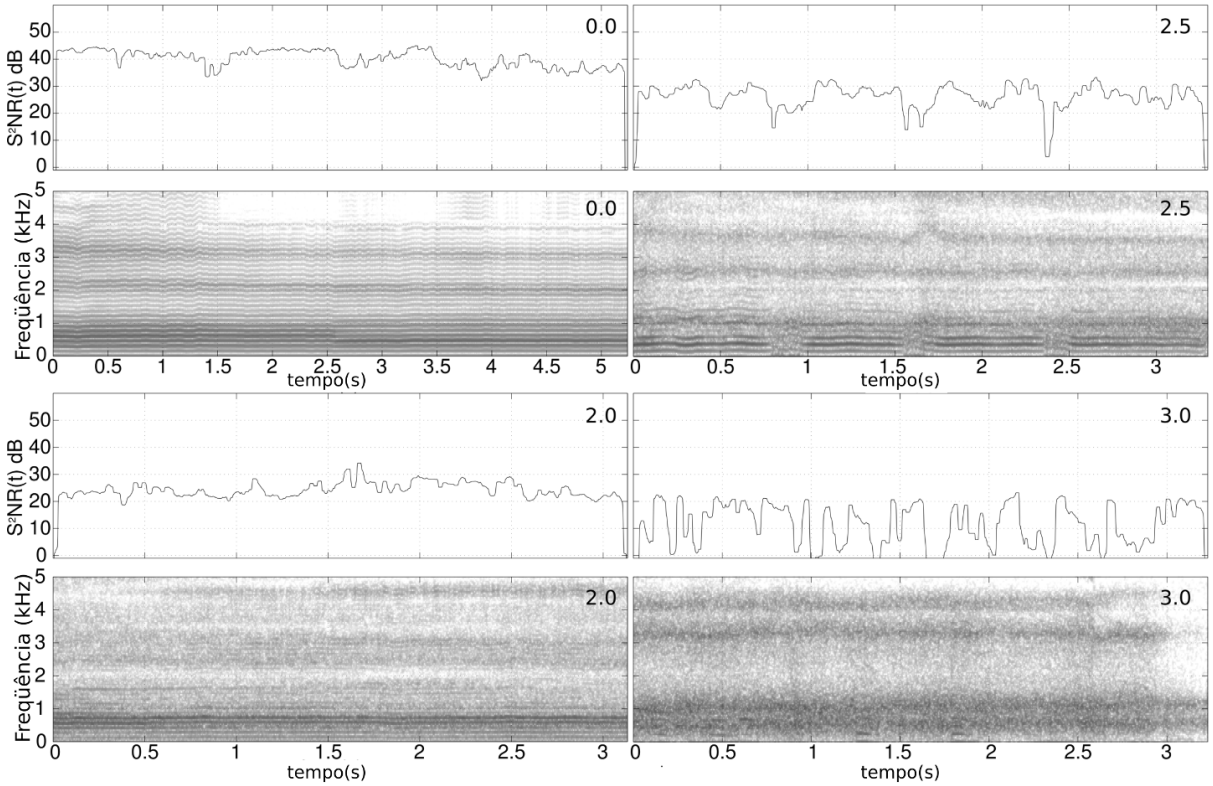


Figura 2.12 – Espectrograma e medida da S^2NR correspondente. Dados de vogal /a/ natural. A avaliação perceptiva está indicada por número no topo direito das figuras. Os eixos de freqüência estão limitados em 5 kHz para permitir visualização da estrtura harmônica. Adaptado de Vieira, Sansão e Yehia (2014).

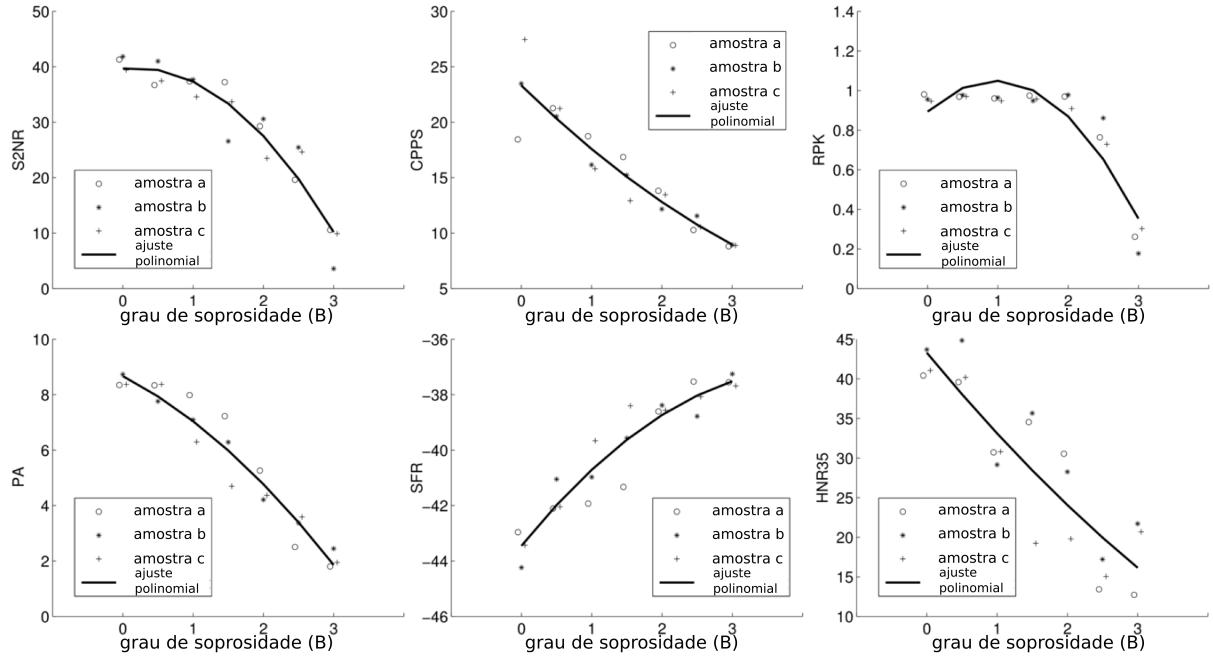


Figura 2.13 – Medidas de pertubação e medidas perceptivas de soprosidade. Cada grau perceptivo apresenta três amostras (a,b,c) obtidos de diferentes falantes. Adaptado de Vieira, Sansão e Yehia (2014).

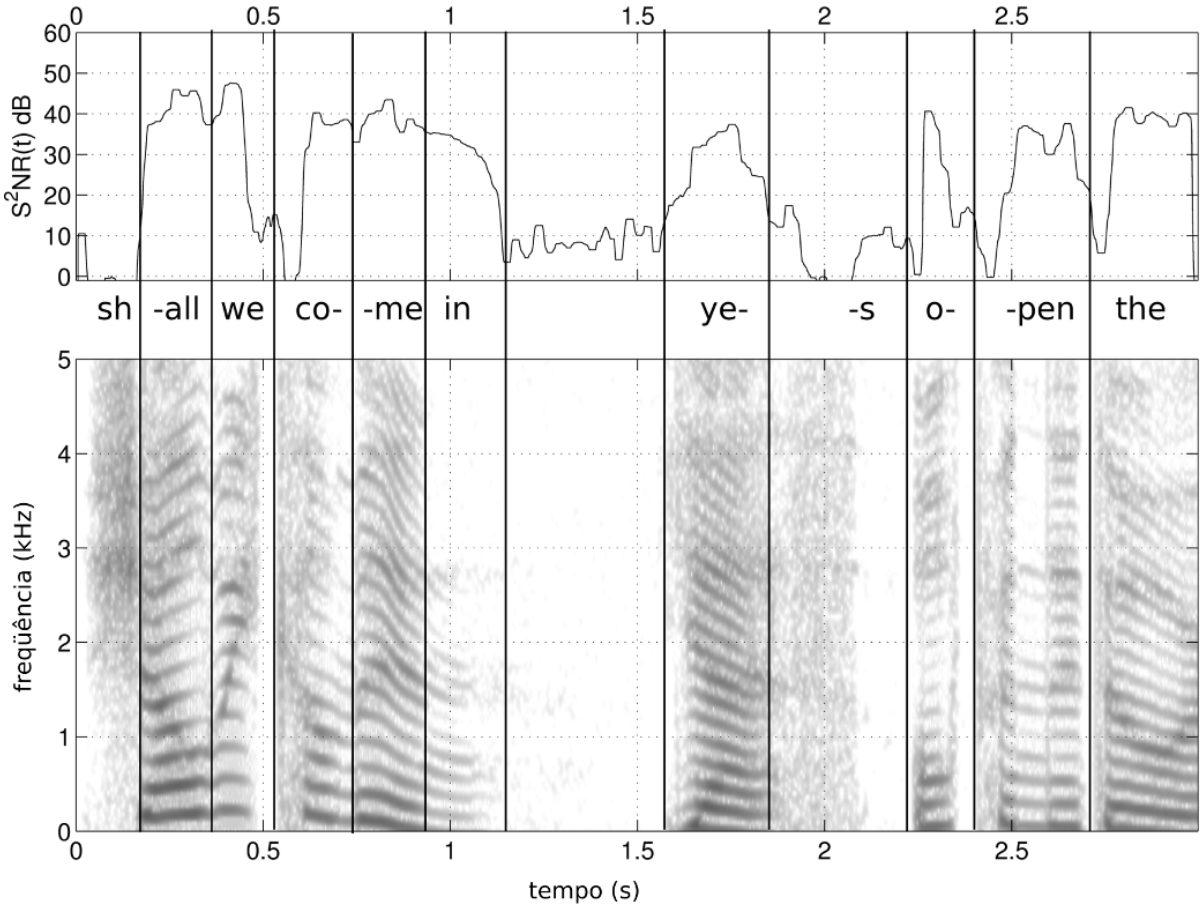


Figura 2.14 – S^2NR em fala corrente. Fragmento de "Shall we come in? Yes, open the door ..."lido por um falante nativo do inglês. Figura inferior mostra o espectrograma de banda estreita e a superior a evolução temporal dos valores S^2NR. Adaptado de Vieira, Sansão e Yehia (2014).

seguindo o decaimento da f_0 (entre 0,75 s e 1,0s, por exemplo), pode sugerir que a $S^2NR(t)$ depende da freqüência fundamental. No entanto, uma inspeção mais próxima revela que a $S^2NR(t)$ seguiu de uma redução na excitação vocal, como pode ser vista na atenuação das linhas harmônicas no intervalo mencionado. O método tem potencial aplicação em fala corrente mas possíveis influências de fatores linguísticos precisam ser investigados em detalhe.

2.4.7 Observações finais

Este capítulo descreveu a aplicação de um algoritmo para medida da relação sinal-ruído em sinais de fala. O método foi sistematicamente avaliado com vogais sintéticas corrompidas por perturbações fonatórias e níveis de ruído conhecidos a priori. Valores da S^2NR foram precisos dentro de uma faixa de $\pm$3,2 dB para maior parte das medidas dentro da faixa de 5-40 dB. O algoritmo foi robusto ao *jitter*, apresentou certa dependência ao *shimmer*, se apresentou preciso nos sinais com maior nível de ruído (faixa de 5-35 dB) e pode ser usado como preditor da soprosidade. O método não foi sensível a erros

de rastreamento da freqüência fundamental e pode potencialmente ser aplicado em fala corrente. Investigações futuras podem melhorar a sintonia do algoritmo para reduzir a influência do tipo de vogal e validar seu uso em fala corrente, mas comparado com os métodos existentes, o comportamento da S^2NR foi similar ou superior. Em particular, S^2NR não foi paradoxal no sentido que os valores medidos não dependiam fortemente dos valores crescentes de *jitter* e *shimmer*.

3 Modelagem psicofísica da soprosidade

3.1 Escalas de soprosidade

3.1.1 Escalonamento por diferenças

O escalonamento por diferenças é um método psicofísico usado para estimar diferenças perceptivas em estímulos distribuídos ao longo de uma dimensão física contínua. O objetivo do método é entender como diferenças na escala física são traduzidas como diferenças perceptivas (KNOBLAUCH; MALONEY et al., 2008; KINGDOM; PRINS, 2010).

Deseja-se atribuir valores numéricos para cada estímulo, de forma que a diferença designada reflita a diferença percebida entres os estímulos. Os valores designados formam uma escala projetada para predizer as diferenças percebidas, ou seja, uma "escala de diferenças".

Como exemplo, pode-se citar um hipotético experimento para detecção de limiares em escalas de tons de cinza. A Figura 3.1 mostra uma série de amostras com intensidade progressiva na escala de cinza. A diferença entre os níveis consecutivos é de 10%. O objetivo neste caso é relacionar a variação da intensidade com uma escala de diferença percebida, como no caso hipotético da Figura 3.2 (gerada através da simulação computacional de experimentos psicométricos com parâmetros controlados).

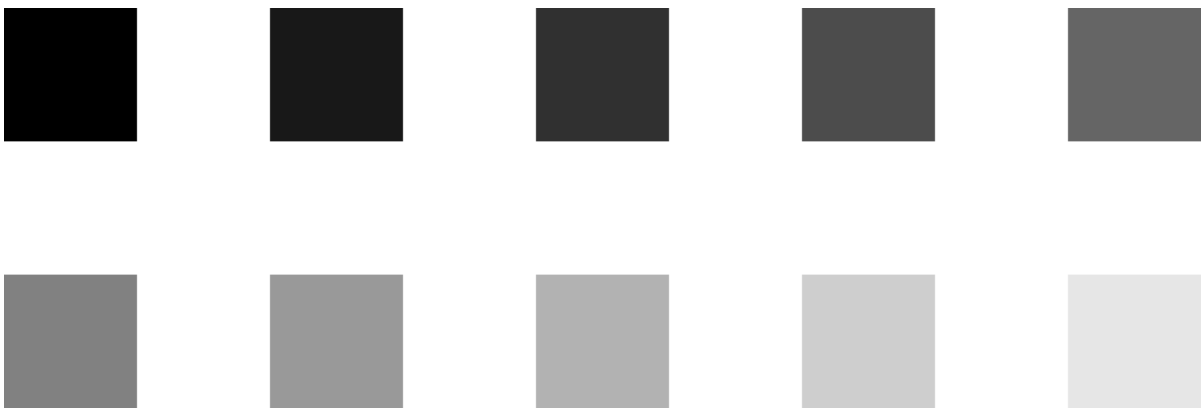


Figura 3.1 – Amostras com variação de intensidade de tons de cinza.

Para o caso hipotético, pode-se perceber que a diferença perceptiva é menor nos extremos da escala física (intensidades menores que 20% e intensidades maiores que 70%) do que na faixa intermediária (30% a 60%).

Neste trabalho, o método será utilizado para obter uma escala de diferenças para

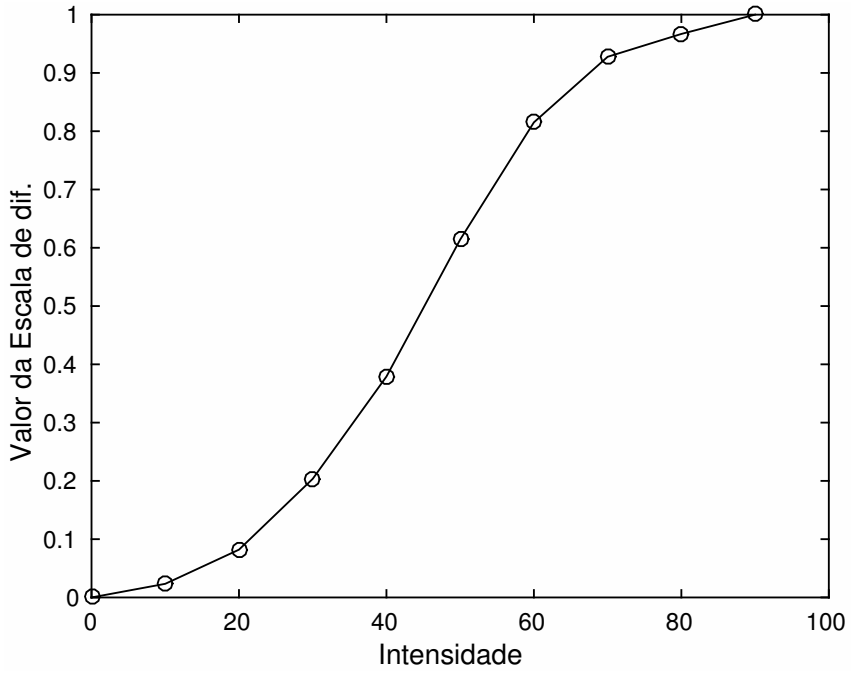


Figura 3.2 – Escala de diferenças em função da intensidade de tons de cinza (gráfico obtido através de simulação computacional).

a percepção da soprosidade vocal com a diferentes níveis de ruído glótico, obtendo uma função similar à Figura 3.2.

3.1.2 Escalonamento de diferenças por máxima verossimilhança (MLDS)

O procedimento de "Escalonamento de diferenças por máxima verossimilhança" (MLDS, *maximum likelihood difference scaling*) foi desenvolvido por Maloney e Yang (2003). O método é descrito brevemente abaixo:

Supondo um conjunto de N estímulos, $S_1, \ldots, S_N$, cada estímulo S_i está relacionado com um parâmetro físico ϕ_i e com um parâmetro de percepção ψ_i. Os estímulos devem ser ordenados de forma monótona crescente ($\phi_1 < \cdots < \phi_N$) ou decrescente ($\phi_1 > \cdots > \phi_N$).

Em sua formulação original (MALONEY; YANG, 2003), a cada avaliação, quatro estímulos são apresentados (uma quadra, na forma de dois pares). O avaliador deve então estimar qual dos pares tem o parâmetro em observação mais distinto. Sendo $(S_i, S_j; S_k, S_l)$, se (S_i,S_j) ou (S_k,S_l) são mais distintos.

Um caso particular de quadra é quando um dos estímulos é comum a ambos pares. Neste caso, são apresentados três estímulos distintos, formando uma tríade. O avaliador é apresentado aos estímulos (S_i, S_j, S_k) e deve avaliar se (S_i, S_j) ou (S_j, S_k) são mais distantes perceptivamente.

A tarefa sempre consiste na percepção da diferença entre dois estímulos, no intervalo perceptivo $l_{ij} = \psi_j - \psi_i$.

Formulação para o caso da quadra

Supondo o caso da quadra, a variável de decisão é dada por:

$$D(i,j;k,l) = l_{ij} - l_{kl} + \epsilon, \tag{3.1}$$

onde ϵ é uma variável aleatória Gaussiana de média nula e desvio padrão σ representando o erro de avaliação.

No caso de $\psi_j - \psi_i > \psi_l - \psi_k$, ter-se-á $D(i,j;k,l) > 0$.

Pode-se diminuir o número de graus de liberdade arbitrando $\psi_1 = 0$ e $\psi_N = 1$, por exemplo. Resta a determinação dos parâmetros $\psi_2, \ldots, \psi_{N-1}$ e σ. A estimação destes parâmetros é feita por máxima verossimilhança.

A probabilidade que $D(i,j;k,l) > 0$, $P[D(i,j;k,l) > 0|\psi_2, \ldots, \psi_{N-1}, \sigma]$ é dada pela função distribuição acumulada $\Phi_\sigma(l_{kl} - l_{ij})$. Desta forma:

$$\Phi_\sigma = P[\epsilon \leq x] = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{u^2}{2\sigma^2}\right) du \tag{3.2}$$

A probabilidade de escolha do outro intervalo, $D(i,j;k,l) < 0$ é $1 - \Phi_\sigma(l_{kl} - l_{ij})$.

Representando a escolha do primeiro intervalo ($D(i,j;k,l) > 0$) por $R = 1$ e do segundo por $R = 0$ ($D(i,j;k,l) < 0$), a probabilidade de resposta r de um sujeito é escrita como:

$$P[r|\psi_2, \ldots, \psi_{N-1}, \sigma] = \Phi_\sigma(\psi_j - \psi_i - \psi_l + \psi_k)^R [1 - \Phi_\sigma(\psi_j - \psi_i - \psi_l + \psi_k)]^{1-R} \tag{3.3}$$

Para uma seqüência de testes $t = 1, \ldots, T$, tem-se uma avaliação R_t para cada t, com uma diferença $\Delta_t = \psi_j - \psi_i - \psi_l + \psi_k$. Assim, a probabilidade de uma seqüência de respostas é dada por:

$$P[R_1, \ldots, R_T|\phi_2, \ldots, \phi_{N-1}, \sigma] = \prod_{t=1}^{T} \Phi_\sigma(\Delta_t)^{R_t} [1 - \Phi_\sigma(\Delta_t)]^{1-R_t} \tag{3.4}$$

A equação 3.4 é equivalente à expressão da verossimilhança $L[\psi_2, \ldots, \psi_{N-1}, \sigma|R_1, \ldots, R_T]$. Determina-se, através de uma procedimento de estimativa por máxima verossimilhança (maximum likelihood estimation - MLE), os parâmetros $\hat{\Psi}_2, \ldots, \hat{\Psi}_{N-1}$ e $\hat{\sigma}$ que maximizam a equação do log natural da verossimilhança:

$$\begin{aligned} LL[\psi_2, \ldots, \psi_{N-1}, \sigma|R_1, \ldots, R_T] &= \log L[\psi_2, \ldots, \psi_{N-1}, \sigma|R_1, \ldots, R_T] \\ &= \sum_{t=1}^{T} \log \left\{ \Phi_\sigma(\Delta_t)^{R_t} [1 - \Phi_\sigma(\Delta_t)]^{1-R_t} \right\} \end{aligned} \tag{3.5}$$

Formulação para o caso da tríade

Em muitos casos, notadamente quando os estímulos são apresentados seqüencialmente, é mais conveniente usar tríades na obtenção da escala de diferenças. Nesta situação, três estímulos são apresentados e o sujeito informa qual dos pares é mais próximo (ou mais distinto).

Os resultados são análogos ao do caso da quadras, e são obtidos ao se incluir um elemento comum aos pares.

Neste caso, a variável de decisão é dada por:

$$D(i,j,k) = l_{ij} - l_{jk} + \epsilon, \tag{3.6}$$

onde ϵ é uma variável aleatória Gaussiana de média nula e desvio padrão σ representando o erro de avaliação.

No caso de $\psi_j - \psi_i > \psi_k - \psi_j$, ter-se-á $D(i,j,k) > 0$. A probabilidade de escolha deste intervalo é dada por $\Phi_\sigma(l_{jk} - l_{ij})$. A escolha do intervalo complementar é $1 - \Phi_\sigma(l_{jk} - l_{ij})$.

Representando a escolha do primeiro intervalo por $R = 1$ e do segundo por $R = 0$, a probabilidade de resposta r de um sujeito é escrita como:

$$P[r|\psi_2,\ldots,\psi_{N-1},\sigma] = \{\Phi_\sigma[\psi_j - \psi_k - (\psi_i - \psi_j)]\}^R \{1 - \Phi_\sigma[\psi_j - \psi_k - (\psi_i - \psi_j)]\}^{1-R} \tag{3.7}$$

Para uma seqüência de testes $t = 1,\ldots,T$, tem-se uma avaliação R_t para cada t, com uma diferença $\Delta_t = 2\psi_j - \psi_i - \psi_k$. Assim, a probabilidade de uma seqüência de respostas é dada por:

$$P[R_1,\ldots,R_T|\phi_2,\ldots,\phi_{N-1},\sigma] = \prod_{t=1}^{T} \Phi_\sigma(\Delta_t)^{R_t}[1 - \Phi_\sigma(\Delta_t)]^{1-R_t} \tag{3.8}$$

A equação 3.8 é equivalente à expressão da verossimilhança $L[\psi_2,\ldots,\psi_{N-1},\sigma|R_1,\ldots,R_T]$. Determina-se através de uma procedimento de estimativa por máxima verossimilhança (maximum likelihood estimation - MLE), os parâmetros $\hat{\Psi}_2,\ldots,\hat{\Psi}_{N-1}$ e $\hat{\sigma}$ que maximizam a equação do log natural da verossimilhança, como na equação 3.5.

Escolha das tríades/quadras

Sendo p o número de estímulos, nos casos de tríades e quadras, $i = 3$ e $i = 4$, respectivamente, o número possível de tríades/quadras é dado por:

$$\binom{p}{i} = \frac{p!}{i!(p-i)!}. \tag{3.9}$$

Uma determinada tríade/quadra deve ser apresentada ao informante mais de uma vez. Desta forma, o número de avaliações necessárias pode tornar o experimento extremamente longo.

O experimento pode ser abreviado ao se impedir a apresentação de estímulos distantes em relação à grandeza física, sem prejuízo na convergência do método de maximização da verossimilhança (KNOBLAUCH; MALONEY, 2012).

Avaliação do intervalo de confiança das escalas

O erro das escalas obtidas através da MLDS pode ser estimado através do método do *bootstrap* (KNOBLAUCH; MALONEY, 2012), permitindo estabelecer intervalos de confiança para um determinado experimento.

3.1.3 Estudos pilotos e experimentos

Uma série de estudos pilotos e experimentos, com diferentes tipos de estímulos, é realizada para obter diferentes escalas de diferenças relativas à percepção auditiva do ruído em variados tipos de sinais de teste. Os experimentos serão assim designados:

- Estudo piloto 1: Tom puro com adição controlada de ruído rosa,
- Estudo piloto 2: Tom puro com adição controlada de ruído branco gaussiano,
- Estudo piloto 3: Ruído rosa (sem harmônicos),
- Estudo piloto 4: Voz sintética com relação sinal-ruído controlada (ruído adicionado à fonte glótica, 12 níveis),
- Experimento 1: Voz sintética com relação sinal-ruído controlada (ruído adicionado à fonte glótica, 7 níveis),
- Experimento 2: Voz humana com características predominantemente soprosas, classificadas perceptivamente.

Os estudos pilotos são assim designados por envolverem apenas um informante na avaliação das amostras.

Metodologia dos experimentos

Procedimento

A obtenção dos dados para estabelecimento das escalas de diferenças de percepção através da MLDS foi feita através de experimentos psicométricos de escolha forçada.

Neste trabalho, os estímulos foram feitos através da reprodução de amostras sonoras em fones de ouvido, apresentados ao avaliador de forma intervalada e seqüencial. Este fato tornou o método das tríades mais indicado para a obtenção das respostas (KNOBLAUCH; MALONEY, 2012), pois diminuiu o número de amostras executadas em uma avaliação e o número de referências que o informante deveria levar em consideração antes de tomar sua decisão.

Para cada experimento específico, p estímulos foram escolhidos. Escolheram-se então as tríades que geram a seqüência de apresentação ao usuário, incluindo repetições e eliminando tríades com comparações triviais (com estímulos perceptivamente distantes).

A seqüência de apresentação das tríades foi feita de forma aleatória. Os três estímulos (a, b, c) foram apresentados em ordem crescente de perturbação. O avaliador deveria determinar se b estava mais próximo perceptivamente de a ou de c. A reprodução da tríade de estímulos podia ser repetida a critério do avaliador antes da tomada da decisão.

Implementação

As rotinas computacionais de reprodução de amostras, telas de interface com usuário e registro de decisão através de comandos do teclado foram implementadas usando as ferramentas do *toolbox* Psychtoolbox-3 (KLEINER et al., 2007) rodando no GNU/Octave (EATON et al., 2014) no sistema operacional Debian GNU/Linux 8.

As escalas são obtidas com auxílio do pacote MLDS (KNOBLAUCH; MALONEY et al., 2008) para o software estatístico livre R (R Core Team, 2014).

Fones de ouvido

Nestes estudos pilotos/experimentos, foram utilizados fones de ouvido intra-auriculares sem marca definida (genéricos) adquiridos em um lote de 10 unidades. Deste lote, foram selecionados cinco fones com os seguintes critérios: menor assimetria entre os lados direito e esquerdo, com resposta em freqüência com menor variação na faixa de 500 a 4000 Hz.

Os dados foram obtidos através de calibração submetendo os fones a tons puros gerados no computador e reproduzidos pela placa de som do mesmo. A emissão sonora dos fones foi medida através do medidor de pressão sonora da marca Tenmars, modelo TM-103, acoplado aos fones através de um adaptador feito de isopor. Os valores medidos estão na Tabela 3.1.

Para determinar a resposta média relativa dos fones, considerou-se o valor médio da resposta em 1 kHz como referência.

Tabela 3.1 – Nível de pressão sonora dos fones utilizados (dBA) com sinais de teste. Valor de referência para medida relativa: valor médio da resposta em 1 kHz.

Fone	500 Hz	1 kHz	2 kHz	4 kHz
1 (Esquerdo)	105,5	107,2	109,2	106,0
1 (Direito)	106,1	108,1	110,1	107,1
2 (Esquerdo)	108,0	108,3	110,3	107,0
2 (Direito)	106,3	105,8	110,5	102,5
3 (Esquerdo)	103,8	105,1	108,5	105,2
3 (Direito)	106,5	107,9	111,8	103,9
4 (Esquerdo)	108,1	109,4	112,1	105,6
4 (Direito)	107,9	109,0	113,0	105,6
5 (Esquerdo)	108,5	108,9	113,3	107,4
5 (Direito)	106,8	107,4	111,3	107,5
Média	106,8	107,7	111,0	105,8
Relativo	-1,0	0,0	3,3	-1,9
Desvio	1,1	1,1	1,3	1,2

Para fins de comparação, a mesma calibração foi realizada com três fones marcas conhecidas: dois intra-auriculares, Samsung (modelo HS330) e Motorola (modelo SJYN1181B), e um supra-auricular marca Philips (modelo SHP2000). Os valores medidos estão na Tabela 3.2. Nota-se que na faixa de freqüência medida, os fones genéricos do lote em questão têm em média menor variação em sua resposta.

Tabela 3.2 – Nível de pressão sonora dos fones de referência (dBA) com sinais de teste, valor médio dos lados esquerdo e direito. Valor de referência para medida relativa: valor médio da resposta em 1 kHz.

Fone	500 Hz	1 kHz	2 kHz	4 kHz
Samsung (intra)	106,3	107,3	117,7	107,0
Relativo	-1,0	0,0	10,4	-0,3
Motorola (intra)	102,4	107,2	115,0	112,8
Relativo	-4,8	0,0	7,8	5,6
Philips (supra)	87,7	92,1	93,4	86,0
Relativo	-4,4	0,0	1,3	-6,1

Estudo piloto 1: tom puro com ruído rosa aditivo

No primeiro estudo piloto, buscou-se obter a escala para um tom puro (220 Hz) quando existe a adição de ruído rosa.

Outro efeito investigado é o passo na relação sinal-ruído entre os estímulos, repetindo o experimento para estímulos separados por 10 dB, 5 dB e 3 dB.

Passo 10 dB

Para o passo de 10 dB, seis estímulos foram gerados (variando de 0 a 50 dB). Foi estabelecida um distância máxima de 3 níveis para comparação e 9 repetições por tríade. Cada estímulo tem duração de 1,5 s.

Um informante (sexo masculino, 35 anos, sem perdas auditivas) participou deste estudo com 180 comparações de tríades necessárias. A escala de diferenças estimada pela MLDS e o intervalo de confiança de 95% são mostrados na Figura 3.3a.

Passo 5 dB

Para o passo de 5 dB, onze estímulos foram gerados (variando de 0 a 50 dB). Foi estabelecida um distância máxima de 3 níveis para comparação e 2 repetições por tríade. Cada estímulo tem duração de 1,5 s.

O mesmo informante participou no estudo com 242 comparações de tríades necessárias.

A escala de diferenças (e respectivo intervalo de confiança de 95%) estimada pela MLDS está mostrada na Figura 3.3b.

Passo 3 dB

Para o passo de 3 dB, onze estímulos foram gerados (variando de 7 a 40 dB). Foi estabelecida um distância máxima de 3 níveis para comparação e 2 repetições por tríade. Cada estímulo tem duração de 1,5 s.

O informante efetuou 242 comparações de tríades. A escala de diferenças estimada está na Figura 3.3c.

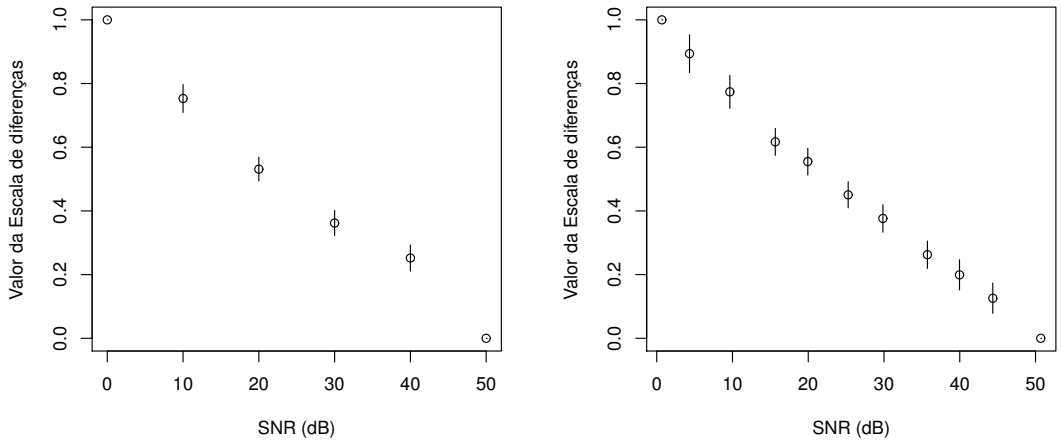
Estudo piloto 2: tom puro com ruído branco gaussiano aditivo

Nesta etapa, buscou-se obter a escala para um tom puro (220 Hz) quando existe a adição de ruído branco gaussiano (AWGN).

São gerados 12 estímulos com relação sinal-ruído controlada, variando de 7 dB até 40 dB, com passo de 3 dB. Cada estímulo tem duração de 1,5 s.

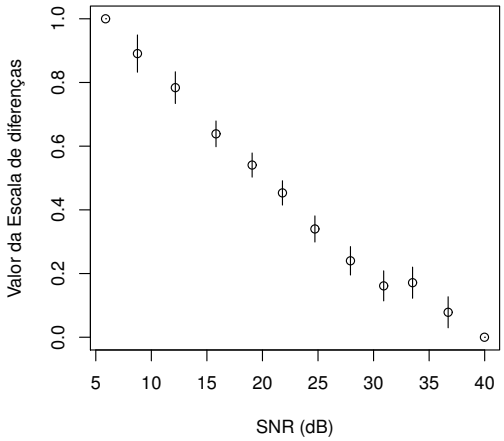
A freqüência do tom puro foi escolhida por ser próxima da freqüência fundamental da voz feminina.

Um número excessivo de tríades, limitou-se a distância máxima da SNR entre estímulos para 9 dB (3 níveis). Cada tríade teve 3 repetições. O informante do estudo piloto anterior participou deste experimento, totalizando 456 avaliações de tríades. A



(a) Passo de 10 dB

(b) Passo de 5 dB



(c) Passo de 3 dB

Figura 3.3 – Tom puro (220 Hz) com ruído rosa aditivo. Escala de diferenças em função da SNR, escala padrão.

escala de diferenças estimada pela MLDS e os respectivos intervalos de confiança estão na Figura 3.4.

Estudo piloto 3: ruído rosa

Neste estudo piloto, desejou-se investigar a percepção do ruído rosa puro, sem a presença de um sinal harmônico como referência.

Desta forma, não foi possível usar a relação sinal-ruído, isto é, harmônico-ruído como atributo físico. No seu lugar, usou-se a tensão de saída logarítmica, medida em dBV[1].

Neste experimento, variou-se o nível do ruído de -53 dBV até aproximadamente

[1] Considerando a saturação da placa de som em 0 dBV, equivalente à uma saída de $1V_{rms}$.

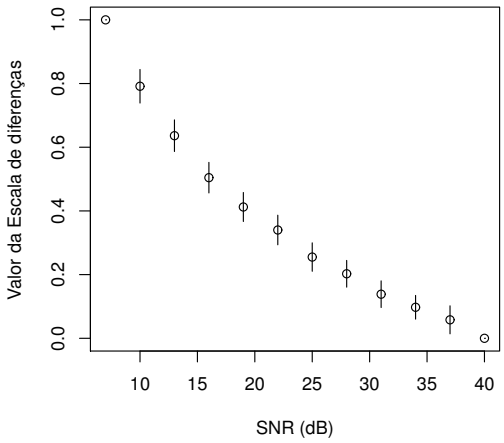


Figura 3.4 – Tom puro (220 Hz) com ruído branco aditivo. Escala de diferenças em função da SNR, escala padrão.

-12 dBV, com passos de 8 dB, obtendo 6 níveis de estímulo. Estabeleceu-se 9 repetições por tríade e distância máxima de 3 níveis para comparação.

O informante dos casos anteriores participou deste estudo com 180 comparações de tríades necessárias. A escala de diferenças estimada pela MLDS e os respectivos intervalos de confiança estão na Figura 3.5.

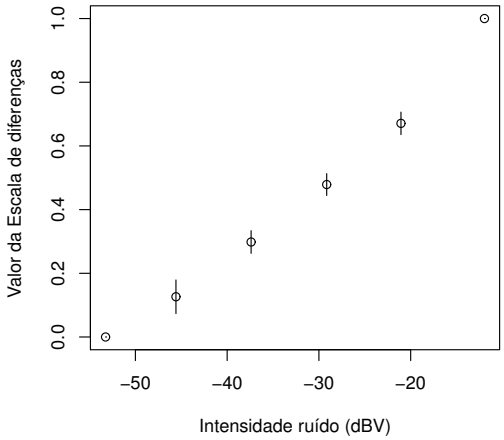


Figura 3.5 – Ruído rosa puro. Escala de diferenças em função da potência do ruído (em dBV), escala padrão.

Estudo piloto 4: voz soprosa sintética (12 níveis)

Passou-se a considerar neste estudo piloto sinais mais complexos que o tom puro com ruído aditivo. Desejou-se obter uma escala de diferenças para sinais sintético com característica de voz soprosa.

Utilizando o modelo fonte-filtro, adicionou-se ruído controlado na fonte do sinal e filtrou-se com a vogal desejada. Os estímulos usados neste experimento tinham freqüência fundamental de 220 Hz, *jitter* de 0,2 % e vogal /a/. Cada elocução tinha 3 segundos de duração.

As amostras de vogal sustentada utilizadas nestes estudos e experimento têm a percepção de naturalidade discutida no trabalho desenvolvido por Santos et al. (2018).

Foram gerados 12 estímulos, com relação sinal-ruído variando de 7 dB até 40 dB. Foram efetuadas 3 repetições por tríade, distância máxima de 3 níveis para comparação.

Um informante participou do estudo, com 456 comparações de tríades efetuadas. O resultado obtido pela MLDS está exibido na Figura 3.6.

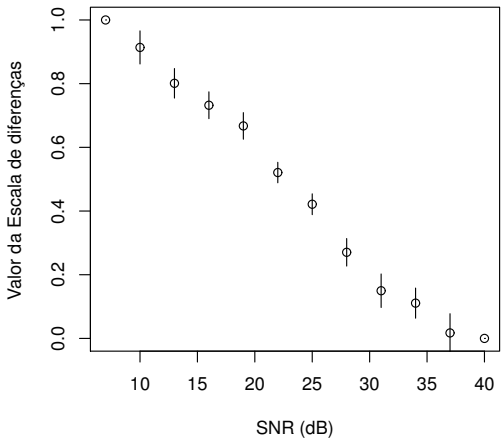


Figura 3.6 – Voz sintética ($f_0 = 220Hz$) com ruído glótico. Estímulos separados por 3 dB. Escala de diferenças em função da SNR, escala padrão.

Experimento 1: voz soprosa sintética (7 níveis)

Com base no resultado do "Estudo piloto 4", desenvolveu-se o "Experimento 1" com o objetivo de aumentar número de participantes.

Considerando o tempo de duração do estudo anterior (aproximamente 2 horas, com as repetições) e a escala de diferenças obtida, diminuiu-se o número de estímulos, escolhendo as amostras com relação sinal-ruído de referência em 7, 10, 16, 22, 28, 34 e 40 dB.

Sete voluntários participaram do experimento (faixa etária de 22-55 anos, 5 homens, 2 mulheres, sem perda auditiva conhecida).

Estabeleceu-se uma distância máxima de 3 níveis para as comparações. Cada voluntário efetuou 66 avaliações (2 repetições por tríade).

O resultado considerando todos os voluntários obtido pela MLDS está exibido na Figura 3.7. Para fins de comparação, exibe-se também o resultado do voluntário que participou do "Estudo piloto 4" na Figura 3.8.

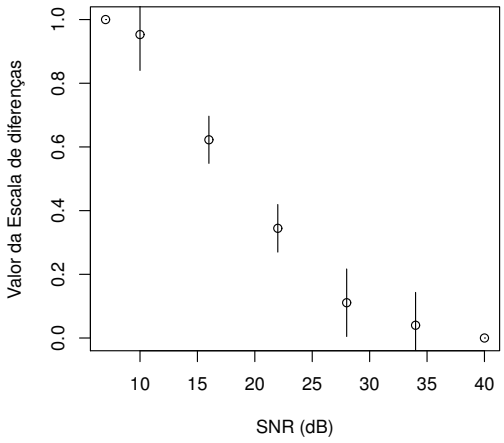


Figura 3.7 – Voz sintética ($f_0 = 220Hz$) com ruído glótico. Estímulos separados por 6 dB. Escala de diferenças em função da SNR, escala padrão. $N = 7$.

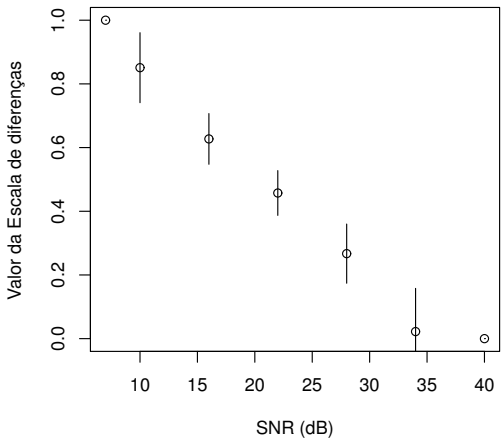


Figura 3.8 – Voz sintética ($f_0 = 220Hz$) com ruído glótico. Estímulos separados por 6 dB. Escala de diferenças em função da SNR, escala padrão. Informante comum ao estudo piloto 4.

Experimento 2: voz soprosa humana

Em extrapolação do experimento anterior, os estímulos sintéticos foram substituídos por amostras de voz humana com variada característica de soprosidade.

As amostras utilizadas foram gravações da vogal /a/ de diferentes sujeitos. Neste caso, não se tinha uma variável de controle nos estímulos. No entanto, o conjunto de sete

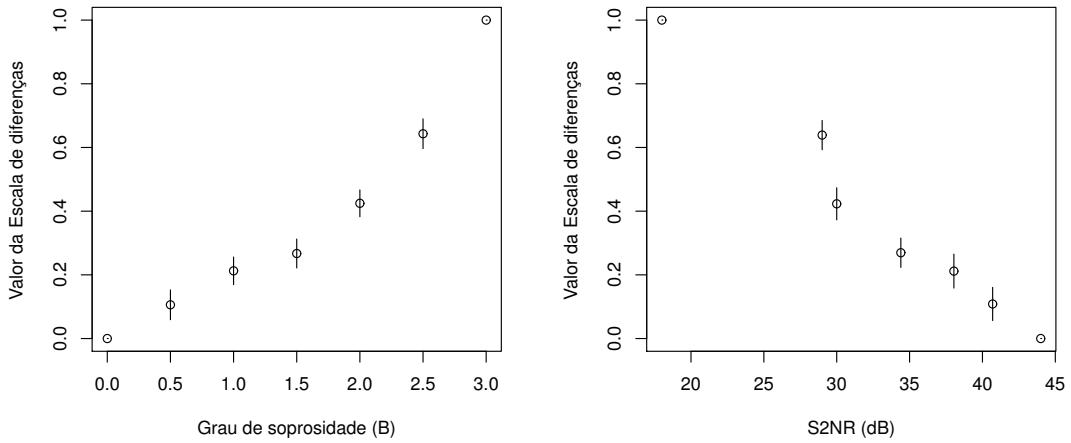
estímulos foi previamente perceptivamente classificado em soprosidade, 0 (ausente) a 3 (extremo) (VIEIRA; SANSÃO; YEHIA, 2014).

Sete voluntários participaram do experimento (faixa etária de 22-55 anos, 5 homens, 2 mulheres, sem perda auditiva conhecida), os mesmos do experimento anterior.

Estabeleceu-se uma distância máxima de 3 níveis para as comparações. Cada voluntário efetuou 66 avaliações (2 repetições por tríade).

O resultado obtido pela MLDS está exibido na Figura 3.9a.

Neste caso não se conhece *a priori* a relação sinal-ruído das amostras. Para obter-se um eixo correlacionado ao parâmetro físico, calculou-se a S^2NR em um trecho representativo do estímulo, gerando a Figura 3.9b.



(a) Escala de diferenças em função da classificação perceptiva, escala padrão.

(b) Escala de diferenças em função da S^2NR, escala padrão.

Figura 3.9 – Voz humana com variado grau de soprosidade.

3.1.4 Discussão dos resultados

Estudo piloto 1: tom puro com ruído rosa aditivo

No primeiro experimento, ao se determinar um passo de 10 dB na relação sinal-ruído dos estímulos, pretendeu-se verificar a viabilidade da utilização do método da MLDS com este tipo de sinal. Tal passo seria supostamente maior que a mínima diferença perceptível do ruído neste tipo de estímulo.

Observando os resultados exibidos no gráfico da Figura 3.3a e na Tabela 3.3, nota-se que os 6 estímulos testados apresentam valores na escala de diferenças bem definidos e sem sobreposição considerando os intervalos de confiança calculados através do *bootstrap*.

Os estímulos extremos têm seus valores na escala arbitrados, portanto não apresentam variação. Os estímulos intermediários apresentaram intervalos de confiança de 95% na faixa de ±0,04 do valor determinado.

Nestas condições, considerando a faixa de variação da relação sinal-ruído de 0 a 50 dB, é possível afirmar que é possível distinguir seis ou mais classes diferentes na percepção do ruído adicionado a um tom puro.

Tabela 3.3 – Análise de bootstrap para o caso do tom puro com ruído rosa (passo 10 dB)

Estímulo	SNR (dB)	Média	Desvio	IC 95%
1	50	0,00	-	-
2	40	0,25	0,02	0,04
3	30	0,36	0,02	0,04
4	20	0,54	0,02	0,04
5	10	0,76	0,02	0,04
6	0	1,00	-	-

Procedeu-se diminuindo o passo entre os estímulos para 5 dB, resultado exibido na Figura 3.3b e na Tabela 3.4. A faixa dos estímulos é a mesma do experimento com passo de 10 dB, com o valor dos extremos arbitrados pelo método em 0 e 1.

Analisando os valores obtidos, observa-se que a curva apresenta monotonicidade, e as separações entre os estímulos ficam entre 0,06 e 0,16 na escala de diferenças. Levando em conta os respectivos intervalos de confiança de 95%, alguns níveis de estímulo apresentam superposição dentro da faixa, como ocorre por exemplo em 35 dB ($0{,}26 \pm 0{,}04$) e 40 dB ($0{,}20 \pm 0{,}04$).

Tabela 3.4 – Análise de bootstrap para o caso do tom puro com ruído rosa (passo 5 dB)

Estímulo	SNR (dB)	Média	Desvio	IC 95%
1	50	0,00	-	-
2	45	0,13	0,03	0,05
3	40	0,20	0,02	0,05
4	35	0,26	0,02	0,04
5	30	0,38	0,02	0,04
6	25	0,45	0,02	0,04
7	20	0,56	0,02	0,04
8	15	0,62	0,02	0,04
9	10	0,78	0,02	0,04
10	5	0,90	0,03	0,05
11	0	1,00	-	-

Considerando a faixa de sobreposição pequena entre os estímulos, diminuiu-se o passo dos estímulos para 3 dB. Para não tornar o experimento mais longo, o número de níveis foram mantidos, mas modificou-se os extremos inferior e superior para 7 dB e 40 dB, respectivamente.

Na estimativa da escala, os estímulos de 34 dB e 31 dB ficaram sobrepostos. No entanto, na faixa de relação sinal-ruído inferior aos 30 dB não ocorreu sobreposição, nem mesmo na faixa de confiança.

Este resultado sugere que o valor da relação sinal-ruído tem impacto na mínima diferença perceptível, isto é, com relação sinal-ruído mais baixo seria possível distinguir menores diferenças.

Tabela 3.5 – Análise de bootstrap para o caso do tom puro com ruído rosa (passo 3 dB)

Estímulo	SNR (dB)	Média	Desvio	IC 95%
1	40	0,00	-	-
2	37	0,08	0,02	0,04
3	34	0,17	0,02	0,04
4	31	0,16	0,02	0,04
5	28	0,24	0,02	0,04
6	25	0,34	0,02	0,04
7	22	0,45	0,02	0,04
8	19	0,54	0,02	0,04
9	16	0,64	0,02	0,04
10	13	0,79	0,02	0,04
11	10	0,89	0,03	0,06
12	7	1,00	-	-

Estudo piloto 2: tom puro com ruído branco gaussiano aditivo

Para este experimento, mudou-se a característica espectral do ruído, desejando-se comparar a escala obtida em relação ao ruído rosa.

Escolheu-se o passo de 3 dB e a faixa de 7 a 40 dB de relação sinal-ruído.

Considerando os resultados exibidos no gráfico da Figura 3.4 e na Tabela 3.6, obteve-se uma relação monotônica para a escala de diferenças em função da SNR. Todavia ao se analisar os valores em conjunto com seus intervalos de confiança, verifica-se superposição dos estímulos na faixa de 25-40 dB, fato que não ocorre na faixa de 7-25 dB.

Este resultado é consistente com o visto no ruído rosa, sugerindo diferenças mínimas perceptíveis menores em SNR mais baixa.

Estudo piloto 3: ruído rosa

O intuito deste experimento foi obter uma escala de diferenças na percepção de ruído puro, sem referências de sinais harmônicos simultâneos.

A diferença de nível dos estímulos foi de aproximadamente 8 dB. Os resultados exibidos na Figura 3.5 e na Tabela 3.7 mostram que é possível obter uma escala de diferenças para o ruído rosa sem a presença de sinais harmônicos para comparação.

Tabela 3.6 – Análise de bootstrap para o caso do tom puro com ruído branco aditivo

Estímulo	SNR (dB)	Média	Desvio	IC 95%
1	40	0,00	-	-
2	37	0,06	0,02	0,04
3	34	0,10	0,02	0,04
4	31	0,14	0,02	0,04
5	28	0,21	0,02	0,04
6	25	0,26	0,02	0,04
7	22	0,34	0,02	0,04
8	19	0,42	0,02	0,04
9	16	0,51	0,02	0,04
10	13	0,64	0,02	0,04
11	10	0,79	0,03	0,06
12	7	1,00	-	-

Na faixa testada não ocorreu superposição entre os estímulos, mesmo considerando os intervalos de confiança.

Tabela 3.7 – Análise de bootstrap para o caso do ruído rosa

Estímulo	Ruído (dBV)	média	desvio	IC 95%
1	-53,3	0,00	-	-
2	-45,6	0,13	0,02	0,04
3	-37,4	0,30	0,02	0,04
4	-29,2	0,48	0,02	0,04
5	-21,1	0,67	0,02	0,04
6	-12,0	1,00	-	-

Estudo piloto 4: voz soprosa sintética com 12 níveis

Considerando que a voz humana tem características harmônicas mais ricas que um tom puro, realizou-se um experimento incluindo voz sintética simulando uma vogal /a/ feminina com ruído glótico sobreposto na fonte, antes do filtro do trato vocal.

Considerando o passo de 3 dB na SNR dos estímulos, foram obtidos o gráfico da Figura 3.6 e a Tabela 3.8. A resposta é monotônica e assim como nos casos de tom puro com ruído adicionado, ocorreu superposição dentro do intervalo de confiança com SNR superiores a 30 dB.

Mesmo com o ruído filtrado pelo trato vocal foi possível obter a escala de diferenças com passo de 3 dB, notadamente na faixa de relação sinal-ruído inferior a 30 dB, faixa correlacionada à percepção acústica da soprosidade.

Tabela 3.8 – Análise de bootstrap para o caso de voz sintética

Estímulo	SNR (dB)	Média	Desvio	IC 95%
1	40	0,00	-	-
2	37	0,02	0,03	0,06
3	34	0,12	0,03	0,06
4	31	0,15	0,03	0,06
5	28	0,28	0,02	0,04
6	25	0,43	0,02	0,04
7	22	0,52	0,02	0,04
8	19	0,67	0,02	0,04
9	16	0,73	0,02	0,04
10	13	0,80	0,02	0,05
11	10	0,92	0,03	0,06
12	7	1,00	-	-

3.1.5 Experimento 1: voz soprosa sintética, 7 níveis

Neste experimento, o objetivo foi obter uma escala de diferenças utilizando respostas fornecidas por diferentes avaliadores ($N = 7$). A curva obtida tem comportamento monotônico, que pode ser visto na Figura 3.7.

As amostras com SNR baixo (7-10 dB) e SNR (34-40dB) ficaram próximas, dentro do intervalo de confiança obtido via *bootstrap*. Na faixa intermediária, as amostras apresentaram maior distância perceptiva, sugerindo a possibilidade de se alocar mais amostras nesta faixa de SNR.

Ao se considerar o corte de apenas um avaliador, na Figura 3.8, também notou-se o comportamento monotônico e é possível perceber superposição de intervalos de confiança apenas na região de relação sinal-ruído mais alta. O resultado é coerente com o obtido na Figura 3.6, quando foi feito o estudo com 12 níveis.

Tabela 3.9 – Análise de bootstrap para o caso de voz sintética, 7 níveis, $N = 7$

Estímulo	SNR (dB)	Média	Desvio	IC 95%
1	40	0,00	-	-
2	34	0,04	0,05	0,10
3	28	0,12	0,05	0,10
4	22	0,35	0,04	0,08
5	16	0,62	0,04	0,08
6	10	0,94	0,05	0,10
7	7	1,00	-	-

Experimento 2 : voz soprosa humana

Neste experimento, tentou-se obter a escala de diferenças a partir de um conjunto de estímulos previamente classificados perceptivamente em relação à predominância da

soprosidade, em faixas que variam de amostras sem perturbação até níveis extremos.

Ressalva-se que não se tem controle do parâmetro físico em questão e as amostras são gravações de diversos sujeitos (diferentes tratos, freqüências fundamentais), condições para a adequada utilização da MLDS.

Em adição à classificação perceptiva prévia, foram feitas medidas acústicas dos estímulos para efeito de comparação (S^2NR).

Feitas as considerações, foram obtidos os gráficos exibidos na Figura 3.9 e a Tabela 3.10

Observando o gráfico da Figura 3.9a, verifica-se que o gráfico é monotônico. As amostras grau 1 e 1,5 apresentam superposição dentro dos intervalos de confiança. Todavia as amostras de grau 0, 0,5 e 1,0 não apresentam superposição entre elas, assim como a faixa de 1,5 a 3, sugerindo a possibilidade de alocação de mais níveis de classificação nesta faixa (onde ocorre maior perturbação).

Nota-se também que a maior diferença nos níveis de ruído dos estímulos (estimados pela S^2NR) está entre as amostras de grau 2,5 e 3,0 (cerca de 11,0 dB), apesar da distância perceptiva entre as referidas amostras ter sido uma das menores entre estímulos adjacentes (0,07) neste experimento.

Tabela 3.10 – Análise de bootstrap para o caso de voz predominantemente soprosa humana

Estímulo	B	S2NR (dB)	média	desvio	IC 95%
1	0,0	44,0	0,00	-	-
2	0,5	40,7	0,11	0,03	0,05
3	1,0	38,1	0,21	0,02	0,05
4	1,5	34,4	0,27	0,03	0,06
5	2,0	30,0	0,43	0,02	0,05
6	2,5	29,0	0,63	0,02	0,05
7	3,0	18,0	1,00	-	-

3.1.6 Observações sobre as escalas obtidas

Nos experimentos descritos, foram obtidas escalas de diferenças para diferentes tipos de sinal, abrangendo tom puro com ruído aditivo (rosa e branco) e sinais complexos como a voz sintética com características de soprosidade. Nestes casos foi possível obter uma escala relacionada à uma grandeza física (a relação sinal-ruído) controlada.

Nos experimentos envolvendo tom puro e ruído branco/rosa (passo de 3dB), os resultados obtidos indicaram que em determinadas condições (SNR inferior a 30 dB) é possível a discriminação de estímulos com diferença de 3 dB. Em SNR maiores tal discriminação é ainda possível, mas com grau de confiança menor. Ao se considerar uma

diferença de 6 dB (2 passos), não ocorre sobreposição na escala. As mesmas considerações podem ser estendidas ao experimento com voz sintética.

3.2 Parametrização da função psicométrica da soprosidade

Na seção anterior, foram obtidas escalas de diferenças para a percepção do ruído utilizando o método da MLDS. Seu funcionamento pode ser explicado pela integração das mínimas diferenças perceptíveis (*jnd*) com a variação do parâmetro físico estudado (KINGDOM; PRINS, 2010).

No entanto, apesar de fornecer uma escala e o intervalo de confiança para o valor do estímulo, o método não fornece explicitamente o valor da *jnd*. Para tanto, é necessária a estimativa da função psicométrica.

Ao contrário da MLDS, que obtém uma estimativa global da escala, função psicométrica é específica para um determinado nível da grandeza física percebida, *i.e.* determina-se o limiar de percepção de ruído quando a SNR é 20 dB, 30 dB, em procedimentos separados.

3.2.1 Descrição de uma função psicométrica

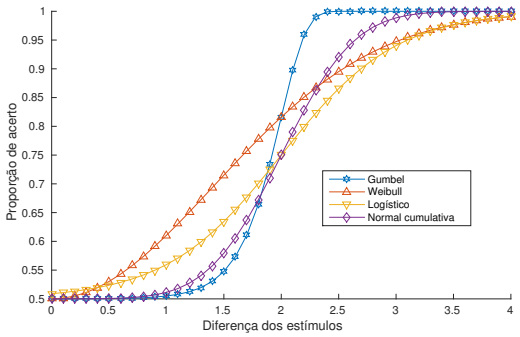
Uma função psicométrica relaciona a variação de uma determinada grandeza física com a capacidade do sujeito perceber corretamente esta variação (KINGDOM; PRINS, 2010).

A estimativa é feita através de seguidos testes de alternativa forçada (resposta sim ou não, maior ou menor). Com as respostas, determina-se o índice de acerto para cada valor da escala física. Diversas funções podem ser usadas para aproximar a função psicométrica: Função normal cumulativa, logística, Weibull e Gumbel são exemplos, curvas de forma sigmoidal.

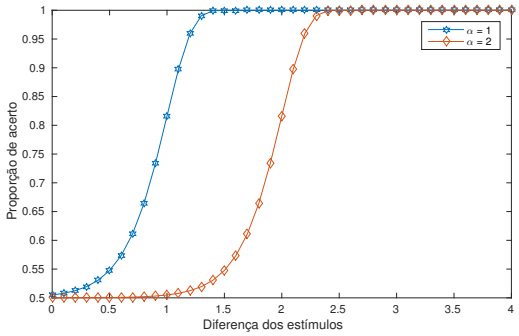
Os parâmetros de ajuste das curvas psicométricas são: limiar (α), inclinação (β), fator de adivinhação (*guessing rate*, γ) e fator de lapso (*lapse rate*, λ). A Figura 3.10a mostra as funções citadas com parâmetros $\alpha = 2$, $\beta = 2$ e $\gamma = 0.5$.

A Figura 3.10b mostra a influência do valor de α na função Gumbel (variação do limiar). Já a Figura 3.10c exemplifica a variação de β (inclinação) no mesmo tipo de função.

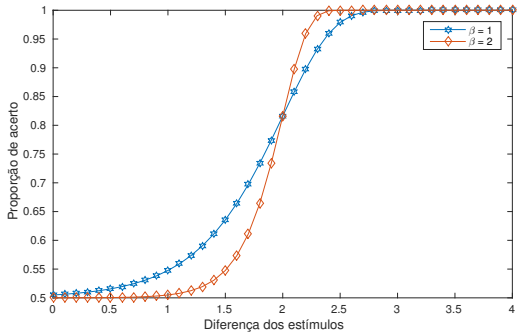
A escolha da função de ajuste pode ser feita de duas formas: baseadas em teorias *a priori* do tipo de função ou *a posteriori* buscando o melhor ajuste entre as funções (KINGDOM; PRINS, 2010).



(a) Exemplos de tipos de funções psicofísicas



(b) Variação de α na Função Gumbel



(c) Variação de β na Função Gumbel

Figura 3.10 – Funções psicofísicas e variações paramétricas na Gumbel

3.2.2 Procedimentos adaptivos

A realização de determinados experimentos psicofísicos pode se tornar um procedimento demorado e cansativo, dependendo do número de tarefas a serem realizadas pelo observador.

Um número excessivo de tarefas triviais, por exemplo, pode cansar o observador e deteriorar seu desempenho nas tarefas que exigem maior atenção.

A introdução de procedimentos adaptativos permitem que o experimento seja adequado em tempo de execução para que sejam executados testes que forneçam as informações mais relevantes para o ajuste das funções psicométricas.

Essa classe de procedimentos força que o maior número de observações sejam executadas em regiões próximas aos limiares perceptivos.

O primeiro método proposto denominado *UP/DOWN* (DIXON; MOOD, 1948) apresenta os estímulos e verifica se resposta do avaliador está correta. No caso de resposta correta, a distância dos estímulos diminui, aumentando no caso contrário.

Uma variação do método é o *UP/DOWN* transformado (WETHERILL; LEVITT, 1965). A alteração é incluir um efeito de memória na variação dos estímulos, incluindo regras como "dimimuir a distância a cada dois/três acertos do avaliador e aumentar a cada erro". Os critérios de parada podem ser definidos como um número total de ensaios

ou número máximo de reversões.

Este método ainda pode ser incrementado incluindo diferentes pesos/passos ao aumentar/diminuir as distâncias (KAERNBACH, 1991). Todavia, esta metodologia oferece apenas o limiar em um determinado ponto, não ajustando uma função psicofísica.

Outros métodos determinam o limiar ajustando a função psicofísica em tempo de execução dos ensaios. Estes são o " melhor PEST" (PENTLAND, 1980) e "QUEST" (WATSON; PELLI, 1983). Um ajuste do modelo é calculado após cada ensaio, servindo para selecionar a intensidade de estímulo do próximo passo, repetindo o procedimento até o critério de parada (KINGDOM; PRINS, 2010).

No caso do "PEST", deve-se fixar valores para β, γ e λ, além do tipo da função psicofísica. No "QUEST", os parâmetros são usados para inicialização, mas são modificados em tempo de execução para o ajuste da função psicofísica.

Neste trabalho, o método "QUEST" será usado na implementação fornecida pelo *toolbox* Palamedes (KINGDOM; PRINS, 2010).

3.2.3 Experimentos

Os experimentos foram conduzidos com dois tipos de estímulos:

- Experimento 1: tom puro com ruído rosa aditivo,
- Experimento 2: voz sintética com ruído adicionado à fonte glótica.

O objetivo foi investigar os limiares de percepção da variação da relação sinal-ruído nos referidos estímulos. O parâmetro controlado é a relação sinal-ruído em cada amostra de teste.

Metodologia dos experimentos

Procedimento

Os dados para cálculo dos limiares foram obtidos através de testes psicométricos de escolha forçada.

Os estímulos foram apresentados aos avaliadores por meio de fones de ouvido, conectados diretamente à placa de som do computador programado para a execução do experimento.

A tarefa básica do avaliador em cada ensaio, após a apresentação intervalada de dois estímulos, foi definir qual apresentava menor nível de ruído. Como os níveis eram conhecidos, foi possível afirmar se a resposta estava correta ou incorreta e posteriormente calcular o próximo passo.

Para cada experimento, diferentes níveis de relação sinal-ruído de referência foram testados, já que a função psicométricas são definidas localmente (*i.e.*: os limiares para $SNR = 40$ dB podem ser diferentes de $SNR = 10$ dB). Desta forma, para cada nível foi usado um estímulo de referência fixo nas comparações e outro variável indicado pelo método. A ordem de apresentação do par era aleatória.

Implementação

As rotinas computacionais de reprodução de amostras, telas de interface com usuário e registro de decisão através de comandos do teclado foram implementadas usando as ferramentas do *toolbox* Psychtoolbox-3 (KLEINER et al., 2007) rodando no GNU/Octave (EATON et al., 2014) no sistema operacional Debian GNU/Linux 8. As análises e o controle do experimento adaptativos utilizaram os procedimentos do pacote Palamedes 1.8.2(KINGDOM; PRINS, 2010).

Exemplo de uso do método

A Figura 3.11 mostra a sucessão dos ensaios em uma execução do método. Foram executados 25 ensaios, os círculos preenchidos indicam respostas corretas e os círculos vazios, as incorretas.

Com as respostas, calcula-se a função psicofísica correspondente utilizando o ajuste por máxima verosimilhança de uma função do tipo Gumbel (KINGDOM; PRINS, 2010). A Figura 3.12 mostra a função estimada para este teste.

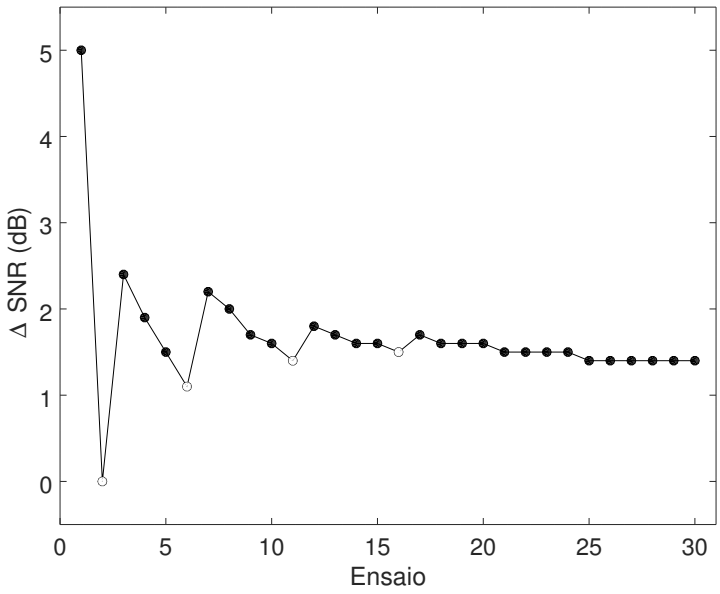


Figura 3.11 – Execução do método adaptativo com voz sintética e $SNR_{ref} = 30$ dB.

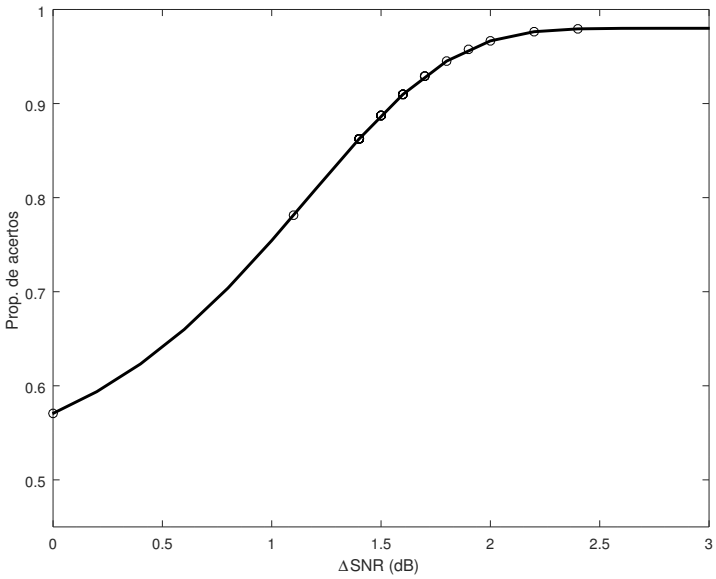


Figura 3.12 – Função psicofísica estimada com voz sintética e $SNR_{ref} = 30$ dB.

Experimento 1: tom puro com ruído rosa aditivo

No primeiro experimento para detecção dos limiares, os estímulos foram tons puros (220 Hz) com ruído rosa (Gaussiano) aditivo.

Os níveis de SNR testados foram 10 dB, 20 dB, 30 dB, 40 dB e 50 dB. Para cada nível um procedimento de detecção foi executado.

Além dos estímulos de referência, gerou-se também os estímulos de comparação com diferença de SNR em 0,1 dB na faixa de 0 a 60 dB (600 amostras). Os estímulos tinham 1,5 s de duração.

Três informantes participaram do experimento, produzindo os resultados da Figura 3.13.

Considerando que cada ensaio tem um valor de limiar médio μ_i com desvio σ_i, para os M ensaios tem-se um limiar médio associado μ, com respectivo desvio σ, calculados pelas expressões:

$$\mu = \frac{\sum_i \mu_i}{M} \tag{3.10}$$

$$\sigma = \left(\frac{\sum_i \sigma_i^2}{M} + \frac{\sum_{i<j} (\mu_i - \mu_j)^2}{M^2} \right)^{1/2} \tag{3.11}$$

Desta forma, o resultado combinado dos informantes é apresentado na Figura 3.14.

Experimento 2: voz sintética com ruído glótico

O procedimento anterior foi repetido para amostras de voz sintética soprosa.

Utilizando o modelo fonte-filtro, adicionou-se ruído controlado na fonte do sinal e filtrou-se com a vogal desejada. Os estímulos usados neste experimento tinham freqüência fundamental de 220 Hz, sem *jitter* e vogal /a/. Cada elocução tinha 1,5 segundos de duração. Neste experimento as referências foram posicionados em 10 dB, 20 dB, 30 dB e 40 dB.

Três informantes participaram do experimento, produzindo os resultados da Figura 3.15 e de forma combinada na Figura 3.16.

3.2.4 Discussão de resultados

Experimento 1: tom puro com ruído rosa

Neste experimento, o objetivo foi estimar o limiares de percepção de ruído combinado ao tom puro, em 5 níveis diferentes de SNR.

Os resultados obtidos estão exibidos nas Figuras 3.13, 3.14 e na Tabela 3.11.

A Figura 3.13 mostra o resultado individual de cada avaliar, nas regiões das respectivas regiões de SNR de referência, com o valor médio do limiar determinado e o respectivo valor de desvio padrão.

A Figura 3.14 mostra o agrupamento dos resultados dos avaliadores. Neste caso, é possível perceber uma tendência de diminuição no valor de limiar com o aumento da relação sinal-ruído. Os limiares médios determinados variam de 0,8 dB a aproximadamente 2,0 dB.

Tabela 3.11 – Limiares detectados pelo método adaptativo: tom, $N = 3$, grandezas em dB

	1		2		3		Médio	
Estímulo	Limiar	desvio	Limiar	desvio	Limiar	desvio	Limiar	desvio
50	0,55	0,16	0,88	0,17	1,00	0,20	0,81	0,26
40	1,56	0,36	1,20	0,20	0,71	0,28	1,15	0,45
30	1,29	0,48	1,40	0,40	1,98	0,25	1,55	0,49
20	1,53	1,83	1,83	0,11	2,60	0,20	1,98	0,51
10	1,92	0,78	2,40	0,30	1,47	0,19	1,93	0,62

Experimento 2: voz sintética com ruído glótico

Com a mesma metodologia do experimento anterior, mas utilizando como estímulos amostras de voz sintéticas com ruído glótico adicionado na fonte, buscou-se estimar os limiares de percepção do ruído nesta situação.

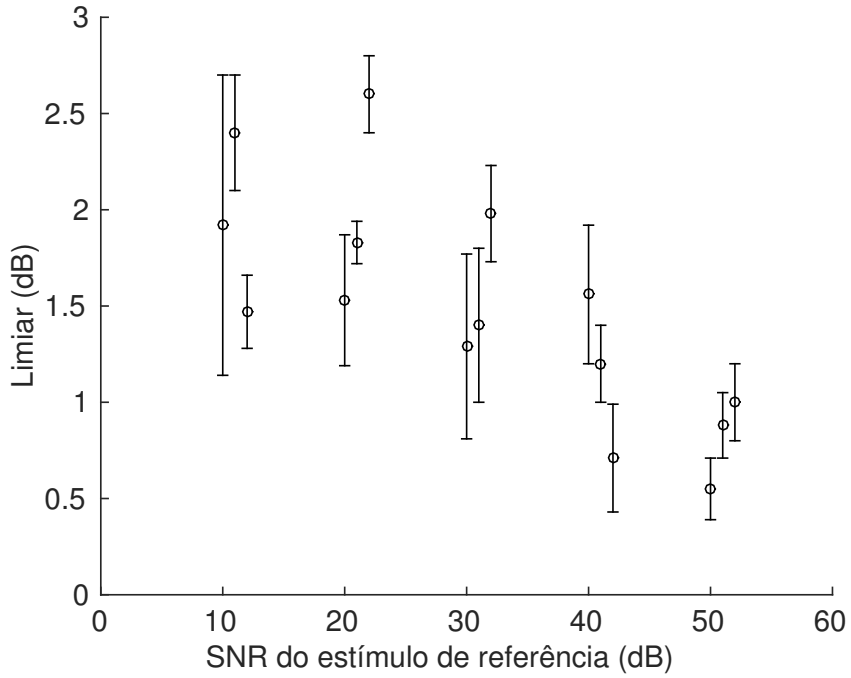


Figura 3.13 – Limiares detectados pelo método adaptativo: tom, $N = 3$

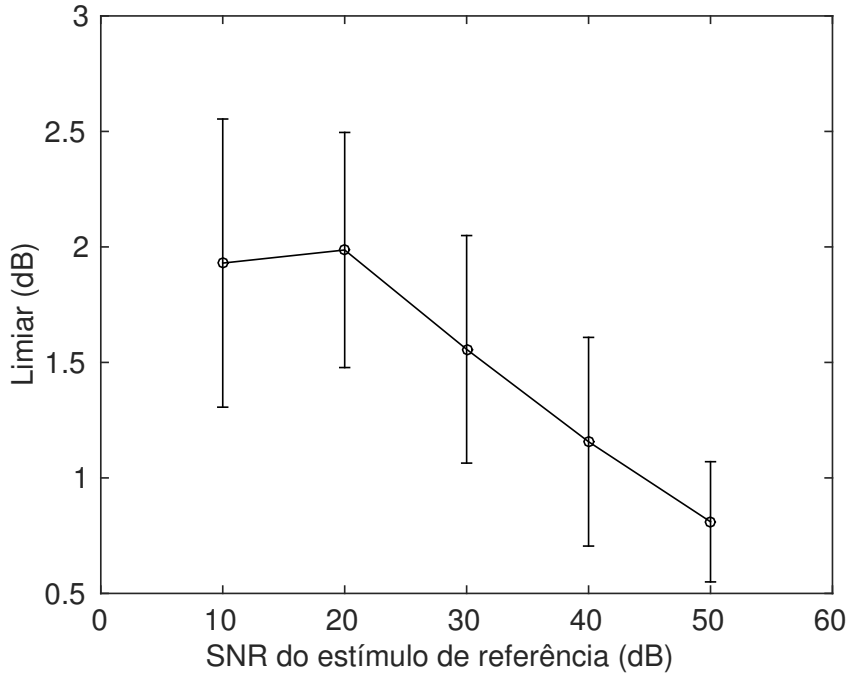


Figura 3.14 – Limiares detectados pelo método adaptativo: tom, observadores combinados

Os resultados obtidos estão exibidos nas Figuras 3.15 (resultados individuais dos avaliadores), 3.16 (resultados agrupados) e na Tabela 3.12.

Neste experimento observa-se que os limiares nas faixas de relação sinal-ruído de 20, 30 e 40 dB têm valores próximos (em média, 1,3 a 1,7 dB). O limiar na faixa de 10 dB apresentou média de 2,3 dB.

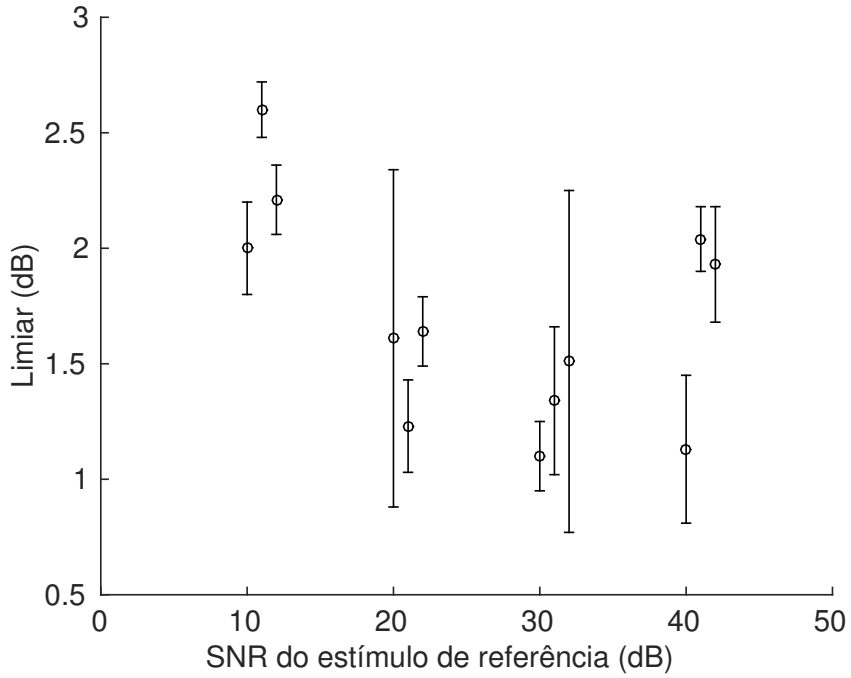


Figura 3.15 – Limiares detectados pelo método adaptativo: voz sintética, $N = 3$

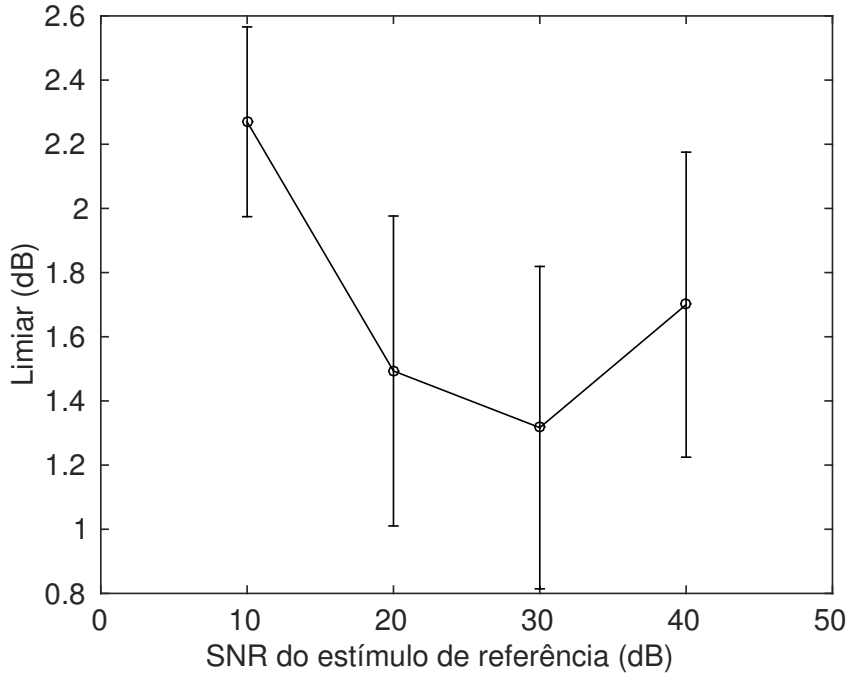


Figura 3.16 – Limiares detectados pelo método adaptativo: voz sintética, observadores combinados

Observações sobre os limiares obtidos

Nos experimentos descritos, estimou-se limiares com estímulos baseados em tons puros e voz sintética. No primeiro caso, a média de limiares cobriu o intervalo de 0,8 a 2,0 dB, e no segundo, 1,3 a 2,3 dB.

Os resultados, mesmo considerando os erros de estimação, implicam que é possivel

Tabela 3.12 – Limiares detectados pelo método adaptativo: voz sintética, $N = 3$, grandezas em dB

	1		2		3		Médio	
Estímulo	Limiar	desvio	Limiar	desvio	Limiar	desvio	Limiar	desvio
40	1,13	0,32	2,04	0,14	1,93	0,14	1,70	0,47
30	1,10	0,15	1,34	0,32	1,51	0,74	1,31	0,50
20	1,61	0,73	1,23	0,20	1,64	0,15	1,49	0,48
10	2,00	0,20	2,60	0,12	2,21	0,15	2,27	0,30

a discriminação de amostras separadas de 3 dB em relação sinal-ruído.

3.3 Considerações finais do capítulo

As características extraídas das escalas de diferenças e dos limiares de percepção da soprosidade aqui desenvolvidos servirão de subsídios no desenvolvimento do método de avaliação comparativa por busca binária que será descrito no próximo capítulo.

Outra consideração que deve ser feita é que alguns efeitos importantes não foram quantificados: a variação de escalas e limiares na presença de perturbações simultâneas na freqüência fundamental, a variação da freqüência fundamental (caso de vozes masculinas comparadas com femininas). Estes efeitos podem ser caracterizados por um método que avalia variações conjuntas, como a *Maximum Likelihood Conjoint Measurement* (KNOBLAUCH; MALONEY, 2012).

4 Avaliação comparativa da percepção de soprosidade

4.1 Introdução

A avaliação perceptiva-auditiva da qualidade de voz pelos profissionais é uma atividade freqüente na clínica. Diversos protocolos são estabelecidos e descritos na literatura, sendo os mais difundidos os métodos VPAS (LAVER et al., 1981) e GRBAS(HIRANO, 1981).

No entanto, diversas limitações dos referidos métodos foram explicitadas ao longo do anos, relativos à baixa concordância inter e intra-avaliadores, a validade de escalas e o caráter subjetivo da avaliação, como por exemplo a dependência de treinamento e experiência do julgador(KREIMAN; GERRATT, 1998; KREIMAN; GERRATT; ITO, 2007).

Uso de âncoras de referência como maneira de aumentar a consistência inter e intra-avaliadores vem sendo estudado (YIU; CHAN; MOK, 2007; AWAN; LAWSON, 2009; EADIE; KAPSNER-SMITH, 2011). Boa parte dos trabalhos, porém, concentra-se no treinamento dos avaliadores (BARSTIES et al., 2017; BRINCA et al., 2015), visando reforço no aprendizado de referências internas, sem o intuito do uso de âncoras no processo de avaliação em si.

A abordagem tradicional, sem o uso de referências, pode ser comparada ao fenômeno do ouvido absoluto, uma habilidade muito rara que permite que uma pessoa tenha a capacidade de identificar uma determinada nota musical de forma isolada (sem um tom de referência). Estudos estimam que o "ouvido absoluto" ocorre em apenas 1 de 10.000 pessoas (BACHEM, 1955). Na música, por exemplo, esta limitação da maioria dos seres humanos é compensada através do chamado "ouvido relativo", permitindo a afinação de instrumentos comparando com outros que geram tons de referência (e.g. diapasão).

O uso de âncoras visaria introduzir o conceito de "ouvido relativo" na avaliação da qualidade de voz. Neste trabalho, uma nova metodologia para a avaliação da qualidade de voz soprosa baseada em âncoras é proposta.

4.2 Estudos preliminares: ordenação de voz soprosa

4.2.1 Estudo preliminar: ordenação de voz soprosa sintética

Método

O primeiro experimento da avaliação comparativa busca ordenar, por grau de distúrbio, amostras de voz sintética feminina (vogal /a/, $f_0 = 220$ Hz). Para dar naturalidade às amostras, incluiu-se *jitter* de 0,2% conforme descrito na sub-seção 2.3.1, e ruído branco gaussiano é adicionado ao fluxo glótico, sendo, posteriormente, filtrado pelo trato vocal. As amostras geradas têm relação sinal-ruído conhecida.

No caso de se n amostras, se for feita a comparação de cada par, são necessárias $n(n-1)/2$ comparações, ou seja, tem-se um procedimento com complexidade $O(n^2)$. Para diminuir o número de comparações necessárias, utiliza-se um método ótimo de ordenação, a Ordenação Rápida (*Quicksort*).

O método de *Quicksort* (SEDGEWICK; WAYNE, 2011) é um algoritmo de "dividir para conquistar". O método inicialmente divide um lista maior em duas sub-listas menores: elementos inferiores e elementos superiores. O algoritmo então ordena recursivamente as sub-listas.

Os passos são:

- Tomar um elemento, denominado pivô da lista;
- Reordenar a lista de forma que todos os elementos menores que o pivô venham antes dele, enquantos os elementos maiores venham depois (para o caso de valores iguais, é indiferente a posição). Após este particionamento, o pivô estará na sua posição final.
- Aplicar recursivamente os passos acima para sub-listas de elementos com menor valor e, separadamente, para sub-listas de elementos com valor maior.

O critério de parada para recursão são listas de tamanho zero ou um, que não necessitam ser ordenadas. O método tem complexidade média $O(n \log n)$.

O código fonte da listagem 4.1 mostra uma implementação em Pascal, na qual foi baseada a implementação utilizada neste trabalho. O vetor a ser ordenado é a.

O algoritmo da ordenação rápida exige apenas que se tenha um definição de métrica, isto é, se um item é maior ou igual ao outro. Esta decisão é passada ao avaliador, que é inquirido através da interface computacional. Esta é a única diferença da *Quicksort* tradicional, onde o computador (por hardware ou software) faz a comparação.

Listing 4.1 – Código fonte do *Quicksort* em Pascal

```
procedure quicksort (l, r : integer);
var v, t, i, j : integer;
begin
   if r>l then
   begin
      v:=a[r]; i:=l-1;j:=r;
      repeat
         repeat i:=i+1 until a[i]>=v;
         repeat j:=j-1 until a[j]<=v;
         t:=a[i]; a[i]:=a[j]; a[j]:=t;
      until j <= i;
      a[j]:=a[i]; a[i]:=a[r]; a[r]:=t;
      quicksort(l, i-1);
      quicksort(i+1,r);
   end
end;
```

Implementação

Os experimentos descritos foram implementados utilizando GNU/Octave (EATON et al., 2014) e o pacote Psychtoolbox (KLEINER et al., 2007).

O referido pacote oferece facilidades como captura de teclado, exibição de estímulos visuais em tela cheia e reprodução de áudio com baixa latência.

A tarefa básica, que se repete em todos os experimentos, é escutar um par de amostras (amostras A e B) e avaliar se: "A apresenta maior distúrbio que B?", "B apresenta maior distúrbio que A?", "não é possível distinguir". Caso o avaliador não tenha uma decisão, ou se distraiu durante a execução das amostras, é permitida a repetição da reprodução.

Resultados e discussão

Em uma primeira tentativa, varreu-se uma faixa de relação sinal-ruído de 0-42 dB, com passo de 7 dB entre os estímulos ($n = 7$). A ordem inicial das amostras foi aleatória, para evitar tendências. A tarefa, neste caso, foi trivial (não houve dúvida na discriminação das amostras), e para o avaliador que participou deste experimento (o autor do trabalho), a ordem final obedeceu a relação sinal-ruído em duas execuções repetidas do experimento.

O passo seguinte foi aumentar o número de amostras (e diminuir o passo), fazendo $n = 14$ e o passo de 3,0 dB, valor próximo ao estimado no Capítulo 3 para a *jnd* da percepção do ruído em voz sintética. A faixa varrida de SNR foi de 2 a 44 dB.

Neste caso, a tarefa foi mais complexa que a tentativa anterior, tanto em relação

ao número de avaliações necessárias, quanto ao momento da avaliação, já que em alguns casos a separação entre os estímulos é menor, requerendo maior atenção do avaliador.

O avaliador fez duas execuções do teste de ordenação. A ordem encontrada em cada execução foi (ordenando os estímulos pelo nível de perturbação, referidos pela SNR):

- Execução 1: [41, 44, 35, 38, 32, 29 , 26, 23, 20, 17, 14, 11, 8, 5, 2] dB
- Execução 2: [41, 44, 38, 35, 32, 29 , 26, 23, 20, 17, 14, 11, 8, 5, 2] dB.

Ou seja, duas inversões de posição na primeira execução (41 e 44, 35 e 38), e uma na segunda (41 e 44).

Este experimento indicou que é possível ordenar amostras com diferentes níveis de relação sinal-ruído. No entanto, este método apresenta uma limitação em sua aplicação: quando a diferença dos estímulos nas comparações está próxima à mínima distância perceptível, podem ocorrer erros na avaliação, gerando instabilidade no método da *QuickSort*, possível fonte das inversões encontradas no teste anterior.

O experimento indicou também que para estes estímulos sintéticos, a resolução possível está próxima de 3 dB, em acordo com os experimentos do Capítulo 3, obtidos nos experimentos com MLDS e na estimação da função psicométrica da soprosidade.

4.2.2 Estudo preliminar: Ordenação de voz soprosa humana

Método

Seguindo a metodologia apresentada na sub-seção anterior, desejou-se ordenar por grau de distúrbio, as amostras já referidas na sub-seção 2.3.6. Estas amostras foram previamente avaliadas em uma escala de 7 níveis (3 amostras por nível), através de um processo de ordenação análogo ao descrito neste trabalho.

Cada nível de soprosidade recebe valor entre 0 e 3,0, com passo de 0,5. Como cada nível tem 3 amostras, estas são referidas como a, b e c. Desta forma, a amostra "b" com nível de soprosidade de 1,5 é designada como $B(1{,}5; b)$.

O conjunto de amostras é:

$B(0{,}0; a)$	$B(0{,}5; a)$	$B(1{,}0; a)$	$B(1{,}5; a)$	$B(2{,}0; a)$	$B(2{,}5; a)$	$B(3{,}0; a)$
$B(0{,}0; b)$	$B(0{,}5; b)$	$B(1{,}0; b)$	$B(1{,}5; b)$	$B(2{,}0; b)$	$B(2{,}5; b)$	$B(3{,}0; b)$
$B(0{,}0; c)$	$B(0{,}5; c)$	$B(1{,}0; c)$	$B(1{,}5; c)$	$B(2{,}0; c)$	$B(2{,}5; c)$	$B(3{,}0; c)$

Neste estudo, buscou-se confirmar a avaliação prévia, desta vez testando procedimento da ordenação rápida. Dois voluntários participaram do experimento.

Resultados e discussão

Como no caso de voz sintética, boa parte das comparações foi trivial, pois são muito discrepantes os níveis de perturbação. Os extremos da escala foram os mais fáceis de serem discriminados e desta forma suas posições após a ordenação foram consistentes com a avaliação anteriormente feita.

A concordância nas classes entre os dois voluntários foi alta, inclusive em relação a uma amostra com posição incoerente com a pré-avaliação. Esta amostra, $B(1{,}0; c)$, tinha sido avaliada como grau 1,0 anteriormente, e na ordenação foi enquadrada como grau 2,0 por ambos avaliadores.

Para verificar a razão da incoerência, foi-se verificar a forma de onda do sinal. Constatou-se que esta amostra apresentava instabilidade em termos de perturbação, com um trecho inicial com alta perturbação e na seqüência um trecho estável, com menor influência do ruído. Este fenômeno indica a possibilidade de mascaramento quando existem flutuações ao longo da execução (isto é, a perturbação inicial faz percepção crer em um nível de percepção mais alto ao longo da elocução).

Ressalta-se a necessidade de se quantificar as mínimas diferenças perceptíveis, perturbações de natureza episódica podem afetar o julgamento e a possibilidade de mascaramento quando ocorrem perturbações distintas simultâneas (tal como ocorre nas medidas objetivas).

Os métodos de ordenação fornecem elementos que corroboram a utilidade da avaliação comparativa. No entanto, não existe nestes casos uma métrica evidente para avaliar os métodos. As análises aqui feitas foram qualitativas e com pequena amostra. A ordenação por si não irá indicar a classificação do grau de soprosidade da amostra. A confusão na ordenação, no entanto, é um indício que duas amostras pertencem a uma mesma classe.

4.3 Classificação de voz soprosa por busca em árvore binária

Nesta seção, um método para classificação de voz soprosa baseado na busca em árvores binárias é proposto.

Para o desenvolvimento deste método, baseou-se nos resultados da caracterização psicofísica da soprosidade, desenvolvida no capítulo 3 e nas propriedades e operações existentes nas árvores binárias, estrutura recorrente na computação (SEDGEWICK; WAYNE, 2011).

O método será descrito nas sub-seções seguintes.

4.3.1 Árvore binária

Uma árvore binária de busca (*binary search tree*) é uma estrutura de dados de árvore binária baseada em nós (SEDGEWICK; WAYNE, 2011). Na maior parte dos casos, por convenção, todos os nós da sub-árvore esquerda possuem um valor numérico inferior ao nó raiz e todos os nós da subárvore direita possuem um valor superior ao nó raiz. Esta forma de estruturação permite a chamada "busca binária".

Alguns conceitos sobre árvores binárias (exemplificados pela Figura 4.2) :

- Nós: são todos os itens armazenados na árvore
- Raiz: é o nó do topo da árvore (e.g.: nó 1,5)
- Filhos - são os nós que vêm depois dos outros nós (nó 2,5 é filho de 1,5).
- Pais - são os nós que vêm antes dos outros nós (2,5 é pai de 2,0)
- Folhas - são os nós que não têm filhos; são os últimos nós da árvore (o nós 1,25 e 1,75 são folhas).

4.3.2 Busca em árvore binária

O procedimento de busca em árvore binária é simples inicia-se examinando o nó raiz. Se o valor encontrado é igual ao da raiz, a busca é bem sucedida, e retorna-se o nó. Se o valor é inferior ao da raiz, a busca segue pela sub-árvore da esquerda. No caso de valor superior, a busca segue pela sub-árvore da direita. O procedimento é repetido até que o valor procurado seja encontrado ou se atinja uma sub-árvore nula (sub-árvore inexistente), retornando um valor nulo na função, isto é, não existe tal registro na árvore.

4.3.3 Método

A árvore binária é formada em seus nós (exceto as folhas) por âncoras com grau de soprosidade pré-estabelecido. Busca-se inserir amostras de teste em uma árvore binária onde todos os nós têm uma amostra de referência como âncora. Portanto, as amostras âncora são comparadas perceptivamente com a amostra de teste.

Neste trabalho, foram utilizadas 3 ou 7 âncoras, pois permitem a distribuição das mesmas como nós de uma árvore binária equilibrada. As âncoras têm gradação de 0,0 a 3,0, com passo de 0,5. As árvores binárias para os casos de 3 e 7 âncoras estão mostradas nas Figura 4.1 e 4.2, respectivamente. Os nós que possuem âncoras são representados por quadrados. As folhas não têm âncoras associadas.

O procedimento de busca (classificação) consiste, inicialmente na comparação da amostra de teste com a amostra do nó central (a saber, grau 1,5). Se a amostra teste tem

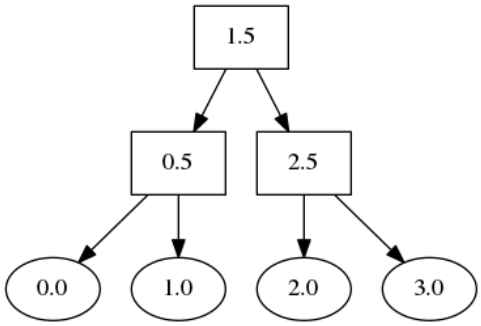


Figura 4.1 – Árvore no caso de 3 âncoras. As âncoras estão representadas por quadrados.

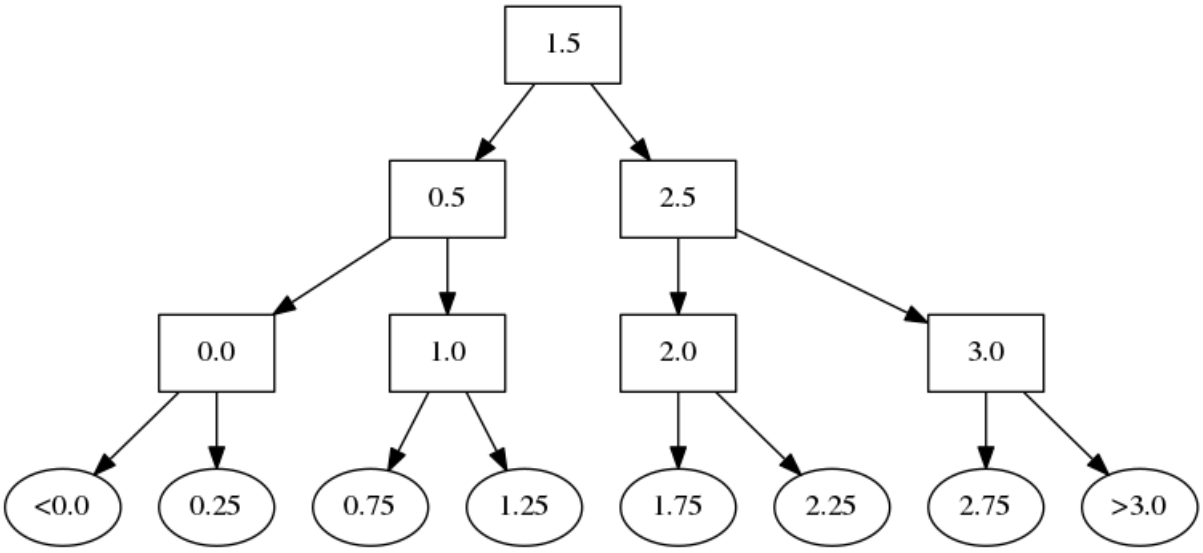


Figura 4.2 – Árvore no caso de 7 âncoras. As âncoras estão representadas por quadrados.

mesmo nível de perturbação que o nó âncora (ou não é possível a distinção), retorna-se o valor da âncora. Caso seja superior, segue-se para fazer a comparação com o nó da direita (grau 2,5). Caso seja inferior, segue-se para o nó da esquerda (grau 0,5). As comparações seguem até que não se consiga distinguir a diferença ou até se atingir as folhas da árvore.

Com as 7 âncoras (0,0; 0,5; 1,0; 1,5; 2,0; 2,5; 3,0) têm-se 4 níveis de profundidade. Assim, é possível chegar a 15 classificações, a saber ($<$ 0,0; 0,0; 0,25 ; 0,5; 0,75; 1,0; 1,25; 1,5; 1,75; 2,0; 2,25; 2,5; 2,75; 3,0; $>$ 3,0). Na prática, os níveis $<$ 0,0 e $>$ 3,0 são desconsiderados e, se a classificação atinge os referidos valores, são considerados como 0,0 e 3,0, respectivamente, com 13 níveis de classificação de fato. Uma amostra teste é classificada com até três comparações distintas.

Já com 3 âncoras (0,5; 1,5; 2,5) têm-se 3 níveis de profundidade, podendo chegar-se a 7 classificações: (0,0; 0,5; 1,0; 1,5; 2,0; 2,5; 3,0). A amostra teste é classificada em até duas comparações distintas. No protocolos tradicionais, que usam escalas discretas, como GRBAS e VPAS, o número de classificações é de 4 e 7, respectivamente.

Um questão importante é o estabelecimento das amostras âncoras para o nós. Neste

trabalho, as âncoras foram escolhidas com base nos experimentos do Capítulo 3.

A escolha de 7 âncoras (e os respectivos 13 níveis possíveis) para estudo da soprosidade pode parecer excessiva, mas é justificável pelos limiares encontrados no estudo psicofísico para vogais sintéticas. Por exemplo, considerando-se uma faixa dinâmica de 0-42 dB, alocam-se âncoras distantes entre si de 7 dB, atingindo-se a resolução entre os nós de 3,5 dB, em conformidade com os valores determinados anteriormente.

4.3.4 Implementação

As rotinas computacionais de reprodução de amostras, telas de interface com usuário e registro de decisão através de comandos do teclado foram implementadas usando as ferramentas do *toolbox* Psychtoolbox-3 (KLEINER et al., 2007) rodando no GNU/Octave (EATON et al., 2014) no sistema operacional Debian GNU/Linux 8.

O métodos e estrutura de dados para implementação da árvore binária foram baseados nos algoritmos mostrados em Sedgewick e Wayne (2011).

O tratamento estatístico foi feito usando o software R (R Core Team, 2014), em especial, o pacote psych (REVELLE, 2018).

Os testes foram realizados com os fones de ouvido intra-auriculares sem marca definida caracterizados no Capítulo 3.

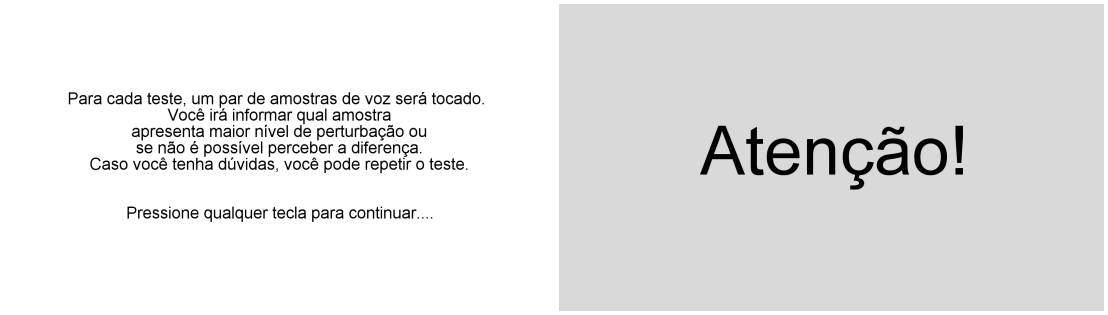
A Figura 4.3 mostra exemplos de telas exibidas ao usuário durante a execução dos testes. Nesta implementação, os julgadores não foram informados de maneira explícita que estavam executando um procedimento de busca binária. Uma versão alternativa desta implementação poderia apresentar como interface gráfica estruturas semelhantes às Figuras 4.1 e 4.2, que não foi testada neste trabalho.

4.3.5 Estudos pilotos e experimentos

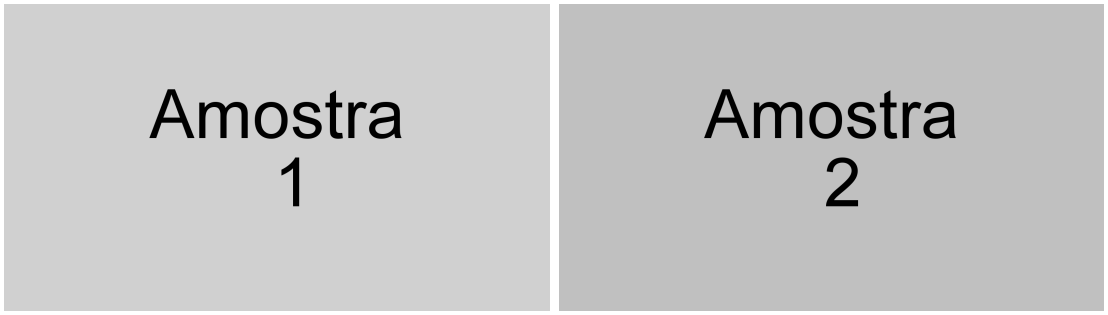
Uma série de estudos pilotos e experimentos, com diferentes tipos de estímulos de teste e âncoras foram realizados para avaliar a aplicação do método da busca binária na avaliação da soprosidade vocal, em vogais /a/ sustentadas.

Os pilotos são experimentos limitados, realizados por apenas um avaliador (o autor do trabalho) e serão assim designados:

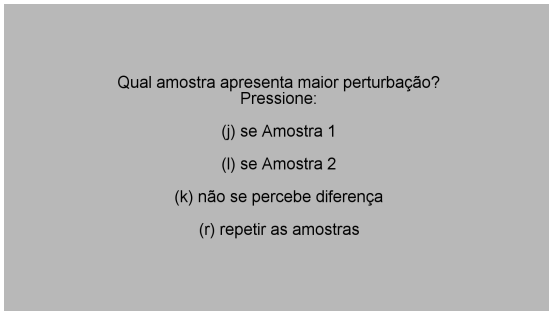
- Piloto 1: Avaliação de voz sintética ($f_0 = 220$ Hz), 3 âncoras sintéticas ($f_0 = 220$ Hz);
- Piloto 2: Avaliação de voz sintética ($f_0 = 220$ Hz), 7 âncoras sintéticas ($f_0 = 220$ Hz);



(a) Tela de instruções inicial (b) Alerta para início da tarefa (alerta sonoro)



(c) Reprodução da amostra 1 (d) Reprodução da amostra 2



(e) Tela de entrada de decisão

Figura 4.3 – Exemplo de funcionamento da plataforma de testes de percepção

- Piloto 3: Avaliação de voz soprosa natural, 3 âncoras naturais (f_0 não controlado);
- Piloto 4: Avaliação de voz soprosa natural, 7 âncoras naturais (f_0 não controlado);
- Piloto 5: Avaliação de voz soprosa natural, 7 âncoras sintéticas ($f_0 = 220$ Hz);
- Piloto 6: Avaliação de voz sintética ($f_0 = 120$ Hz), 7 âncoras sintéticas ($f_0 = 220$ Hz).

O intuito de desenvolvimento dos estudos pilotos foi testar a viabilidade do método da busca binária, verificando a consistência intra-julgador em um estudo limitado.

Os experimentos ampliam os pilotos 1, 2, 3 e 4 para um maior número de avaliadores:

- Experimento 1: Avaliação de voz sintética (f_0 = 220 Hz), 3 âncoras sintéticas (f_0 = 220 Hz);
- Experimento 2: Avaliação de voz sintética (f_0 = 220 Hz), 7 âncoras sintéticas (f_0 = 220 Hz);
- Experimento 3: Avaliação de voz soprosa natural, 3 âncoras naturais (f_0 não controlado);
- Experimento 4: Avaliação de voz soprosa natural, 7 âncoras naturais (f_0 não controlado).

Estudos pilotos

Piloto 1: Avaliação de voz sintética (f_0 = 220 Hz), 3 âncoras sintéticas (f_0 = 220 Hz)

No primeiro piloto, desejou-se classificar voz sintética com 3 âncoras sintéticas. A árvore tem o formato da Figura 4.1, sendo âncoras de 1.5, 0.5 e 2.5, amostras com relação sinal-ruído de 21, 35 e 7 dB, respectivamente.

Todas as amostras tinham a mesma freqüência fundamental ($f_0 = 220Hz$), vogal /a/ e *jitter* de 0,2%.

Cada ensaio envolvia a classificação de 14 amostras distintas, com valores de SNR variando entre 2 e 44 dB (com passo de 3 dB). As amostras para classificação tinham a mesma f_0, vogal e *jitter* das âncoras.

O avaliador (autor do trabalho) repetiu o experimento em 3 ocasiões. O resultado das avaliações está apresentado na Figura 4.4a. A correlação entre as avaliações está mostrada na Figura 4.5a e dispersão na Figura 4.6a.

As amostras classificadas têm relação sinal-ruído crescente (o item 1 tem SNR de 2 dB, o item 14 apresenta SNR de 44 dB). Pode-se perceber que dispersão nas três execuções não foi superior em nenhum caso a meio grau na escala definida de soprosidade.

Nota-se que as execuções não são idênticas, mas apresentam alto grau de correlação entre si, 0,96, 0,98 e 0,98.

Piloto 2: Avaliação de voz sintética (f_0 = 220 Hz), 7 âncoras sintéticas (f_0 = 220 Hz)

O caso anterior foi modificado para incluir 7 âncoras sintéticas. A árvore tomou a forma da Figura 4.2, com âncoras de 1,5 (nó pai, primeiro nível), 0,5 e 2,5 (segundo nível), 0,0, 1,0, 2,0 e 3,0 (terceiro). As âncoras utilizadas no primeiro e segundo níveis foram as mesmas do piloto 1. Para o terceiro nível, as âncoras 0,0, 1,0, 2,0 e 3,0, tinham relação sinal-ruído de 42 dB, 28 dB, 14 dB, 1.0 dB.

As mesmas amostras de classificação do caso anterior foram utilizadas.

O avaliador repetiu o experimento em 3 ocasiões. O resultado das avaliações está apresentado na Figura 4.4b. A correlação entre as avaliações está mostrada na Figura 4.5b e a dispersão na Figura 4.6b.

No piloto 2, a resolução possível é maior que no piloto 1, 0,25 contra 0,5 na avaliação da soprosidade, fazendo que o valor médio das avaliações ficasse mais uniformemente distribuído na escala de percepção. Em conseqüência, a dispersão é ligeiramente menor e a correlação apresentada também foi próxima da unidade.

Os pilotos 1 e 2 mostraram alta concordância intra-avaliador, com coeficiente de correlação intraclasse $ICC(1) = 0{,}97 \pm 0.3$ e $ICC(1) = 0{,}98 \pm 0.2$, respectivamente, para um intervalo de confiança de 95%.

Para testar a consistência inter-julgadores do método, os pilotos 1 e 2 foram expandidos nos experimentos 1 e 2, com um número maior de participantes.

Piloto 3: Avaliação de voz soprosa natural, 3 âncoras naturais

Nos pilotos anteriores, utilizou-se voz sintética onde foi possível o controle dos parâmetros tanto nas amostras âncora quanto nas amostras de classificação.

Neste caso, a abordagem foi diferente, utilizando amostras de voz natural com características predominantemente soprosas, já referidas nos capítulos anteriores desta tese.

As amostras de classificação e âncoras eram naturais, e portanto, tinham parâmetros distintos entre si, além de possíveis ocorrências de perturbações simultâneas à soprosidade durante a elocução, podendo afetar no julgamento dos avaliadores.

Utilizou-se a classificação prévia para as 3 amostras âncoras, escolhidas as amostras $B(1{,}5; b)$, $B(0{,}5; a)$ e $B(2{,}5; c)$.

As amostras de teste para classificação tinham as seguintes classificações prévia, na ordem dos itens 1 ao 14: $B(0{,}0; b)$, $B(0{,}0; c)$, $B(0{,}5; b)$, $B(0{,}5; c)$ $B(1{,}0; b)$, $B(1{,}0; c)$, $B(1{,}5; a)$, $B(1{,}5; c)$, $B(2{,}0; a)$, $B(2{,}0; c)$, $B(2{,}5; a)$, $B(2{,}5; b)$, $B(3{,}0; b)$, $B(3{,}0; c)$. As amostras foram apresentadas aleatoriamente.

Foram feitas três execuções deste teste. O *boxplot* está mostrado na Figura 4.4c, assim como a correlação (Figura 4.5c) e a dispersão (Figura 4.6c).

As correlações foram 0,94, 0,95 e 0,90 entre as execuções, valores inferiores aos casos com voz sintética. O ICC(1) foi encontrado no intervalo de 0,82 e 0,97 para um intervalo de confiança de 95%. Nota-se todavia que as classificações ficaram dentro de 0,5 grau, exceto a amostra 8 que variou entre 1 e 2,5.

Em relação à classificação prévia, a diferença também ficou em 0.5 na maioria das amostras, exceto nas amostras 7 e 12, além da amostra 8 com maior dispersão.

Piloto 4: Avaliação de voz soprosa natural, 7 âncoras naturais

Ampliou-se o número de âncoras neste piloto, incluindo o terceiro nível, com âncoras em $B(0{,}0; a)$, $B(1{,}0; a)$, $B(2{,}0; b)$ e $B(3{,}0; a)$. As amostras de classificação foram as mesmas do item anterior.

Como no piloto 3, foram feitas três execuções deste teste. O *boxplot* está mostrado na Figura 4.4d, assim como a correlação está na Figura 4.5d e a dispersão na Figura 4.6d.

As correlações foram 0,96, 0,91 e 0,94. O ICC(1) foi encontrado na faixa de 0,85 a 0,98 para um intervalo de confiança de 95%. . Os resultados são comparáveis aos vistos no piloto 3, levando-se em conta a pequeno número de execuções aqui apresentadas.

As amostras 5, 6, 9 e 12 apresentaram as maiores dispersões, próximas a um grau de soprosidade.

Nota-se que a maior variabilidade das amostras âncoras e das amostras de avaliação adicionam maior nível de dificuldade na avaliações

Os pilotos 3 e 4 serão expandidos em número de avaliadores distintos nos experimentos 3 e 4.

Piloto 5: Avaliação de voz soprosa natural, 7 âncoras sintéticas ($f_0 = 220Hz$)

No piloto 5, experimentou-se o uso de âncoras sintéticas na avaliação de voz natural. As âncoras utilizadas foram as mesmas do piloto 2 e as amostras de teste são as mesmas do piloto 3 e 4.

O *boxplot* está mostrado na Figura 4.4e, a correlação está na Figura 4.5e e a dispersão na Figura 4.6e.

Novamente, três execuções foram realizadas. As correlações entre as execuções foram 0,93, 0,95, 0,95. O ICC(1) ficou localizado entre 0,81 e 0,97 para um intervalo de confiança de 95%

No gráfico de *boxplot*, nota-se um vazio na faixa de percepção de 1,5 a 2,5, e um maior acúmulo nas faixas inferiores de soprosidade.

Isso da decorre da distribuição das amostras naturais em questão não ser a mesma das âncoras sintéticas (igualmente espaçadas em SNR). Uma maneira de se equilibrar essa alocação seria usar procedimentos como a MLDS para compatibilizar as escalas.

Piloto 6: Avaliação de voz sintética ($f_0 = 120$ Hz), 7 âncoras sintéticas ($f_0 = 220$ Hz)

No piloto 6, testou-se a avaliação de amostras com freqüência fundamental diferente das âncoras, especificamente, as âncoras tinham $f_0 = 220$ Hz e amostras de classificação $f_0 = 120$ Hz, sempre sintéticas com parâmetros controlados.

As âncoras utilizadas foram as mesmas dos pilotos 2 e 5. As amostras de classificação têm SNR variando de 0 dB (amostra 1) a 39 dB (amostra 14) (com passos de 3dB).

O *boxplot* está mostrado na Figura 4.4f, assim como a correlação está na Figura 4.5f e a dispersão na Figura 4.6f.

As correlações entre as três execuções ficaram em 0,96, 0,96 e 0,99, respectivamente. O ICC(1) ficou entre 0,94 e 0,99 para um intervalo de confiança de 95%.

Considerando os resultados obtidos neste estudo, pode-se observar que âncoras com diferente f_0 não impedem a aplicação do método. No entanto, deve-se ressaltar que a f_0 pode modificar o limiar de percepção, fato que não foi quantificado neste trabalho.

Experimentos

Experimento 1: Avaliação de voz sintética ($f_0 = 220$ Hz), 3 âncoras sintéticas ($f_0 = 220$ Hz)

Neste experimento, repetiu-se a configuração do estudo piloto 1, anteriormente descrito, com um número maior de participantes para a avaliação da confiabilidade inter-julgadores.

Fizeram parte deste experimento nove voluntários ($N = 9$), sem experiência prévia em avaliação de voz ou formação na área, faixa etária de 20 a 27 anos, estudantes de Engenharia da Universidade Federal de São João del-Rei (Campus Alto Paraopeba).

As respostas dos voluntários foram tabuladas e geraram os gráficos da Figura 4.7. A Figura 4.7a mostra a correlação entre os participantes e a Figura 4.7b mostra os respectivos *boxplots*.

No gráfico da Figura 4.7b nota-se que a amostras com menor relação sinal-ruído (1-6) apresentaram a menor dispersão. As amostras 8, 10 e 11 apresentaram dispersão máxima de 1,0 grau. A amostra com maior dispersão foi a 9. As amostras com maior relação sinal-ruído tiveram dispersão de até 0,5 grau.

Da Figura 4.7a pode-se notar que os informantes X4 e X9 apresentaram desempenho inferior na tarefa, por apresentarem correlações inferior a 0,8.

Os coeficientes de correlação intraclasse (ICC) encontrados estão na Tabela 4.1, indicando os limites superior e inferior para o intervalo de confiança de 95%.

O coeficiente de concordância de Kendall encontrado foi $W = 0{,}80$, a correlação média de $R = 0{,}84$.

Tabela 4.1 – Coeficientes de correlação intraclasse para experimento 1

	ICC	inferior	superior
ICC(1)	0,82	0,68	0,92
ICC(2)	0,82	0,68	0,92
ICC(3)	0,82	0,69	0,92
ICC(1,k)	0,98	0,95	0,99
ICC(2,k)	0,98	0,95	0,99
ICC(3,k)	0,98	0,95	0,99

Experimento 2: Avaliação de voz sintética ($f_0 = 220$ Hz), 7 âncoras sintéticas ($f_0 = 220$ Hz)

Neste experimento, repetiu-se a configuração do estudo piloto 2, anteriormente descrito, com um número maior de participantes para a avaliação da confiabilidade inter-julgadores.

Fizeram parte deste experimento quatorze voluntários ($N = 14$), sem experiência prévia em avaliação de voz ou formação na área, faixa etária de 20 a 27 anos, estudantes de Engenharia da Universidade Federal de São João del-Rei (Campus Alto Paraopeba).

As respostas dos voluntários foram tabuladas e geraram os gráficos da Figura 4.8. A Figura 4.8a mostra a correlação entre os participantes e a Figura 4.8b mostra o respectivo *boxplot*.

As dispersões das amostras de 1 a 8, 11, 13 e 14 foram inferiores a 0,5 grau de soprosidade. As amostras 9 e 10 apresentaram maior dispersão, apesar da maior parte das resposta terem ficado em uma faixa de 0,5 grau.

Os coeficientes de correlação intraclasse (ICC) encontrados estão na Tabela 4.2, indicando os limites superior e inferior para o intervalo de confiança de 95%.

Tabela 4.2 – Coeficientes de correlação intraclasse para experimento 2

	ICC	inferior	superior
ICC(1)	0,89	0,80	0,95
ICC(2)	0,89	0,80	0,95
ICC(3)	0,89	0,80	0,95
ICC(1,k)	0,99	0,98	1,00
ICC(2,k)	0,99	0,98	1,00
ICC(3,k)	0,99	0,98	1,00

O coeficiente de concordância de Kendall encontrado foi $W = 0{,}87$, a correlação média de $R = 0{,}91$.

Os índices de correlação, ICC e concordância calculados foram superiores aos do caso de 3 âncoras, indicando melhor desempenho dos participantes nesta tarefa.

Experimento 3: Avaliação de voz soprosa natural, 3 âncoras naturais

Neste experimento, repetiu-se a configuração do piloto 3, onde fez-se a análise de voz soprosa natural com 3 âncoras naturais. Para expandir o estudo, 6 voluntários participaram do experimento.

As respostas dos voluntários foram tabuladas e geraram os gráficos da Figura 4.9. A Figura 4.9a mostra a correlação entre os participantes e a Figura 4.9b mostra o respectivo *boxplot*.

Os coeficientes de correlação intraclasse (ICC) encontrados estão na Tabela 4.3, indicando os limites superior e inferior para o intervalo de confiança de 95%.

Tabela 4.3 – Coeficientes de correlação intraclasse para experimento 3

	ICC	inferior	superior
ICC(1)	0,80	0,65	0,92
ICC(2)	0,80	0,65	0,92
ICC(3)	0,83	0,68	0,93
ICC(1,k)	0,96	0,92	0,99
ICC(2,k)	0,96	0,92	0,99
ICC(3,k)	0,97	0,93	0,99

O coeficiente de concordância de Kendall encontrado foi $W = 0{,}84$, a correlação média de $R = 0{,}85$.

Experimento 4: Avaliação de voz soprosa natural, 7 âncoras naturais

O experimento 4 é extensão do piloto 4, com 5 voluntários como julgadores. Neste caso, foram usada 7 âncoras naturais para comparação de amostras de voz natural.

As respostas dos voluntários foram tabuladas e geraram os gráficos da Figura 4.10. A Figura 4.10a mostra a correlação entre os participantes e a Figura 4.10b mostra o respectivo *boxplot*.

Os coeficientes de correlação intraclasse (ICC) encontrados estão na Tabela 4.4, indicando os limites superior e inferior para o intervalo de confiança de 95%.

O coeficiente de concordância de Kendall encontrado foi $W = 0{,}92$, a correlação média de $R = 0{,}91$.

Tabela 4.4 – Coeficientes de correlação intraclasse para experimento 3

	ICC	inferior	superior
ICC(1)	0,88	0,76	0,95
ICC(2)	0,88	0,76	0,95
ICC(3)	0,89	0,79	0,96
ICC(1,k)	0,97	0,94	0,99
ICC(2,k)	0,97	0,94	0,99
ICC(3,k)	0,98	0,95	0,99

As correlações foram superiores ao caso com 3 âncoras, como ocorrera nos experimentos com voz sintética. Porém, o número de participantes foi reduzido nos experimentos com voz natural, podendo ser a variação devida a um efeito amostral.

Uma queixa comum nos experimentos com voz natural foi a dificuldade de ser comparar amostras com âncoras muito diferentes. Avaliadores mais experientes potencialmente terão um melhor desempenho, mas este efeito não foi testado neste trabalho.

4.4 Considerações finais do capítulo

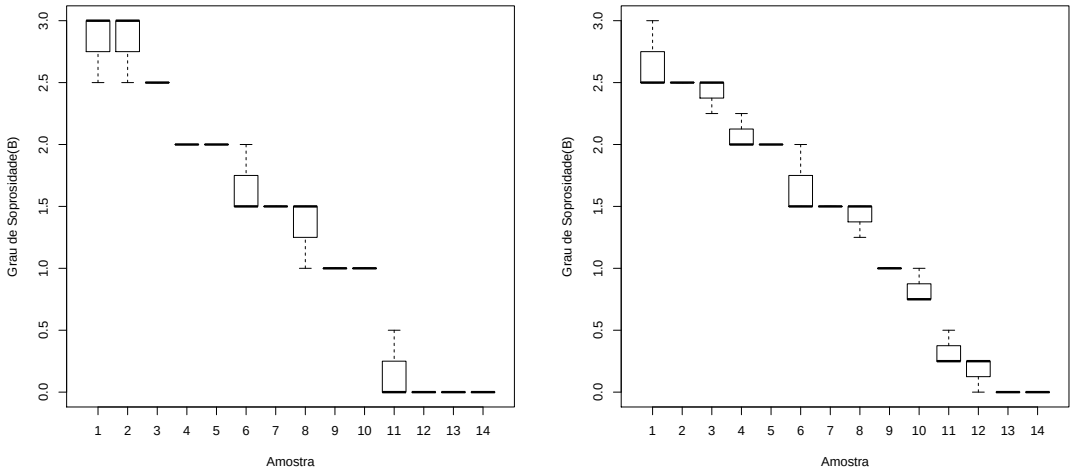
Neste capítulo, demonstrou-se um método simples para avaliação de voz soprosa, com potencial aplicação clínica.

Nos experimentos aqui propostos, o método da busca em árvore binária apresentou níveis elevados de confiabilidade inter e intra-avaliadores, sobretudo nas avaliações envolvendo voz sintética.

A comparação com outras metodologias é difícil, pela variedade de protocolos aplicados e mesmo da diversidade de índices utilizados para medição do desempenho. Maryn et al. (2009) tabularam alguns destes métodos com dados sobre as metodologias.

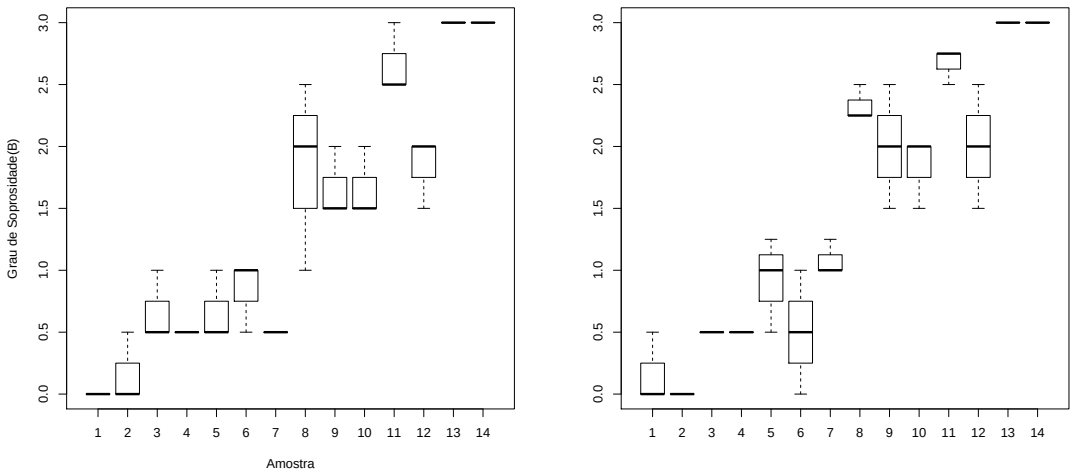
Os resultados obtidos em Kreiman, Gerratt e Berke (1994) podem servir de referência para comparação com o presente estudo, quando foi testado um grupo de 8 julgadores experientes, avaliando a soprosidade em escalas de 7 pontos. Com duas avaliações para cada amostra, a correlação média entre as leituras foi de 0,81, variando na faixa de 0,63-0,92 entre os participantes. Na comparação inter-julgadores, a correlação média foi 0,69, variando entre 0,44 e 0,86.

As correlações obtidas no presente trabalho foram superiores, em geral, às do estudo citado. Deve-se ainda aumentar o número de sujeitos nos experimentos com voz real, visando dar maior significância aos resultados.



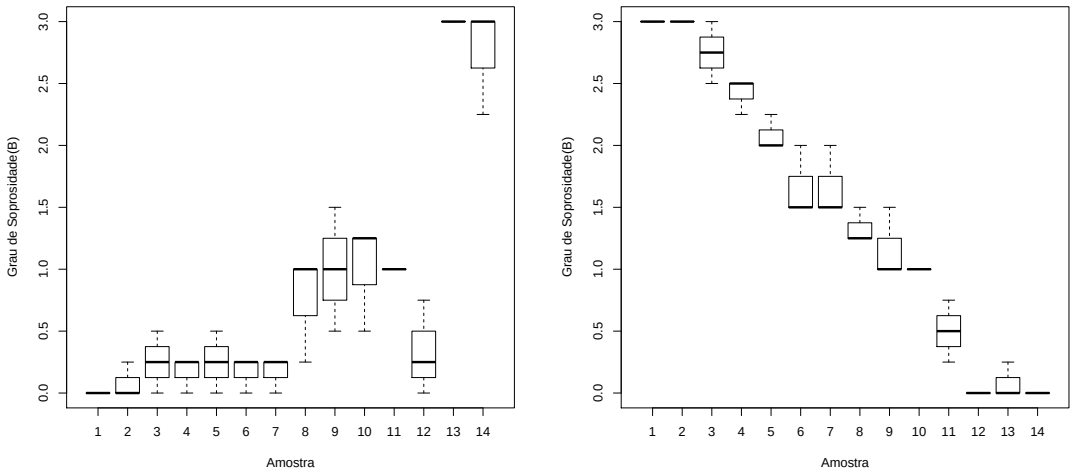
(a) Piloto 1

(b) Piloto 2



(c) Piloto 3

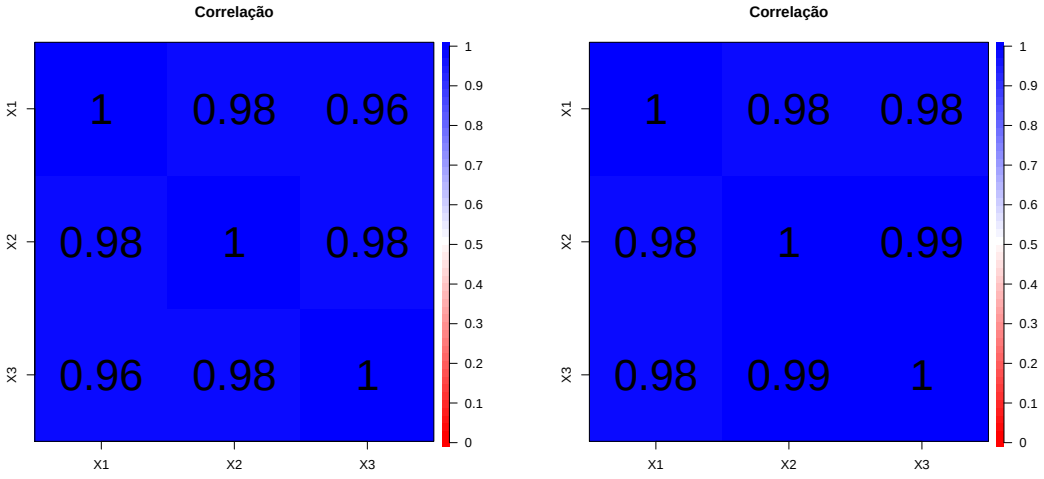
(d) Piloto 4



(e) Piloto 5

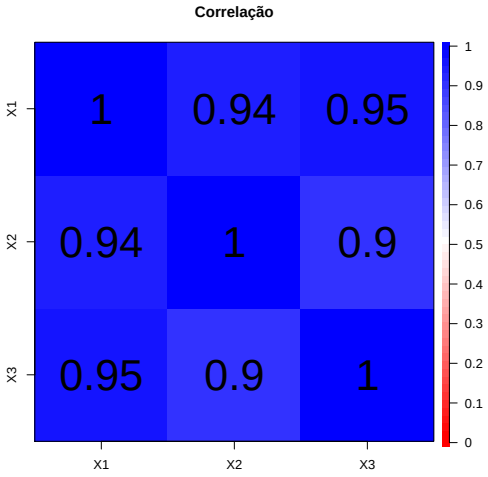
(f) Piloto 6

Figura 4.4 – Estudos de casos: *boxplot* das classificações

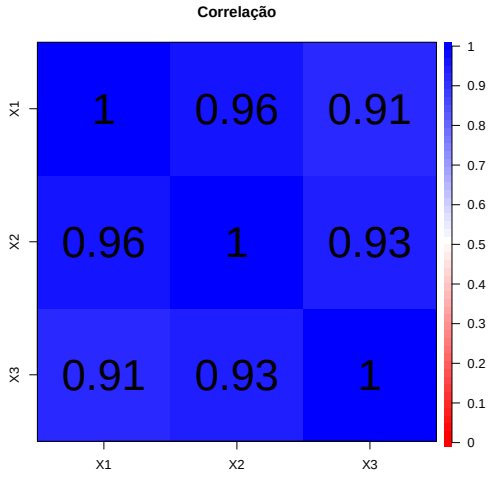


(a) Piloto 1

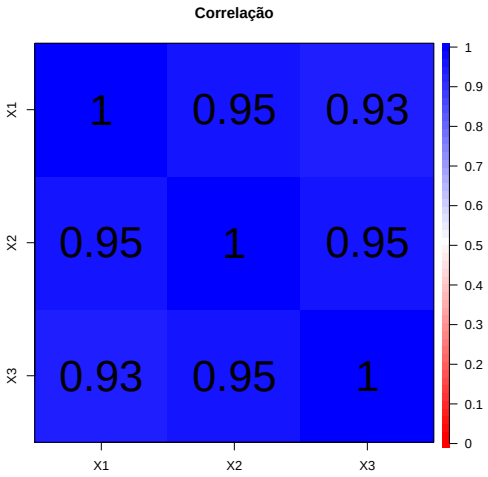
(b) Piloto 2



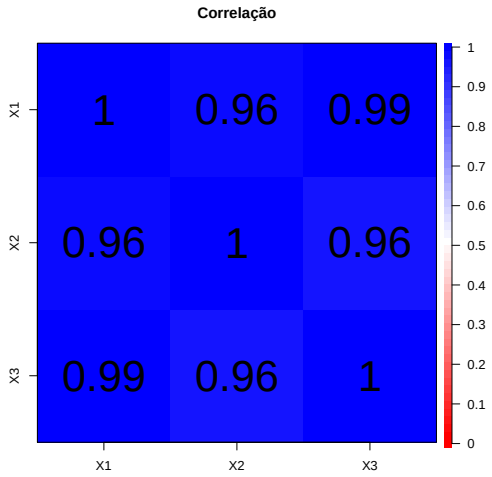
(c) Piloto 3



(d) Piloto 4

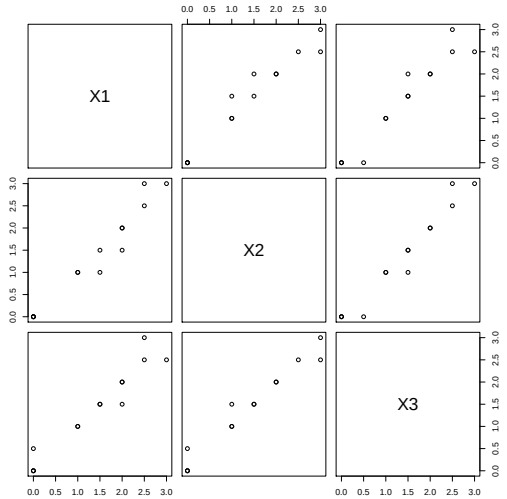


(e) Piloto 5

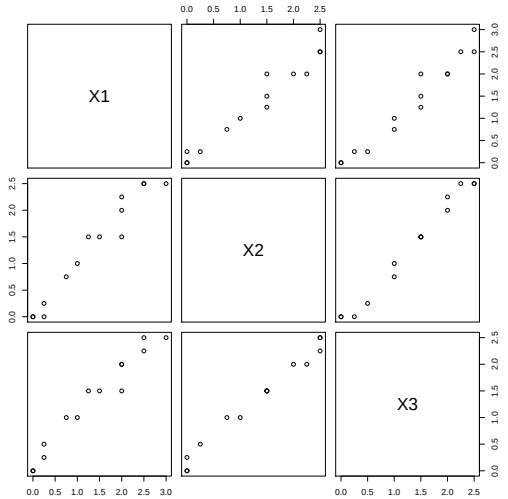


(f) Piloto 6

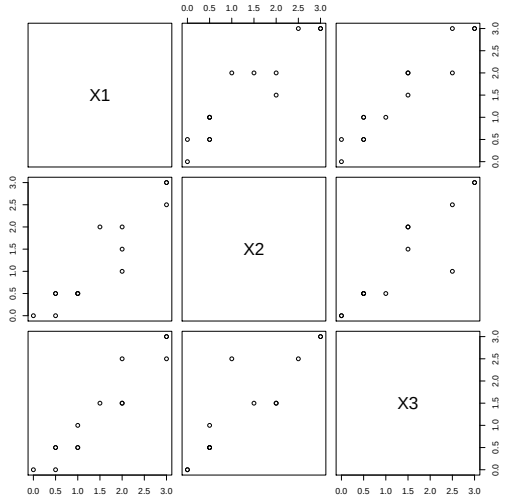
Figura 4.5 – Estudos pilotos: correlação entre as execuções



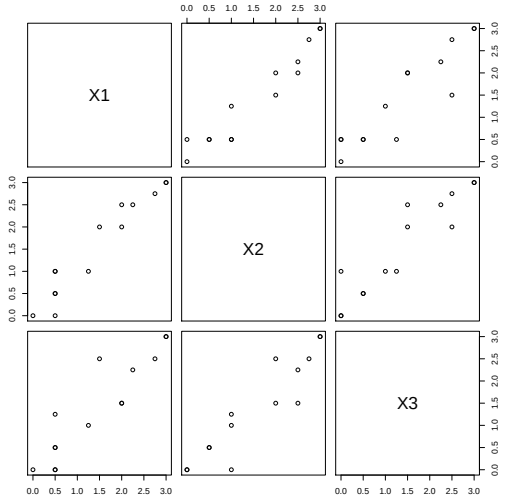
(a) Piloto 1



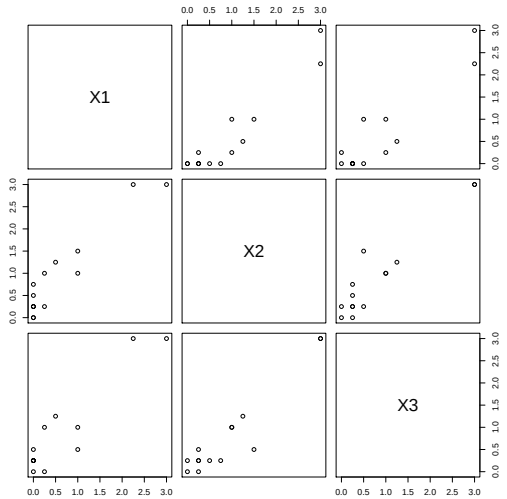
(b) Piloto 2



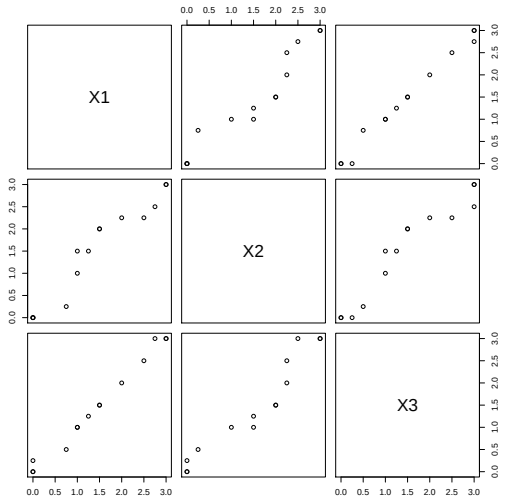
(c) Piloto 3



(d) Piloto 4

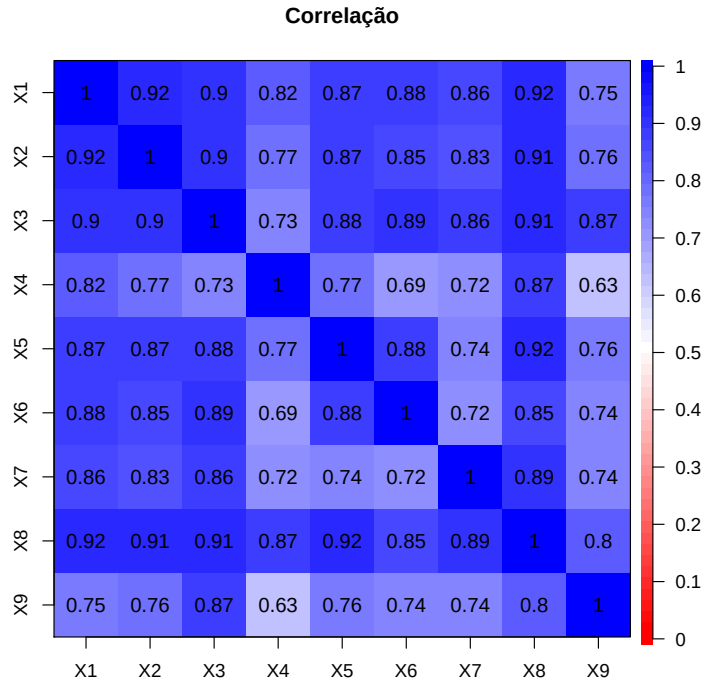


(e) Piloto 5

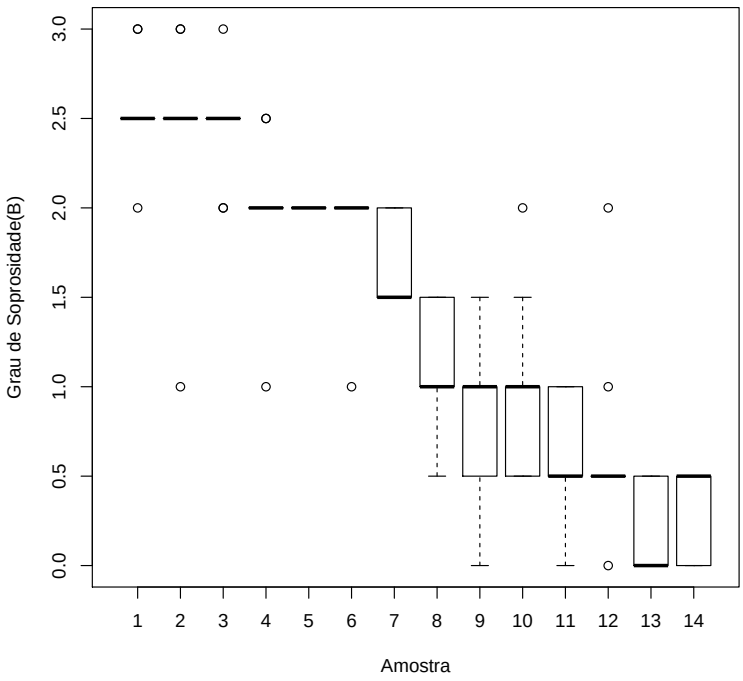


(f) Piloto 6

Figura 4.6 – Estudos pilotos: dispersão entre as execuções

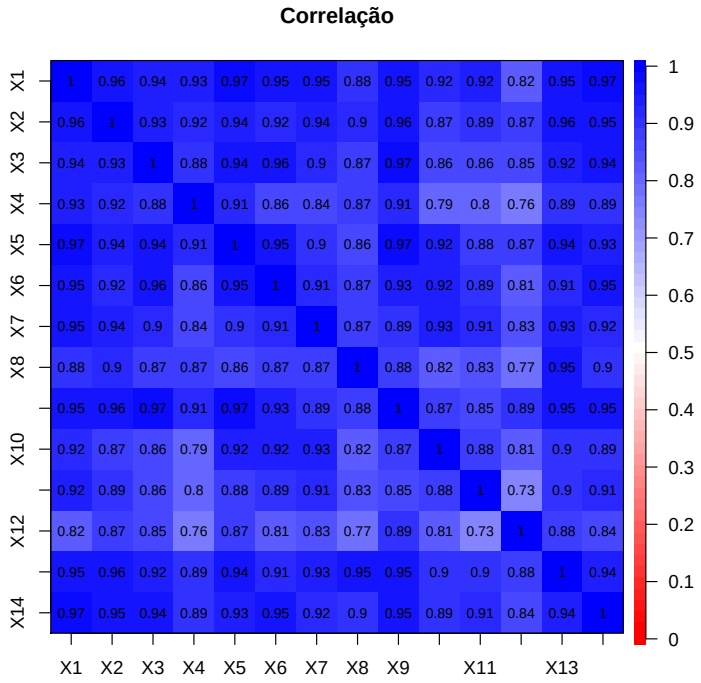


(a) Correlação

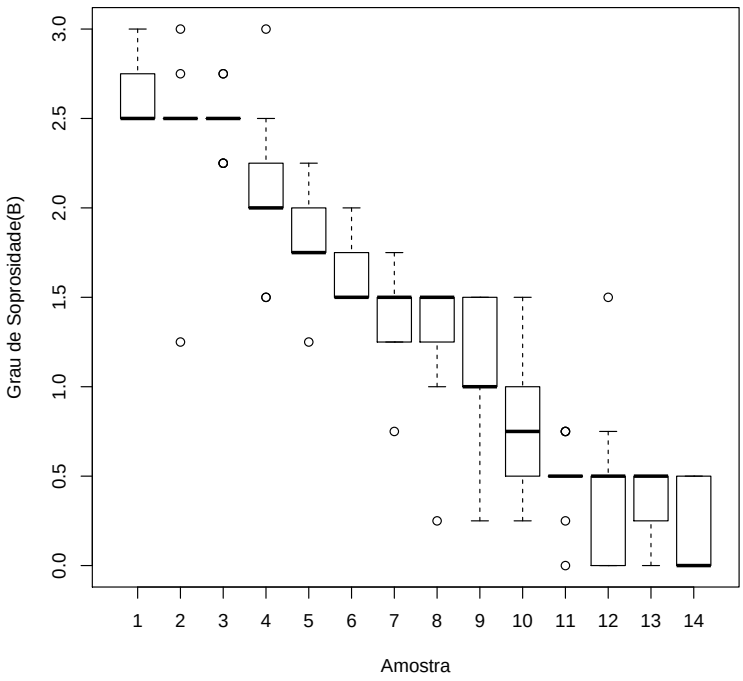


(b) *Boxplot*

Figura 4.7 – Experimento 1: Voz sintética com três âncoras

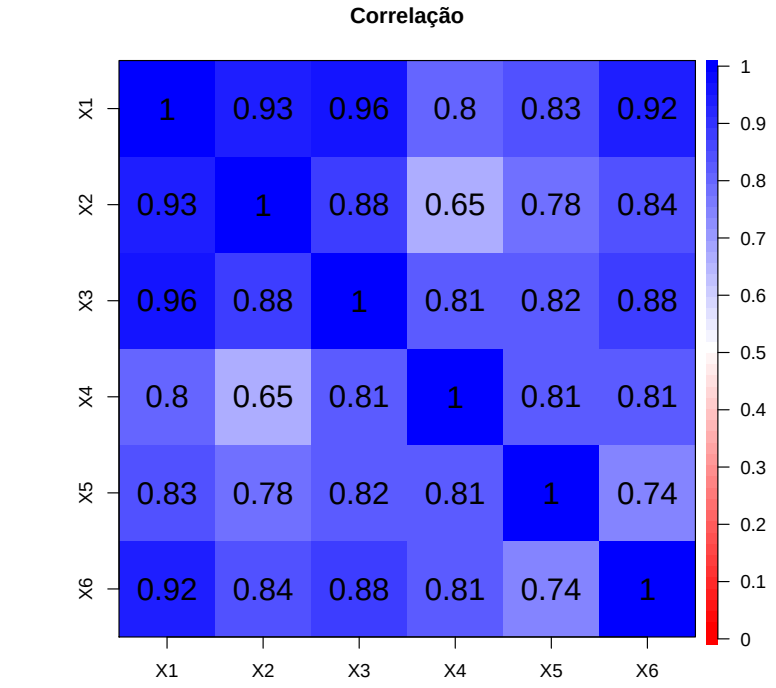


(a) Correlação

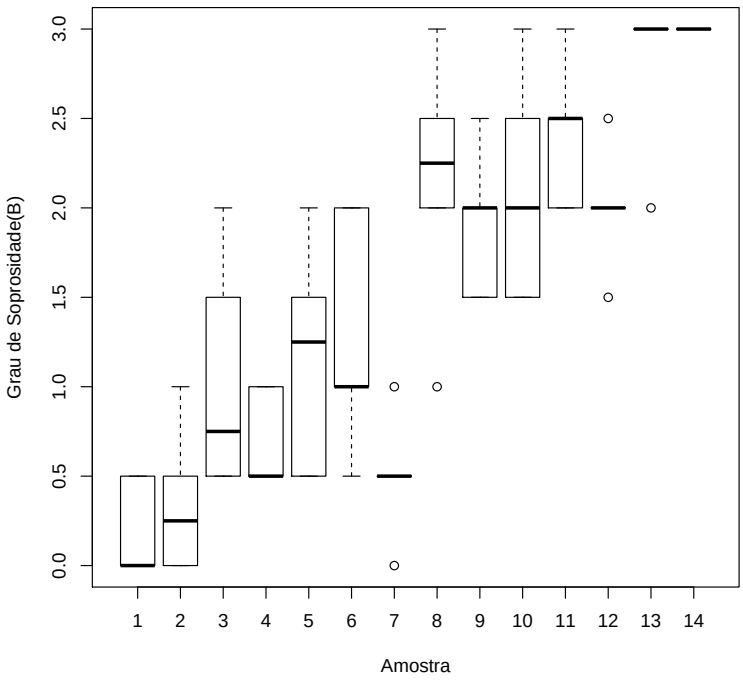


(b) *Boxplot*

Figura 4.8 – Experimento 2: Voz sintética com sete âncoras

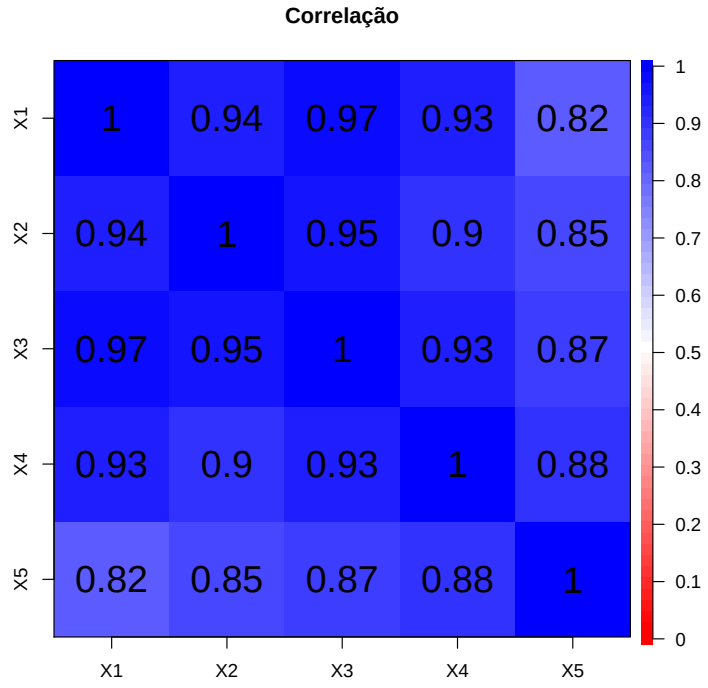


(a) Correlação

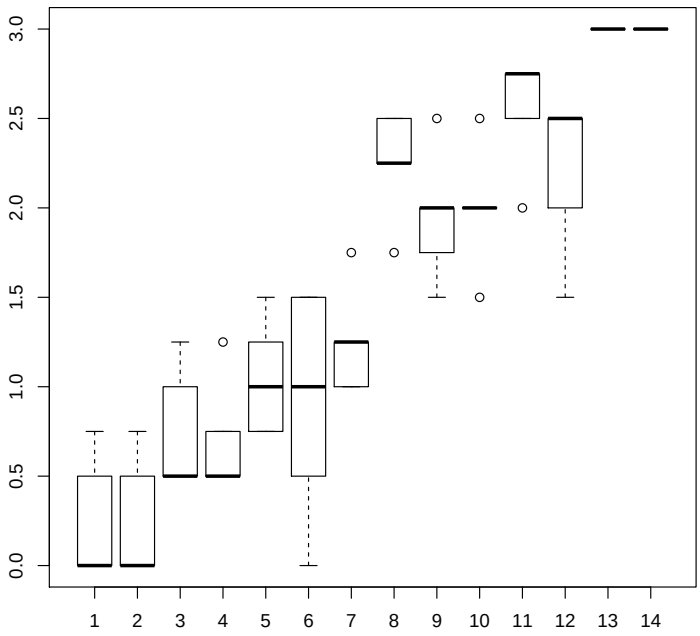


(b) *Boxplot*

Figura 4.9 – Experimento 3: Voz natural com três âncoras



(a) Correlação



(b) *Boxplot*

Figura 4.10 – Experimento 4: Voz natural com sete âncoras

5 Conclusão

5.1 Considerações finais

Este trabalho investigou a soprosidade vocal em duas frentes distintas: das medidas objetivas e da avaliação perceptivo-auditiva.

Em relação às medidas objetivas, uma plataforma para avaliação dos correlatos acústicos da soprosidade foi desenvolvida, para a realização da avaliação de desempenho das medidas de soprosidade em vogais sintéticas com valores de relação sinal-ruído, *jitter* e *shimmer* controlados, com o intuito de caracterizar a robustez das medidas às perturbações que costumam ocorrer na voz disfônica.

Dos testes na plataforma, a relação sinal-ruído espectrográfica (S^2NR) emergiu como a medida acústica mais robusta às perturbações, principalmente em faixas com relação sinal-ruído mais baixa (inferiores a 30 dB).

Uma vez conhecidas as características dos métodos de estimativa numérica da soprosidade, tem-se a perspectiva da dimensão física do fenômeno.

A caracterização da dimensão perceptiva foi desenvolvida com a estimativa das escalas de diferenças de percepção e na estimativa das funções psicofísicas da soprosidade, relacionando uma grandeza física (relação sinal-ruído em amostras de voz sintética) com a mínima capacidade de percepção de variação da mesma. Desta forma, foi possível estimar o número de classes alocáveis na dimensão perceptiva.

O trabalho foi complementado pela proposta de um método de avaliação perceptivo-auditiva da voz soprosa, baseado em comparações em uma árvore binária, visando uma maior confiabilidade nas avaliações inter e intra-julgadores.

E finalmente, pode-se destacar que a metodologia desenvolvida para a caracterização da dimensão perceptiva não restringe-se à soprosidade vocal, podendo ser aplicada às outras dimensões comumente utilizadas na avaliação da voz disfônica.

5.2 Limitações

Nos Capítulos 3 e 4 alguns experimentos foram executados com um número baixo de informantes, e foram tratados como estudos de casos, limitando a validade estatística dos resultados obtidos.

No desenvolvimento deste trabalho, utilizaram-se majoritariamente vozes femininas, tanto em amostras geradas por síntese quanto nas amostras de voz natural e nas âncoras

da avaliação comparativa. Essa escolha foi feita pelo fato da soprosidade ser um traço característico mais comum entres a mulheres do que em homens (KLATT; KLATT, 1990). No entanto, o efeito da variação da freqüência fundamental não deve ser desprezado nos estudos de percepção.

O conjunto de amostras de voz natural utilizado neste trabalho é limitado. Mesmo que sejam vozes com característica predominante de soprosidade, a ocorrência de perturbações simultâneas não pode ser evitada, podendo influenciar nos estudos de percepção e classificação.

A utilização de vogais sustentadas nas análises e experimentos de percepção é outra limitação que deve ser notada, já que não explorou-se em experimentos o uso de amostras de fala corrente.

5.3 Propostas de continuidade do trabalho

Algumas propostas para continuidade deste trabalho podem ser elencadas:

- Aumento do número de informantes nos Experimentos 3 e 4 do Capítulo 4, de Busca em árvore binária com âncoras de voz natural;
- Aumento do número de informantes nos experimentos de parametrização da função psicométrica da soprosidade do Capítulo 3. Além disso, deve se explorar os limiares com freqüência fundamental típica de falantes masculinos ($f_0 = 120Hz$).
- Desenvolvimento de um aplicativo para dispositivos móveis que implemente o método de busca binária, permitindo a gravação de pacientes ou a importação de amostras externas, visando a difusão da técnica entre os profissionais.
- Estudo do efeito de perturbações simultâneas com a soprosidade, como *jitter* e *shimmer*, com a elaboração de experimentos com a *Maximum Likelihood Conjoint Measurement* (KNOBLAUCH; MALONEY, 2012)
- Aplicação da metodologia para determinação de escalas e parametrização de funções psicométricas para outras dimensões perceptivas da voz disfônica.

Referências

AWAN, S. N.; LAWSON, L. L. The effect of anchor modality on the reliability of vocal severity ratings. *Journal of Voice*, v. 23, n. 3, p. 341 – 352, 2009. ISSN 0892-1997. Disponível em: <http://www.sciencedirect.com/science/article/pii/S089219970700135X>. Citado na página 74.

BACHEM, A. Absolute pitch. *The Journal of the Acoustical Society of America*, ASA, v. 27, n. 6, p. 1180–1185, 1955. Citado na página 74.

BARSTIES, B.; BEERS, M.; CATE, L. T.; BALLEGOOIJEN, K. V.; BRAAM, L.; GROOT, M. D.; KANT, M. V. D.; KRUITWAGEN, C.; MARYN, Y. The effect of visual feedback and training in auditory-perceptual judgment of voice quality. *Logopedics Phoniatrics Vocology*, Taylor & Francis, v. 42, n. 1, p. 1–8, 2017. Citado na página 74.

BAZEN, A. M. Systematic methods for the computation of the directional fields and singular points of fingerprints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, 2002. Citado 3 vezes nas páginas 21, 25 e 26.

BIELAMOWICZ, S.; KREIMAN, J.; GERRATT, B. R.; DAUER, M. S.; BERKE, G. S. Comparison of voice analysis systems for perturbation measurement. *Journal of Speech and Hearing Research*, ASHA, v. 39, n. 1, p. 126, 1996. Citado na página 19.

BRINCA, L.; BATISTA, A. P.; TAVARES, A. I.; PINTO, P. N.; ARAúJO, L. The effect of anchors and training on the reliability of voice quality ratings for different types of speech stimuli. *Journal of Voice*, v. 29, n. 6, p. 776.e7 – 776.e14, 2015. ISSN 0892-1997. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0892199715000089>. Citado na página 74.

CHEN, G.; KREIMAN, J.; GERRATT, B. R.; NEUBAUER, J.; SHUE, Y.-L.; ALWAN, A. Development of a glottal area index that integrates glottal gap size and open quotient. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 133, n. 3, p. 1656–1666, 2013. Citado na página 19.

COX, N. B.; ITO, M. R.; MORRISON, M. D. Data labeling and sampling effects in harmonics-to-noise ratios. *The Journal of the Acoustical Society of America*, ASA, v. 85, n. 5, p. 2165–2178, 1989. Disponível em: <http://link.aip.org/link/?JAS/85/2165/1>. Citado 3 vezes nas páginas 20, 36 e 37.

DAUGMAN, J. G. 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*, OSA, v. 2, n. 7, p. 1160, 1985. Disponível em: <http://josaa.osa.org/abstract.cfm?URI=josaa-2-7-1160>. Citado na página 27.

DEJONCKERE, P.; WIENEKE, G. Spectral, cepstral and aperiodicity characteristics of pathological voices before and after phonosurgical treatment! *Clinical linguistics & phonetics*, Informa UK Ltd UK, v. 8, n. 2, p. 161–169, 1994. Citado 2 vezes nas páginas 15 e 19.

DING, H.; SOON, I. Y.; KOH, S. N.; YEO, C. K. A spectral filtering method based on hybrid wiener filters for speech enhancement. *Speech Communication*, Elsevier, v. 51, n. 3, p. 259–267, 2009. Citado na página 21.

DIXON, W. J.; MOOD, A. M. A method for obtaining and analyzing sensitivity data. *Journal of the American Statistical Association*, Taylor & Francis, v. 43, n. 241, p. 109–126, 1948. Citado na página 66.

EADIE, T. L.; KAPSNER-SMITH, M. The effect of listener experience and anchors on judgments of dysphonia. *Journal of Speech, Language, and Hearing Research*, ASHA, v. 54, n. 2, p. 430–447, 2011. Citado na página 74.

EATON, J. W.; BATEMAN, D.; HAUBERG, S.; WEHBRING, R. *GNU Octave version 3.8.1 manual: a high-level interactive language for numerical computations*. CreateSpace Independent Publishing Platform, 2014. ISBN 1441413006. Disponível em: <http://www.gnu.org/software/octave/doc/interpreter>. Citado 4 vezes nas páginas 52, 68, 76 e 81.

ESKENAZI, L.; CHILDERS, D. G.; HICKS, D. M. Acoustic Correlates of Vocal Quality. *J. Speech Hear. Res.*, v. 33, n. 2, p. 298–306, 1990. Citado 2 vezes nas páginas 15 e 19.

EZZAT, T.; BOUVRIE, J. V.; POGGIO, T. Spectro-temporal analysis of speech using 2-d Gabor filters. In: *INTERSPEECH*. [S.l.: s.n.], 2007. p. 506–509. Citado 2 vezes nas páginas 21 e 27.

FANT, G. Glottal source and excitation analysis. *Speech Transmission Laboratory - Quaterly Progress and Status Report*, v. 1, p. 85–107, 1979. Citado na página 29.

FEIJOO, S.; HERNANDEZ, C. Short-Term Stability Measures for the Evaluation of Vocal Quality. *J. Speech Hear. Res.*, v. 33, n. 2, p. 324–334, 1990. Citado 2 vezes nas páginas 15 e 19.

HILLENBRAND, J. A Methodological Study of Perturbation and Additive Noise in Syntheti cally Generated Voice Signals. *J. Speech Hear. Res.*, v. 30, n. 4, p. 448–461, 1987. Citado 2 vezes nas páginas 19 e 20.

HILLENBRAND, J.; CLEVELAND, R.; ERICKSON, R. Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research*, ASHA, v. 37, n. 4, p. 769, 1994. ISSN 1092-4388. Citado 2 vezes nas páginas 18 e 33.

HILLENBRAND, J.; HOUDE, R. Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of speech and hearing research*, ASHA, v. 39, n. 2, p. 311, 1996. ISSN 1092-4388. Citado na página 33.

HIRANO, M. *Clinical Examination of Voice*. [S.l.]: Springer Verlag, 1981. Citado 2 vezes nas páginas 14 e 74.

HIRANO, M.; HIBI, S.; YOSHIDA, T.; HIRADE, Y. i.; KASUYA, H.; KIKUCHI, Y. Acoustic analysis of pathological voice: Some results of clinical appl ication. *Acta Oto-laryngol.*, v. 105, n. 5-6, p. 432–438, 1988. Citado 3 vezes nas páginas 15, 18 e 19.

HIRAOKA, N.; KITAZOE, Y.; UETA, H.; KA, S. T.; TANABE, M. Harmonic-intensity analysis of normal and hoarse voices. *J. Acoust. Soc. Am.*, ASA, v. 76, n. 6, p. 1648–1651, 1984. Citado na página 19.

HONG, L.; WAN, Y.; JAIN, A. Fingerprint image enhancement: algorithm and performance evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 20, n. 8, p. 777–789, Aug. 1998. Citado 2 vezes nas páginas 21 e 27.

HORII, Y. Fundamental Frequency Perturbation Observed in Sustained Phonation. *J. Speech Hear. Res.*, v. 22, n. 1, p. 5–19, 1979. Citado na página 29.

HORN, T. Image processing of speech with auditory magnitude spectrograms. *Acustica*, Stuttgart [Germany]: S. Hirzel, 1951-1995., v. 84, n. 1, p. 175–177, 1998. Citado na página 21.

JAIN, A.; FARROKHNIA, F. Unsupervised texture segmentation using Gabor filters. *Pattern recognition*, Elsevier, v. 24, n. 12, p. 1167–1186, 1991. ISSN 0031-3203. Citado na página 27.

JELLYMAN, K. A.; EVANS, N. W.; LIU, W. M.; MASON, J. S. Towards a new image-based spectrogram segmentation speech coder optimised for intelligibility. In: *Advances in Multimedia Modeling*. [S.l.]: Springer, 2009. p. 63–73. Citado na página 21.

KAERNBACH, C. Simple adaptive testing with the weighted up-down method. *Attention, Perception, & Psychophysics*, Springer, v. 49, n. 3, p. 227–229, 1991. Citado na página 67.

KANE, M.; WELLEN, C. Acoustical measurements and clinical judgments of vocal quality in children with vocal nodules. *Folia Phoniatrica et Logopaedica*, Karger Publishers, v. 37, n. 2, p. 53–57, 1985. Citado 2 vezes nas páginas 15 e 19.

KASUYA, H.; OGAWA, S.; KIKUCHI, Y. An adaptive comb filtering method as applied to acoustic analyses of pathological voice. *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '86.*, v. 11, p. 669–672, Apr 1986. Citado 3 vezes nas páginas 19, 20 e 41.

KASUYA, H.; OGAWA, S.; MASHIMA, K.; EBIHARA, S. Normalized noise energy as an acoustic measure to evaluate pathologic voice. *The Journal of the Acoustical Society of America*, ASA, v. 80, n. 5, p. 1329–1334, 1986. Disponível em: <http://link.aip.org/link/?JAS/80/1329/1>. Citado 2 vezes nas páginas 19 e 20.

KINGDOM, F.; PRINS, N. *Psychophysics: a practical introduction*. [S.l.]: Academic Press London, 2010. Citado 5 vezes nas páginas 17, 47, 65, 67 e 68.

KLATT, D. H. Software for a cascade/parallel formant synthesizer. *the Journal of the Acoustical Society of America*, Acoustical Society of America, v. 67, n. 3, p. 971–995, 1980. Citado na página 30.

KLATT, D. H.; KLATT, L. C. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of Acoustical Society of America*, v. 87, n. 2, p. 820–857, Feb 1990. Citado 2 vezes nas páginas 19 e 98.

KLEINER, M.; BRAINARD, D.; PELLI, D.; INGLING, A.; MURRAY, R.; BROUSSARD, C. et al. What's new in psychtoolbox-3. *Perception*, Alezzo, v. 36, n. 14, p. 1, 2007. Citado 4 vezes nas páginas 52, 68, 76 e 81.

KNOBLAUCH, K.; MALONEY, L. T. *Modeling psychophysical data in R*. [S.l.]: Springer Science & Business Media, 2012. Citado 4 vezes nas páginas 51, 52, 73 e 98.

KNOBLAUCH, K.; MALONEY, L. T. et al. Mlds: Maximum likelihood difference scaling in r. *Journal of Statistical Software*, v. 25, n. 2, p. 1–26, 2008. Citado 3 vezes nas páginas 17, 47 e 52.

KOVESI, P. D. *MATLAB and Octave Functions for Computer Vision and Image Processing*. 2005. School of Computer Science & Software Engineering, The University of Western Australia. Disponível em: <http://www.csse.uwa.edu.au/∼pk/research/matlabfns/>, acessado em Março de 2014. Citado na página 21.

KREIMAN, J.; GERRATT, B.; ITO, M. When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America*, v. 122, p. 2354, 2007. Citado na página 74.

KREIMAN, J.; GERRATT, B. R. Validity of rating scale measures of voice quality. *The Journal of the Acoustical Society of America*, ASA, v. 104, n. 3, p. 1598–1608, 1998. Disponível em: <http://link.aip.org/link/?JAS/104/1598/1>. Citado 4 vezes nas páginas 14, 15, 17 e 74.

KREIMAN, J.; GERRATT, B. R. Measuring voice quality. In: KENT, R.; BALL, M. J. (Ed.). *Handbook of Voice Quality Measurement*. [S.l.]: Singular, 1999. Citado na página 16.

KREIMAN, J.; GERRATT, B. R.; BERKE, G. S. The multidimensional nature of pathologic vocal quality. *The Journal of the Acoustical Society of America*, ASA, v. 96, n. 3, p. 1291–1302, 1994. Citado na página 89.

KREIMAN, J.; GERRATT, B. R.; KEMPSTER, G. B.; ERMAN, A.; BERKE, G. S. Perceptual Evaluation of Voice Quality: Review, Tutorial, and a Framework for Future Research. *Journal of Speech, Language, and Hearing Research*, v. 36, n. 1, p. 21–40, 1993. Disponível em: <http://jslhr.asha.org/cgi/content/abstract/36/1/21>. Citado 2 vezes nas páginas 15 e 17.

KROM, G. de. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech and Hearing Research*, v. 36, p. 254–266, 1993. Citado 4 vezes nas páginas 19, 20, 35 e 41.

LAVER, J.; HILLER, S.; BECK, J. M. Acoustic waveform perturbations and voice disorders. *Journal of Voice*, Elsevier, v. 6, n. 2, p. 115–126, 1992. Citado na página 19.

LAVER, J.; WIRZ, S.; MACKENZIE, J.; HILLER, S. A perceptual protocol for the analysis of vocal profiles. *Work in Progress*, Edinburgh University: Department of Linguistics, v. 14, p. 139–155, 1981. Citado 2 vezes nas páginas 14 e 74.

LIEBERMAN, P. Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 35, n. 3, p. 344–353, 1963. Citado 2 vezes nas páginas 15 e 19.

MALONEY, L. T.; YANG, J. N. Maximum likelihood difference scaling. *Journal of Vision*, The Association for Research in Vision and Ophthalmology, v. 3, n. 8, p. 5–5, 2003. Citado na página 48.

MARKEL, J.; JR, A. H. G. *Linear Prediction of Speech*. [S.l.]: Springer-Verlag, 1976. Citado 2 vezes nas páginas 29 e 34.

MARTIN, D.; FITCH, J.; WOLFE, V. Pathologic voice type and the acoustic prediction of severity. *Journal of Speech, Language, and Hearing Research*, ASHA, v. 38, n. 4, p. 765–771, 1995. Citado 2 vezes nas páginas 15 e 19.

MARYN, Y.; ROY, N.; BODT, M. D.; CAUWENBERGE, P. V.; CORTHALS, P. Acoustic measurement of overall voice quality: A meta-analysis. *The Journal of the Acoustical Society of America*, v. 126, p. 2619, 2009. Citado 5 vezes nas páginas 18, 32, 33, 35 e 89.

MURPHY, P. J. Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis. *The Journal of the Acoustical Society of America*, ASA, v. 105, n. 5, p. 2866–2881, 1999. Disponível em: <http://link.aip.org/link/?JAS/105/2866/1>. Citado 4 vezes nas páginas 19, 20, 32 e 36.

MURPHY, P. J. Spectral characterization of jitter, shimmer, and additive noise in synthetically generated voice signals. *The Journal of the Acoustical Society of America*, ASA, v. 107, n. 2, p. 978–988, 2000. Disponível em: <http://link.aip.org/link/?JAS/107/978/1>. Citado na página 20.

MURPHY, P. J.; AKANDE, O. O. Noise estimation in voice signals using short-term cepstral analysis. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 121, n. 3, p. 1679–1690, 2007. Citado na página 20.

MUTA, H.; BAER, T.; WAGATSUMA, K.; MURAOKA, T.; FUKUDA, H. A pitch-synchronous analysis of hoarseness in running speech. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 84, n. 4, p. 1292–1301, 1988. Citado 3 vezes nas páginas 15, 19 e 20.

NIXON, M. K.; AGUADO, A. S. *Feature Extraction and Image Processing*. [S.l.]: Newnes, 2002. Citado na página 25.

ORLIKOFF, R. F. The perceived role of voice perception in clinical practice. *Phonoscope*, v. 2, p. 87–106, 1999. Citado 2 vezes nas páginas 14 e 16.

PARSA, V.; JAMIESON, D. Acoustic discrimination of pathological voice: sustained vowels versus continuous speech. *Journal of speech, language, and hearing research*, ASHA, v. 44, n. 2, p. 327, 2001. ISSN 1092-4388. Citado na página 32.

PATEL, S.; SHRIVASTAV, R.; EDDINS, D. A. Developing a single comparison stimulus for matching breathy voice quality. *Journal of Speech, Language, and Hearing Research*, ASHA, v. 55, n. 2, p. 639–647, 2012. Citado na página 30.

PENTLAND, A. Maximum likelihood estimation: The best pest. *Attention, Perception, & Psychophysics*, Springer, v. 28, n. 4, p. 377–379, 1980. Citado na página 67.

PROSEK, R. A.; MONTGOMERY, A. A.; WALDEN, B. E.; HAWKINS, D. B. An evaluation of residue features as correlates of voice disorders. *Journal of communication disorders*, Elsevier, v. 20, n. 2, p. 105–117, 1987. Citado na página 34.

QI, Y. Time normalization in voice analysis. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 92, n. 5, p. 2569–2576, 1992. Citado 2 vezes nas páginas 19 e 20.

QI, Y.; WEINBERG, B.; BI, N.; HESS, W. J. Minimizing the effect of period determination on the computation of amplitude perturbation in voice. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 97, n. 4, p. 2525–2532, 1995. Citado 3 vezes nas páginas 19, 20 e 36.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>. Citado 2 vezes nas páginas 52 e 81.

REVELLE, W. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois, 2018. R package version 1.8.4. Disponível em: <https://CRAN.R-project.org/package=psych>. Citado na página 81.

ROBINSON, H. F. Measuring voice quality. In: BEECH, J.; HARDING, L.; HILTON-JONES, D. (Ed.). *Assessment in Speech and Language Therapy*. [S.l.]: Rooutledge, 1993. Citado na página 14.

SANTOS, P. C. M. dos; VIEIRA, M. N.; SANSÃO, J. P. H.; GAMA, A. C. C. Effect of auditory-perceptual training with natural voice anchors on vocal quality evaluation. *Journal of Voice*, Elsevier, 2018. Citado na página 57.

SCHOENTGEN, J. Quantitative evaluation of the discrimination performance of acoustic features in detecting laryngeal pathology. *Speech Communication*, Elsevier, v. 1, n. 3, p. 269–282, 1982. Citado 2 vezes nas páginas 15 e 19.

SEDGEWICK, R.; WAYNE, K. *Algorithms*. [S.l.]: Addison-Wesley Professional, 2011. Citado 4 vezes nas páginas 75, 78, 79 e 81.

SHUE, Y.-L. *VoiceSauce - A program for voice analysis*. 2010. Disponível em: <http://www.ee.ucla.edu/~spapl/voicesauce/>, acessado em Julho de 2013. Citado na página 35.

SOON, Y.; KOH, S. N. Speech enhancement using 2-d fourier transform. *Speech and Audio Processing, IEEE Transactions on*, IEEE, v. 11, n. 6, p. 717–724, 2003. Citado na página 21.

TITZE, I. R. *Workshop on Acoustic Voice Analysis*. [S.l.]: National Center for Voice and Speech, 1995. Citado na página 14.

TITZE, I. R.; WINHOLTZ, W. S. Effect of Microphone Type and Placement on Voice Perturbation Measurements. *Journal of Speech, Language, and Hearing Research*, v. 36, n. 6, p. 1177–1190, 1993. Disponível em: <http://jslhr.asha.org/cgi/content/abstract/36/6/1177>. Citado na página 20.

VIEIRA, M. N. *Automated Measures of Dysphonias and the Phonatory Effects of Asymmetries in the Posterior Larynx*. Tese (Doutorado) — University of Edinburgh, 1997. Citado 2 vezes nas páginas 32 e 34.

VIEIRA, M. N.; MCINNES, F. R.; JACK, M. A. On the influence of laryngeal pathologies on acoustic and electroglottographic jitter measures. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 111, n. 2, p. 1045–1055, 2002. Citado 2 vezes nas páginas 19 e 34.

VIEIRA, M. N.; SANSÃO, J. P. H.; YEHIA, H. C. Measurement of signal-to-noise ratio in dysphonic voices by image processing of spectrograms. *Speech Communication*, North-Holland, v. 61, p. 17–32, 2014. Citado 15 vezes nas páginas 16, 19, 22, 24, 31, 36, 37, 38, 40, 41, 42, 43, 44, 45 e 59.

WATSON, A. B.; PELLI, D. G. Quest: A bayesian adaptive psychometric method. *Perception & psychophysics*, Springer, v. 33, n. 2, p. 113–120, 1983. Citado na página 67.

WETHERILL, G.; LEVITT, H. Sequential estimation of points on a psychometric function. *British Journal of Mathematical and Statistical Psychology*, Wiley Online Library, v. 18, n. 1, p. 1–10, 1965. Citado na página 66.

WOLFE, V. I.; STEINFATT, T. M. Prediction of vocal severity within and across voice types. *Journal of Speech, Language, and Hearing Research*, ASHA, v. 30, n. 2, p. 230–240, 1987. Citado 2 vezes nas páginas 17 e 20.

YANAGIHARA, N. Significance of harmonic changes and noise components in hoarseness. *Journal of Speech, Language, and Hearing Research*, v. 10, n. 3, p. 531–541, Sep 1967. Citado 2 vezes nas páginas 17 e 20.

YIU, E. M.-L.; CHAN, K. M.; MOK, R. S.-M. Reliability and confidence in using a paired comparison paradigm in perceptual voice quality evaluation. *Clinical linguistics & phonetics*, Taylor & Francis, v. 21, n. 2, p. 129–145, 2007. Citado na página 74.

YUMOTO, E.; GOULD, W. J.; BAER, T. Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of Acoustical Society of America*, v. 71, n. 6, p. 1544–1549, Jun 1982. Citado 2 vezes nas páginas 19 e 20.

ZHANG, Y.; JIANG, J.; WALLACE, S.; ZHOU, L. Comparison of nonlinear dynamic methods and perturbation methods for voice analysis. *The Journal of the Acoustical Society of America*, v. 118, p. 2551, 2005. Citado 2 vezes nas páginas 15 e 19.

ZWICKER, E.; FASTL, H. *Psychoacoustics: Facts and Models*. [S.l.]: Springer-Verlag, 1999. Citado 2 vezes nas páginas 16 e 17.