Faculdade de Engenharia da Universidade do Porto

Ambiente de análise robusta dos principais parâmetros qualitativos da voz

José Manuel dos Santos Lopes

Dissertação realizada no âmbito do
Mestrado Integrado em Engenharia Electrotécnica e de Computadores
Major em Telecomunicações

Orientador: Prof. Dr. Aníbal Ferreira
Co-orientador: Eng. Joaquim Matos

Julho 2008

Resumo

No âmbito desta dissertação foi desenvolvida, em colaboração com a empresa SEEGNAL Research, a aplicação SEEGNAL VoiceStudio que é uma aplicação de análise e diagnóstico da voz. A área em que exigiu maior esforço de investigação e desenvolvimento foi a análise dos principais parâmetros da voz entre os quais *pitch, jitter, shimmer,* relação harmónicos-ruído (HNR – *Harmonic-to-Noise Ratio*).

O VoiceStudio revela-se uma ferramenta importante para apoio nas áreas da terapia da fala, porque constitui uma ferramenta objectiva e não invasiva para análise e diagnóstico da voz.

A análise e o diagnóstico da voz são normalmente efectuados por avaliação perceptiva em que a subjectividade está inerente. Aplicações de análise e diagnóstico da voz trazem a vantagem de permitirem uma avaliação objectiva e fundamentada das características da voz.

Existem várias aplicações de análise e diagnóstico da voz. No entanto apresentam algumas desvantagens, tais como, dificuldade na utilização, dificuldade no cálculo dos parâmetros da voz e pobre representação e interacção gráfica.

No âmbito deste trabalho desenvolveu-se um ambiente que permite rapidamente efectuar uma análise acústica sobre um ficheiro de áudio ou uma captação de áudio, facilitando assim o diagnóstico que indica a existência ou não existência de alguma patologia na voz, tendo por base os principais parâmetros qualitativos da voz.

A robustez na medição dos principais parâmetros qualitativos da voz foi testado no VoiceStudio bem como nos ambientes concorrentes, recorrendo a ficheiros de voz sintéticos com valores pré-definidos de *jitter, shimmer* e relação harmónicos-ruído.

Abstract

In the scope of this thesis, the application SEEGNAL VoiceStudio was developed in cooperation with SEEGNAL Research Company. VoiceStudio is an application of analysis and diagnosis of the voice. The area requiring a higher research and development effort was the analysis of the main parameters of the voice (fundamental frequency, jitter, shimmer, harmonic-to-noise ratio (HNR)).

VoiceStudio reveals to be an important tool for support in speech therapy area, because it is a non-invasive and objective method for analysis and diagnosis of the voice.

The analysis and diagnosis of the voice traditionally has been done using perceptual evaluation that is a subjective method. Software based analysis and diagnosis of the voice allow an objective evaluation based on physical and therefore objective characteristics of the voice.

There are several software packages allowing the analysis and diagnosis of the voice. However they present some disadvantages, such as, difficulty in the use, difficulties in the calculation of the parameters of the voice and poor graphical user interface and interaction.

In the scope of this thesis an environment was developed that is easy to use and it quickly allows an effective analysis on an audio file or an audio record, and makes it easy to identify the existence of some pathology in the voice based in the main qualitative parameters of the voice, as well.

The accuracy and robustness of the VoiceStudio computation of the main qualitative parameters of the voice was evaluated and compared to other competing environments using synthetic files of voice with predefined values of jitter, shimmer and harmonic-to-noise ratio.

Agradecimentos

O desenvolvimento deste trabalho não seria possível sem o apoio e disponibilidade do Prof. Aníbal Ferreira e da ajuda fornecida pela equipa SEEGNAL Research: Joaquim Matos e Filipe Abreu.

A eles o meu profundo agradecimento.

Índice

Lista de figuras

Lista de Tabelas

Abreviaturas e símbolos

ACC	*Audio Communication Coder*
API	*Application Programming Interface*
APQ	*Amplitude Perturbation Quotient*
ASIO	*Audio Stream Input/Output*
CSL	*Computer Speech Lab*
EGG	Electroglotografia
FFT	*Fast Fourier Transform*
HNR	*Harmonic-to-Noise Ratio*
GHNR	*Glottal Harmonic-to-Noise Ratio*
NNE	*Normalized Noise Energy*
PPQ	*Period Perturbation Quotient*
RAP	*Relative Average Perturbation*
SFS	*Speech Filing System*
WAV	forma curta de *WAVEform audio format*

Capítulo 1 Introdução

1.1 O que é a Voz

A voz é o som que resulta do fluxo de ar que é expulso dos pulmões por acção do diafragma, com uma determinada pressão e velocidade, passando pelas pregas vocais (com diferentes padrões de adução[1] e abdução[2]), sendo modulado pelas propriedades de reflexão e configuração do tracto vocal (que inclui a boca, lábios e língua).

1.2 Perturbações na voz

O conceito de perturbação na voz (disfonia) está ligado à qualidade da voz e à sua adequação.

A voz produzida possui características próprias que variam de acordo com o sexo, a pessoa e com a faixa etária, além de reflectir o estado e comportamentos laríngeos, caracterizando o que se chama de qualidade vocal.

A qualidade da voz identifica-se através de preceitos fisiológicos, perceptivos e acústicos. Para uma voz normal são aceites as variações de qualidade de voz: o uso de voz basal, o uso de voz sussurrada, a fadiga vocal e irregularidades no inicio e fim da fonação. A adequação vocal pode configurar uma situação de "desvio/variação" e "estilo", sem prejuízo da qualidade vocal. A adequação está ligada com as diferenças de bio-ritmo, factores sociais, natureza comunicativa e multiplicidade de constituições biológicas.

De um modo geral, verifica-se a existência de disfonia quando [1] a altura tonal (*pitch*), a sensação de intensidade e/ou qualidade vocal são desagradáveis ou inadequados para a idade ou sexo do falante e inaceitáveis do ponto de vista social e/ou profissional; o falante refere desconforto ou dor na fonação; o falante apresenta histórico de queixas vocais e os clínicos encontraram sinais evidentes de disfonia.

As perturbações vocais podem ser causadas pelos seguintes factores: variações hormonais; doenças inflamatórias e infecciosas do tracto

1 Fecho das pregas vocais
2 Abertura das pregas vocais

respiratório superior (cavidades nasal, oral, faríngea e laríngea); refluxo gastroesofágico e faringolaríngeo; consumo de medicamentos e drogas; stress; tabaco; consumo de álcool; condições ambientais; abuso vocal; mau uso vocal; factores de risco profissional; patologia laríngea.

1.3 Avaliação da Voz

Não existe um método único que avalie de forma abrangente e precisa a qualidade vocal [1], sendo por isso necessário recorrer a análises multifactorais que permitam um conhecimento amplo, adequado e eficaz da função laríngea e da qualidade vocal. A literatura apresenta uma grande variedade de técnicas, entre as quais a entrevista (história clínica) a avaliação da fisiologia laríngea, a avaliação perceptiva e avaliação acústica.

A avaliação por entrevista tem por objectivo avaliar factores sociais e profissionais envolventes do indivíduo, assim como do seu quadro emocional, que podem ter influência indirecta na perturbação da voz, assim como na terapia.

A avaliação da fisiologia laríngea corresponde à análise fisiológica da laríngea. De entre os métodos existentes, os mais usados, na prática clínica, são a laringoscopia indirecta, a endoscopia, a estroboscopia e a electroglotografia.

A laringoscopia indirecta é uma técnica de visualização laríngea, através de instrumentos ópticos, utilizando um espelho colocado na orofaringe[3] para o qual se dirige uma fonte de luz. Trata-se de uma técnica economicamente acessível, de aplicação rápida e pouco incómoda para o indivíduo. No entanto é uma observação bidimensional invertida (devido ao espelho) e sem ampliação, existindo dificuldade na visualização da laringe devido às condições anatómicas do indivíduo. A avaliação com produção do som é dificultada devido à língua estar amarrada, podendo ocasionar reflexo de vómito.

Na endoscopia distinguem-se duas técnicas: a endoscopia rígida e a endoscopia flexível.

A endoscopia rígida usa a via oral para fazer a visualização da laringe através de um endoscópio rígido de luz fria. Tem a vantagem de se conseguirem obter imagens amplas, estáveis e nítidas. No entanto é um técnica bastante invasiva, sendo necessário aplicação de anestesia para inibir reflexo de vómito. A produção de som ocorre numa situação incomoda com a boca aberta e língua e pescoço em extensão.

A endoscopia flexível utiliza um fibroscópio flexível via nasal, permitindo a observação das fossas nasais e das cavidades faríngeas e da laringe. É possível a fonação utilizando diferentes comportamentos vocais (vogal sustentada, fala

3 é a parte da faringe que se encontra posteriormente à boca e inferior à nasofaringe (parte da faringe que se encontra com o nariz)

e canto), assim como é possível o acoplamento a um computador permitindo a visualização num monitor, impressão e arquivo. Como desvantagem apresenta o facto de ser o método mais invasivo, de entre os apresentados, e a imagem aparece normalmente escura e distorcida.

A estroboscopia explora o fenómeno fisiológico "persistência da visão". A visão não distingue imagens individuais se apresentadas a uma velocidade superior a 5 imagens por segundo. Assim, usando *"flashes"* de luz com a mesma frequência de vibração das pregas vocais, é observada uma imagem clara e nítida, caso contrário, surge o efeito de movimento lento. Este método é também invasivo servindo-se da endoscopia.

A electroglotografia (EGG) é uma técnica não invasiva que se baseia no facto do tecido humano ser um razoável condutor de electricidade. Assim, usando um circuito eléctrico, não prejudicial para a saúde humana, é possível analisar as modificações da transmissão da corrente eléctrica resultantes da mobilidade de estruturas como, por exemplo, a mobilidade das pregas vocais. A informação da EGG quantifica a área de contacto das pregas vocais, sendo apresentada sob a forma visual, em electroglotogramas, ou parametrizada por medidas de frequência, de regularidade e de contacto das pregas vocais. Esta técnica não interfere com o processo da fala, sendo imune ao ruído ambiente. No entanto, esta técnica apenas permite avaliar o sinal laríngeo, não sendo possível obter informação relativa ao tracto vocal.

Na avaliação perceptiva, o especialista, tipicamente terapeuta da fala ou médico otorrinolaringologista, aprecia as características sonoras da voz do falante, por exemplo em resultado da fonação sustentada de uma vogal, em relação a referências perceptivas, adquiridas pelo especialista durante a sua formação ou exercício profissional, de vozes categorizadas como normais. Há inclusivamente alguns procedimentos de avaliação padronizados que permitem quantificar a severidade das perturbações percepcionadas. É disso exemplo a escala GRBAS (G – avaliação global da disfonia *(grade);* R – rouquidão *(roughness);* B – soprosidade *(breathiness);* A – astenia *(asteny);* S – tensão *(strain*)) ou RASAT (Rouquidão, Aspereza, Soprosidade, Astenia, Tensão) [1].

Os parâmetros da avaliação perceptiva podem ter uma base quantitativa que pode ser correlacionada com outras formas de avaliação (por exemplo, a análise acústica).

Esta avaliação é muito utilizada em ambiente clínico, quer por ser economicamente acessível, quer por ser desta forma que o indivíduo identifica e analisa a sua voz.

Este tipo de avaliação apresenta uma grande limitação nas possíveis divergências na classificação entre avaliadores[4]. [1] As causas desta divergência estão relacionadas com:

- o conceito de normalidade e classificação do tipo de voz patológica, porque não existe uma definição universal e estandardizada de voz

4 E até pelo mesmo avaliador em momentos diferentes.

"normal";

- a qualificação do avaliador, relacionada com a experiência na avaliação perceptiva do avaliador
- o tipo de escalas de avaliação usadas. Existe uma disparidade entre os sistemas de avaliação disponíveis. Algumas escalas usam um sistema descritivo com destaque apenas no nível laríngeo, outras no nível laríngeo e supra laríngeos. Outras, além de incluir os anteriores ainda incluem aspectos gerais relacionados com o comportamento vocal.

A análise acústica é uma técnica que permite de forma não invasiva determinar e quantificar a qualidade vocal do indivíduo. Em comparação com a EGG, que apenas permite a análise do sinal laríngeo, este tipo de análise fornece informação sobre sinal vocal laríngeo e supra laríngeo.

1.4 Estrutura da Dissertação

No capítulo 2 é apresentada a temática da análise acústica, enumerando os aspectos de engenharia implicados, descrevendo os principais parâmetros qualitativos da voz e aplicações de análise acústica existentes.

No capítulo 3 é apresentada a aplicação desenvolvida (VoiceStudio), sendo descrita a análise vozeamento que engloba o ambiente robusto de análise dos parâmetros acústicos.

No capítulo 4 são apresentados os resultados dos testes aos parâmetros qualitativos medidos com a finalidade de comprovar a robustez do ambiente.

No capitulo 5 são apresentadas as conclusões relativas à realização da dissertação, assim como perspectivas de desenvolvimento futuro para o VoiceStudio.

Capítulo 2 Análise Acústica

Neste capítulo, inicialmente é descrito como é efectuado o processo de gravação referindo os registos de voz possíveis. Em seguida são descritas algumas análises sobre o sinal, sendo posteriormente apresentadas aplicações de análise acústica.

2.1 Gravação de Voz

A gravação é o processo pelo o qual o áudio é captado e armazenado. Existem dois tipos de gravação: a gravação analógica e a gravação digital. A gravação analógica é a forma mais tradicional em contexto clínico. No entanto, a gravação digital é a forma mais vantajosa devido à melhor qualidade do sinal, destacando-se na qualidade dos sons mais agudos. A compatibilidade com a tecnologia das aplicações de análise de voz e fala, facilidade de edição, cópia e manipulação de dados e a sua não deterioração, são outras vantagens importantes. A relação sinal ruído (SNR) é mais elevada na gravação digital (96 dB numa placa de som de 16bits) do que na gravação analógica (50 dB), o que representa uma grande vantagem.

A gravação digital consiste na transformação dos sinais acústicos em sinais binários podendo existir ou não compressão de dados. A compressão beneficia em termos de dimensão de armazenamento mas perde na qualidade do sinal, podendo comprometer a análise das características do sinal acústico. Normalmente em análise acústica o armazenamento é efectuado sem compressão, sendo guardado num ficheiro de áudio na melhor qualidade possível. No sistema operativo Windows, o formato de ficheiro de áudio com estas características e normalmente utilizado é o formato WAV. Na aplicação desenvolvida (VoiceStudio) faz-se também uso do formato ACC[5] que é um formato de áudio comprimido que garante a compressão de alta qualidade.

O formato de ficheiro WAV é um formato-padrão de ficheiros do sistema operativo Windows para armazenamento de áudio digital. É um dos formatos de ficheiros de áudio mais utilizado devido à popularidade do sistema operativo Windows, e à grande quantidade de aplicações desenvolvidas para

5 O formato ACC é um formato proprietário de compressão de sinais áudio, da empresa ATC Labs, sendo conceptual e funcionalmente semelhante a outros formatos conhecidos como por exemplo o MP3.

Windows. Apesar de um ficheiro WAV poder conter áudio compactado, o formato mais comum de WAV contém áudio em formato de modulação de impulsos (PCM - *pulse-code modulation*). O formato PCM usa um método de armazenamento de áudio não-comprimido (sem perda de qualidade), assim o formato WAV é utilizado para a qualidade máxima de áudio. Os ficheiros WAV podem ser editados e manipulados com relativa facilidade por um grande número de aplicações existentes.

O formato ACC é um formato proprietário de compressão desenvolvido pela empresa *ATC Labs, Inc*. e permite comprimir os ficheiros para cerca de um décimo do tamanho de um ficheiro WAV, sem perda sensível de qualidade. É especialmente útil quando se pretende arquivar gravações áudio muito longas.

Na gravação digital devem ser considerados 2 parâmetros fundamentais: frequência de amostragem e resolução da amostra. Na amostragem do sinal são retiradas amostras do sinal acústico com uma determinada cadência, a que corresponde a frequência de amostragem. Segundo Nyquist [12], esta frequência tem que ser o dobro da frequência máxima do sinal para não existir distorção do sinal. A resolução da amostra tem uma relação directa com a extensão dinâmica da gravação, isto é, é a diferença entre a amplitude mínima e a amplitude máxima que se consegue gravar sem distorção e ruído do sistema.

Na gravação de fala para investigação ou meio clínico [1] existem diferentes registos de fala, diferindo no tipo (voz sustentada, leitura, conversação, canto), na forma de produção (voz suave, voz habitual, projecção vocal) e na duração (tempo, unidade da fala) com consequências na validade e fiabilidade das medições. Em seguida são descritos tipos de registo de fala utilizados na análise e diagnóstico da voz.

As vogais sustentadas são tradicionalmente usadas para obtenção de dados sobre a qualidade vocal. As características que levam a esta escolha prendem-se com o facto das vogais sustentadas serem mais ou menos estáveis, não contendo variações de entoação e efeitos de coarticulação, e com o facto de serem mais práticas porque são facilmente compreendidas pelo indivíduo e fáceis de produzir. No entanto, não são representativas de comunicação verbal e podem mascarar os efeitos da disfonia face às suas limitações em termos da produção.

O uso de leitura apresenta a vantagem de ser similar ao discurso espontâneo, permitindo a realização de testes e re-testes com elevada consistência, sendo possível a comparação dos dados obtidos em momentos diferentes.

O uso de discurso tem a vantagem dos dados serem mais realistas.

2.2 Análise do Sinal

2.2.1 Espectrografia

A espectrografia consiste na análise do sinal de voz no domínio das

frequências, isto é, são analisados pequenos segmentos de sinal de voz aplicando a transformada de Fourier, obtendo assim o espectro, isto é, o conjunto de frequências presentes no segmento de voz analisado. Esta análise ao longo do tempo reproduz o espectrograma, que é representado sob a forma tridimensional com o tempo no eixo horizontal (eixo dos xx), a frequência no eixo vertical (eixo dos yy), e magnitude no grau de escurecimento de cada ponto no plano x,y, tal como se ilustra na Figura 2.1.

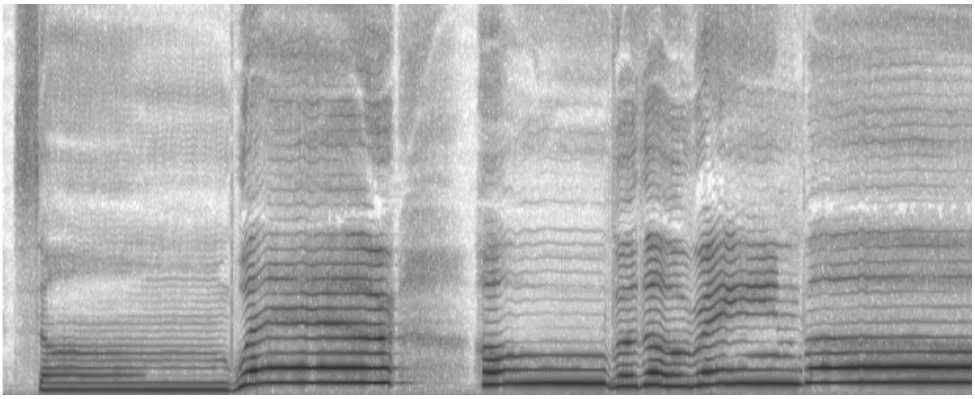


Figura 2.1: Imagem de um Espectrograma

Portanto no espectrograma é possível visualizar a representação do sinal de voz no domínio das frequências, ao longo do tempo, reflectindo as características da voz, a frequência fundamental e os harmónicos correspondentes, a magnitude de cada um dos harmónicos e duração do registo vocal.

A aplicação da espectrografia [1] centra-se na visualização e/ou quantificação da energia não vocal, isto é, o ruído visível entre os harmónicos que corresponde, em muitos casos, à componente aérea que desencadeia a disfonia. Através da gravação de vogais sustentadas é possível analisar visualmente no espectrograma:

- Flutuações da frequência fundamental, e velocidade de flutuação
- Amplitude de flutuação da frequência fundamental
- Zona do espectro onde se encontram os harmónicos com mais informação
- Frequências reforçadas, que permite identificar os formantes (frequências de ressonância existentes devido às várias componentes do tracto vocal).
- O inicio e término de uma fonação.
- Nível relativo de ruído no espectro em determinadas zonas do espectro.

Através da espectrografia vocal, [5] usando soluções tecnológicas é possível arquivar e produzir comparações entre padrões pré-estabelecidos. A sua aplicação oferece melhor compreensão acústica do sinal vocal,

possibilitando a associação entre as análises perceptivo-auditiva e acústica. A espectrografia vocal permite ainda monitorizar a eficácia de um tratamento clínico, comparando os resultados de diferentes procedimentos terapêuticos, em situações temporais diferentes.

A obtenção da representação do sinal no domínio das frequências (espectro) é realizada através de técnicas eficientes do cálculo da transformada de Fourier, vulgarmente chamada por FFT (*Fast Fourier Transform*).

2.2.2 Parâmetros acústicos

Os parâmetros acústicos obtidos pela análise acústica têm a vantagem de descrever a voz de forma objectiva. Com a existência de bases de dados normativas que caracterizam a qualidade vocal, é possível distinguir entre voz normal e patológica, avaliar a voz e sua monitorização do ponto de vista clínico e/ou profissional e diminuir o grau de subjectividade da análise perceptiva ao ser feita a sua correlação com os dados quantitativos.

Actualmente os parâmetros acústicos mais utilizados em aplicações de análise acústica, assim como mais referenciados na literatura, são o *pitch, jitter, shimmer* e HNR.

2.2.2.1 Pitch

O *pitch* da voz, também referenciada como frequência fundamental ou F_0, é o termo usado para referir o parâmetro físico resultante da vibração das pregas vocais por unidade de tempo no comportamento vocal sustentado ou em fala encadeada. Em rigor, o *pitch* denota o correspondente psicofísico (perceptivo) da frequência fundamental (F_0) e é condicionado por outros factores objectivos do sinal de voz como seja a sua intensidade.

O pitch reflecte a eficiência do sistema fonatório, a biomecânica laríngea e a sua interacção com a aerodinâmica.

Este parâmetro pode ser obtido por análise acústica, sendo usualmente descrito pelo valor médio, pelo desvio padrão e pela extensão em frequência.

O valor de F0 depende do sexo do orador e varia em função da sua idade. Na criança o F0 é elevado e encontra-se na faixa dos 260 a 320 Hz. Com o envelhecimento e já depois da fase da puberdade existe uma diminuição do nível médio de F0. Portanto, em adulto nota-se uma variabilidade de F0 quanto ao sexo, em que um adulto feminino tem um valor de F0 mais elevado (140 a 260Hz) do que um adulto masculino (80Hz a 140Hz).

O valor de F0 também difere com o comportamento vocal, isto é, se é produzida uma vogal sustentada, leitura, conversação, canto e contagem de números. A análise mediante uma vogal sustentada varia dependendo da vogal produzida. Assim a produção de vogais altas, como /i/ e /u/ são produzidas normalmente com F0 mais elevado do que as vogais baixas: /a/, /e/ e /o/. A produção de vogais depende da postura da língua, e por esta estrutura ter ligações à estrutura laríngea provoca variações no valor de F0.

No ambiente desenvolvido para o cálculo do valor de *pitch* foi utilizado o algoritmo desenvolvido pela empresa SEEGNAL Research, que avalia as características harmónicas do sinal [20][22].

2.2.2.2 Jitter

O *jitter* é uma medida de curto termo [7] (de um ciclo de vibração das pregas vocais para o ciclo seguinte) de variabilidade não voluntária da frequência fundamental, o que permite determinar o grau de estabilidade do sistema fonatório. Não deve ser confundida com medidas de oscilação de baixa frequência como o vibrato e o tremor.

As fontes desta variabilidade [1] podem ser do tipo neurológicas (sinais neurológicos descoordenados causam variações na contracção muscular), biomecânicas (por exemplo, a pulsação devido à circulação do sangue nos capilares do tecido das cordas vocais), aerodinâmicas (instabilidades no fluxo de ar) e acústicas (pela interacção entre a glote e o tracto vocal). Analisando estas fontes, é previsível e aceitável a presença de um pequeno grau de perturbação e irregularidade no sinal de voz.

A extracção dos valores de *jitter* pode ser efectuada a partir de medidas absolutas ou relativas.

As medidas absolutas têm dimensão temporal (são representadas em unidades de tempo), não estando relacionadas com o valor da frequência ou período fundamental.

As medidas relativas são adimensionais porque estão referenciadas relativamente ao valor do período fundamental. As medidas relativas são mais utilizadas porque traduzem a avaliação da variação dependente da frequência ou período fundamental. Existem várias formas de calcular o *jitter*, mesmo dentro das medidas relativas, em que variam na dimensão da janela utilizada para avaliar o *jitter*. Esta normalização aplica-se no cálculo do *jitter* com as seguintes designações: *jitter* local, RAP (*Relative Average Perturbation*), PPQ5 (PPQ - *Period Perturbation Quocient*), PPQ11.

A equação geral para o cálculo do *jitter* (apenas para medidas relativas) é apresentada na equação (2.1), sendo utilizada para o cálculo do: RAP (p=3); PPQ5 (p=5); PPQ11 (p=11). Relativamente ao cálculo do *jitter* local é utilizada a equação (2.2).

$$Jitter = \frac{\frac{1}{n_T - p + 1} \times \sum_{i=t+1}^{n_T - t} \left| \frac{\sum_{k=i-t}^{i+t} T_{0k}}{p} - T_{0i} \right|}{T_{0med}} \times 100 \,, t = \frac{p-1}{2} \,, T_{0med} = \sum_{i=1}^{n_T} T_{0i} \quad , \tag{2.1}$$

em que n_T corresponde ao número de valores de T_0, k e i correspondem a índices dos valores de T_0, T_{0i} corresponde a um valor de T0 avaliado relativamente à média de p valores vizinhos T_{0k} .

$$Jitter\ Local = \frac{\frac{1}{n_T - 1} \times \sum_{i=1}^{n_T - 1} |T_{0i+1} - T_{0i}|}{T_{0\,med}} \times 100\,,\ T_{0\,med} = \sum_{i=1}^{n_T} T_{0i}\ \text{,} \qquad (2.2)$$

n_T, $T_{0\,i}$ apresenta o mesmo significado que na equação (2.1)

Na equação geral do cálculo do *jitter* (equação (2.1)), é avaliada a existência de *jitter* ciclo a ciclo como um desvio de T_0 (período fundamental) actual relativamente a uma média local de *p* ocorrências adjacentes de T_0.

A medida de *jitter* mais referenciada na literatura e em aplicações de análise acústica é a medida relativa, sendo normalmente utilizado o RAP.

Para uma boa medição do valor de *jitter* [1] é preciso ter em conta os seguintes aspectos:

- Deve ser medido apenas em vogais sustentadas, a um nível confortável de frequência e intensidade, em vez de fala encadeada, devido à existência de perturbações voluntárias da frequência fundamental. Deve-se portanto eliminar o inicio e o fim da fonação.
- Cada vogal tem valores próprios intrínsecos de F0 e por isso deve ser indicada qual a vogal analisada.
- A gravação deverá decorrer em formato digital para maximizar a relação sinal-ruído, assim como deverá ocorrer num ambiente sem grande incidência de ruído.
- Dimensão da amostra, isto é, para obter valores fiáveis de F0 deve-se utilizar uma amostra de voz que contenha pelo menos 110 ciclos [1].

Um dos potenciais interesses sobre as medidas de *jitter* é o diagnóstico diferencial entre voz normal e patológica, dado que estão cientificamente comprovadas diferenças estatisticamente significativas nos valores de *jitter* entre indivíduos com perturbação e/ou patologia vocal e indivíduos sem quaisquer sintomas vocais anómalos.

Quanto aos limiares patológicos, na literatura não existe muito consenso, devido à diversidade de aplicações usadas, às diferentes formas de calcular o *jitter*, e às diferentes bases de dados de vozes patológicas e normais utilizadas. No entanto, em termos gerais, esses limiares situam-se entre 1% e 2%.

2.2.2.3 Shimmer

O *shimmer* é uma medida de curto termo [7] (de um ciclo de vibração das pregas vocais para o seguinte ciclo) de variabilidade não voluntária da amplitude de cada ciclo, quantificando as alterações mínimas da amplitude do sinal, com base em cada ciclo fonatório. O *shimmer* pode ser medido em dB, como uma avaliação da variação logarítmica entre a amplitude em ciclos consecutivos, através da equação (2.3).

$$Shimmer(dB) = \frac{20 \times \sum_{k=0}^{k=n_A-1} |\log_{10} \frac{A_k}{A_{k+1}}|}{n-1} \quad , \tag{2.3}$$

n_A corresponde ao número de ciclos avaliados, k corresponde ao índice do ciclo, A_k corresponde à amplitude de amostra do ciclo k.

A amostra do sinal de áudio utilizada para análise do *shimmer* corresponde à amostra utilizada na marcação do *jitter*.
O *shimmer* também pode ser medido em percentagem (%) como uma avaliação da variação entre amplitudes de ciclos consecutivos, tendo em conta uma média local de *p* ocorrências adjacentes de amplitude, descrito na equação (2.4). Existem vários métodos de cálculo do *shimmer* em percentagem, entre os quais: *shimmer* local, descrito na equação (2.5); APQ5 (APQ – Amplitude Perturbation Quocient) descrito na equação (2.4), em que p=5; APQ11 descrito na equação (2.4), em que p=11.

$$Shimmer = \frac{\frac{1}{n_A - p + 1} \times \sum_{i=t+1}^{n_A - t} |\frac{\sum_{k=i-t}^{i+t} A_k}{p} - A_i|}{A_{med}} \times 100 \, , \, t = \frac{p-1}{2} \, , \, A_{med} = \sum_{i=1}^{n_A} A_i \quad , \tag{2.4}$$

k, n_A e A_i apresentam o mesmo significado que na equação (2.3).
A_i e i tem o mesmo significado que k e A_k respectivamente.

$$Shimmer\ Local = \frac{\frac{1}{n_A - 1} \times \sum_{i=1}^{n_A - 1} |A_{i+1} - A_i|}{A_{med}} \times 100 \, , \, A_{med} = \sum_{i=1}^{n_A} A_i \tag{2.5}$$

i, n_A e A_i apresentam o mesmo significado da equação (2.3).

Existem outros métodos de medição de *shimmer* tais como índice de variabilidade de amplitude (AVI) e o factor de perturbação direccional (DPF, número de vezes que a diferença de amplitude em ciclos consecutivos muda de direcção) mas estes são menos referenciados na literatura e aplicações.

O *shimmer* similarmente ao *jitter* deve respeitar as mesmas condições para ser correctamente medido.

Quantos ao limiares de patologia, similarmente ao *jitter*, não existe muito consenso, mas no entanto são muito utilizados os valores entre 3 e 5%.

2.2.2.4 HNR

O HNR é uma avaliação da relação entre a componente periódica e a componente aperiódica que compõem um segmento sustentado de voz vozeada. A primeira componente decorre da vibração das pregas vocais e a segunda decorre do ruído glótico. A avaliação entre as duas componentes

traduz a eficiência do processo de fonação: quanto maior for a eficiência na utilização do fluxo de ar expelido pelos pulmões em energia de vibração das pregas vocais, e quanto mais íntegro for o ciclo vibratório destas, maior será o valor de HNR. Inversamente, quanto menor for aquela eficiência ou quanto mais anómalo for o ciclo vibratório, maior será o ruído glótico e mais baixo resultará o valor de HNR. Uma voz saudável deve assim caracterizar-se por um valor de HNR elevado, a que se associa a impressão de voz sonora e harmónica. Um baixo valor de HNR denota uma voz apagada e soprosa.

Em termos matemáticos, um sinal vozeado (com estrutura harmónica) no domínio das frequências pode ser representado pela equação (2.6)

$$V(w)=H(w)+N(w) \quad ,$$

em que V(w) corresponde ao sinal de áudio no domínio das frequências, H(w) corresponde à componente harmónica e N(w) à componente de ruído. (2.6)

A relação HNR é, por definição, uma medida logarítmica da relação das energias associadas às duas componentes, o que presume a integração da potência espectral ao longo da gama audível de frequências (ver equação (2.7))

$$HNR=10\times\log_{10}\frac{\int_{w}|H(w)|^2}{\int_{w}|N(w)|^2} \quad . \tag{2.7}$$

A medida logarítmica deve a sua pertinência à boa correlação com a percepção de intensidade sonora (ou volume sonoro). Por outras palavras, o HNR tenta medir a relação entre a percepção da componente periódica de um som vozeado, e a percepção da componente de ruído desse sinal.

Na prática, o cálculo do espectro é realizado através de técnicas eficientes como a FFT. O espectro é calculado não como uma função contínua, mas como uma amostragem desta função pelo que, na prática, o operador integração dá lugar ao somatório (equação (2.8))

$$HNR=10\times\log_{10}\frac{\sum_{k}|H(w_k)|^2}{\sum_{k}|N(w_k)|^2}. \tag{2.8}$$

Há autores [26] que argumentam que a expressão anterior não é apropriada para avaliar o ruído glótico anterior à interacção com o tracto vocal. Por outras palavras, se se pretender avaliar o HNR da fonte glótica em vez do HNR decorrente do sinal de voz, o efeito do tracto vocal deverá ser cancelado (GHNR , *Glottal Harmonic-to-Noise Ratio*).

No contexto da dissertação, pretende-se que a avaliação acústica esteja

em linha com a avaliação perceptiva a partir do sinal de voz, o que implica que o parâmetro HNR inclua da influência do tracto vocal.

O cálculo do HNR encerra uma dificuldade básica: o sinal de áudio captado por um microfone (v(n)), cuja transformada de Fourier é V(w). As componentes H(w) e N(w) encontram-se combinadas no sinal, pelo que é necessário separá-las. Os métodos de cálculo propostos por vários autores diferem, sobretudo, nas técnicas de estimação destas componentes a partir de V(w).

Para melhor se caracterizar o problema e, também, ser possível avaliar o desempenho de aplicações alternativas de análise acústica, implementou-se um modelo de geração voz sintética (ver capítulo "Testes aos principais parâmetros qualitativos da voz"). Deste modo, é possível gerar sinais de voz sintética com valores pré-determinados de jitter, shimmer e HNR. Gerando vozes sintéticas correspondentes à vogal sem componente de ruído glotal (apenas componente harmónica), com determinado ruído glotal, de tal modo que HNR seja igual a 20dB; e sem presença da componente harmónica (apenas componente de ruído) obtém-se os espectros dos sinais representados nas figuras (2.2) , (2.3), (2.4), respectivamente.

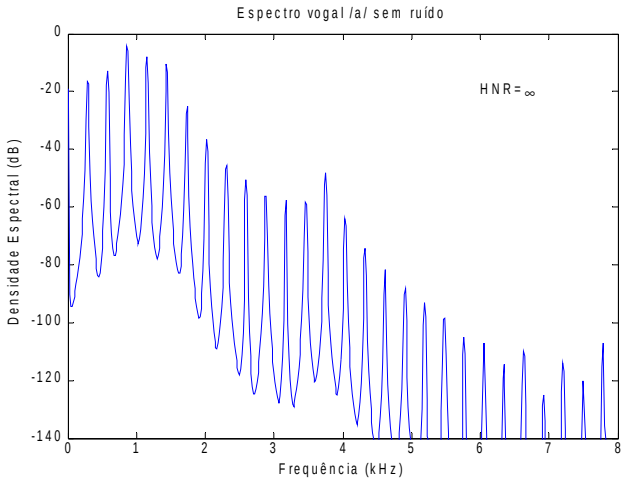


Figura 2.2: Representação do espectro de voz sintética na ausência de ruído glótico.

A Figura (2.2) revela que a regularidade harmónica do espectro é perfeita, sem descontinuidades ou perturbação por ruído.

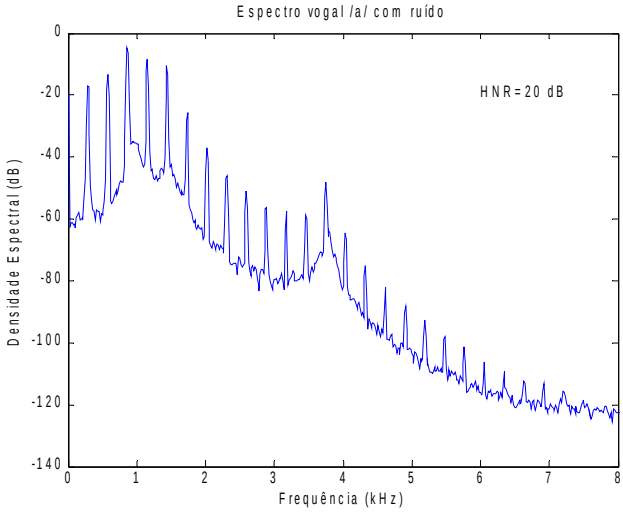


Figura 2.3: Representação do espectro de voz sintética correspondente à vogal /a/ afectada por ruído glótico.

A Figura (2.3) permite clarificar que, em relação à Figura (2.2) , o ruído diminui a pureza da estrutura harmónica, reduzindo muito o destaque de cada harmónico (ou parcial da estrutura harmónica) em relação ao ruído, podendo inclusivamente "apagar" a sua presença.

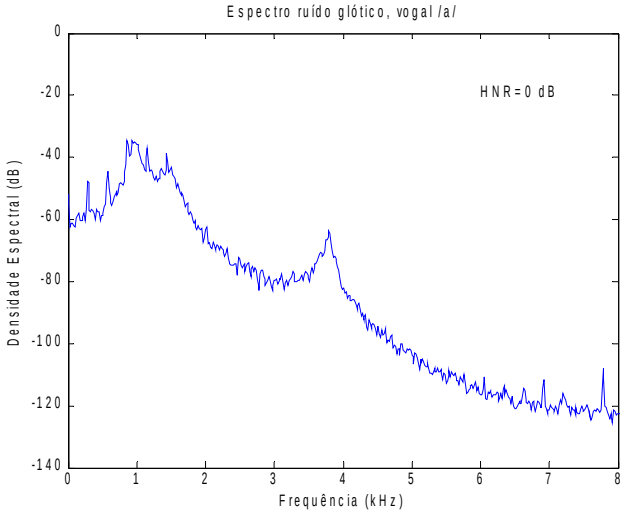


Figura 2.4: Representação do espectro de voz sintética correspondente à vogal /a/ quando a fonte é constituída só por ruído glótico.

Na Figura (2.4) verifica-se que para além de alguns picos espúrios, não há evidência na representação espectral de uma estrutura harmónica.

As figuras (2.2), (2.3) e(2.4) permitem compreender melhor o desafio no cálculo do HNR: ao captar um sinal de voz, o espectro a que se tem acesso é o

ilustrado na Figura (2.3), ou seja, a V(w). Como base neste dever-se-á usar uma estratégia para obter as componentes ilustradas nas Figuras 3 e 5 de modo a ser viável calcular a sua potência espectral e, a partir desta, a relação HNR.

Exemplos de algoritmos no cálculo do HNR são: o algoritmo utilizado na aplicação Praat (ver "Aplicações de Análise Acústica"); o algoritmo utilizado na aplicação Dr. Speech (ver "Aplicações de Análise Acústica").

O Praat utiliza para o cálculo do HNR um algoritmo desenvolvido por Boersma [23]. A sua abordagem é indirecta já que não realiza a separação de componentes atrás descrita, seguindo antes um procedimento baseado nas propriedades da função auto-correlação[6]. De facto, calculando a autocorrelação (AC) do sinal de voz representado por v(n), obtém-se a equação (2.9)

$$AC_V(\tau)=\sum_{(n)} v(n)\times v(n+\tau)=AC_H(\tau)+2\times CC_{H,N}(\tau)+AC_N(\tau) \quad , \qquad (2.9)$$

em que o operador CC representa a correlação cruzada e os índices H e N simbolizam, respectivamente, as componentes harmónica e ruído do sinal de voz. Dado que as componentes harmónica e ruído se consideram independentes e, portanto, não correlacionadas, a sua correlação cruzada é nula pelo que a equação (2.9) se reduz à equação (2.10)

$$AC_V(\tau)=AC_H(\tau)+AC_N(\tau) \quad . \qquad (2.10)$$

Por definição, quando o parâmetro τ é nulo, a função $AC_V(\tau)$ exibe um máximo global que traduz a potência total do sinal de voz que, por sua vez, resulta da soma de potência das componentes harmónica e de ruído (equação (2.11))

$$AC_V(0)=AC_H(0)+AC_N(0) \quad . \qquad (2.11)$$

Admitindo que o ruído é branco (ou de densidade espectral plana), a função $AC_N(\tau)$ é nula para $\tau\neq 0$. Por outro lado, admitindo estacionaridade, a função $AC_H(\tau)$ é periódica e, em particular, o valor de $AC_H(0)$ repete-se quando τ é múltiplo inteiro do período fundamental da voz, $T=1/F_0$. Por outras palavras, a função autocorrelação do sinal de voz (vozeada) exibe máximos locais para valores de τ múltiplos inteiros do período fundamental. Assim, para encontrar a relação HNR basta basta calcular a função autocorrelação do sinal de voz, identificar o primeiro máximo local e ler o valor correspondente à potência da componente harmónica ($AC_V(T)=AC_H(T)=AC_H(0)$). O valor de potência correspondente à componente de ruído determina-se usando a

6 auto-correlação é o resultado da correlação de um sinal com ele mesmo.

equação (2.11). O valor HNR calcula-se, finalmente, a partir da equação (2.12)

$$HNR = 10 \times \log_{10} \frac{AC_V(T)}{AC_V(0) - AC_V(T)} \quad . \tag{2.12}$$

Na prática, o cálculo do HNR por esta via indirecta exige alguns cuidados especiais, além de que os pressupostos de estacionaridade e ruído branco, em rigor, não se verificam.

O Dr. Speech utiliza o algoritmo de Yumoto e Gould [24] cuja técnica fundamenta-se num facto simples: dado que a componente de ruído de um sinal de voz pode ser modelizada como ruído aditivo de média nula, a componente harmónica, $v_H(t)$, pode ser estimada somando um número elevado (K) de períodos do sinal, após alinhamento cuidado (equação (2.13))

$$v_H(t) = \frac{1}{K} \sum_k v(t - T_k), 0 \leqslant t \leqslant T \quad . \tag{2.13}$$

A necessidade de alinhamento decorre da circunstância da voz natural exibir sempre algum nível de *jitter*. Por esta razão, na equação (2.13) ter-se-á que Tk≠kT, caso contrário o *jitter* seria nulo e o sinal seria perfeitamente periódico. A componente de ruído pode ser estimada calculando a diferença entre cada período do sinal de voz e o sinal $v_H(t)$, após alinhamento. Deste modo, a relação HNR é obtida usando a equação (2.14)

$$HNR = 10 \times \log_{10} \frac{K \int_0^T |v_H(t)|^2 \, \delta t}{\sum_k \int_0^T |v(t - T_k) - v_H(t)|^2 \, \delta t} \tag{2.14}$$

Sendo uma técnica do domínio do tempo, possui a vantagem de ser computacionalmente simples mas encerra algumas fragilidades, como por exemplo, ser muito sensível a desalinhamentos e não admitir ruído de natureza não-linear. Os autores reconhecem inclusivamente que o método possa ser inválido em casos de disfonia.

Uma outra abordagem de cálculo do HNR foi inicialmente proposta por Krom [27] e subsequentemente modificada por Qi [28]. Esta abordagem baseia-se na propriedade do cepstrum[7] permitir desacoplar a componentes de

7 O cepstrum aqui considerado (cepstrum real) consiste na transformada de Fourier inversa do logaritmo do espectro . Remete portanto para um domínio do tempo que caracteriza a periodicidade existente no espectro.

variação rápida do espectro (relacionadas com os harmónicos) e as componentes de variação lenta do espectro (relacionadas com a envolvente espectral que retrata razoavelmente o perfil do ruído e, portanto, os formantes). Deste modo - identificando os picos do espectro correspondentes às componentes harmónicas e usando diversos passos de filtragem, que permitem obter uma estimativa do espectro do ruído – é possível calcular o HNR através da equação (2.8). Apesar de mais directa, esta abordagem é vulnerável à natureza dos sinais de voz e, em particular, os seus resultados dependem muito da frequência fundamental (F_0). Estes problemas foram subsequentemente minimizados em novos resultados publicados por Murphy [26].

Na presente dissertação foi desenvolvido um novo método de cálculo que procura atender à definição da medida HNR, através da segmentação das componentes harmónica e ruído a partir do sinal de voz captado por um microfone. O novo método combina funções de análise e síntese de sinal com o objectivo de segmentar fisicamente as duas componentes (componente harmónica e de ruído) e permitir a sua reconstrução individual. Esta é uma capacidade de processamento de sinal inovadora e inspira-se em técnicas de estimação precisa das componentes sinusoidais no sinal de voz [20][22], e em técnicas de reconstrução no tempo de sinais a partir de modelos no domínio da frequência [21].

O procedimento matemático subjacente pode ser sintetizado nos seguintes passos principais de processamento de sinal:

1. é efectuada uma análise precisa de todas as componentes sinusoidais existentes no sinal de voz e, a partir destas, é identificada a estrutura harmónica mais plausível do sinal de voz;
2. a estrutura harmónica mais plausível é modelizada parametricamente e é reconstruída no domínio das frequências;
3. a estrutura harmónica reconstruída no domínio das frequências é subtraída ao espectro complexo do sinal de voz (incluindo informação de magnitude e fase) de forma a obter-se um resíduo espectral que corresponde à estimativa do espectro do ruído;
4. as representações espectrais da estrutura harmónica e da estimativa do ruído são usadas quer para determinar a sua potência média, quer para reconstruir os correspondentes sinais no domínio do tempo. As duas medidas de potência assim obtidas são usadas para calcular a relação HNR de acordo com a equação (2.8).

Trata-se de um método directo de análise precisa, segmentação e síntese das componentes harmónica e ruído, o que encerra o potencial de permitir medições mais rigorosas.

2.3 Aplicações de análise acústica

A grande vantagem da utilização do sinal digital de voz está relacionada com o facto de poder ser utilizado em aplicações de edição e análise acústica. Estas aplicações possibilitam a visualização do sinal acústico e das suas características, que variam de aplicação para aplicação, distinguindo-se nas medidas, nas representações disponibilizadas e na eficiência de utilização.

Speech Filing System (SFS), Praat, Dr. Speech, Voxmetria, CSL (Kay Elemetrics) são exemplos de aplicações de análise acústica já utilizadas em ambiente acústico. Em seguida serão analisadas em mais detalhe.

2.3.1 Praat

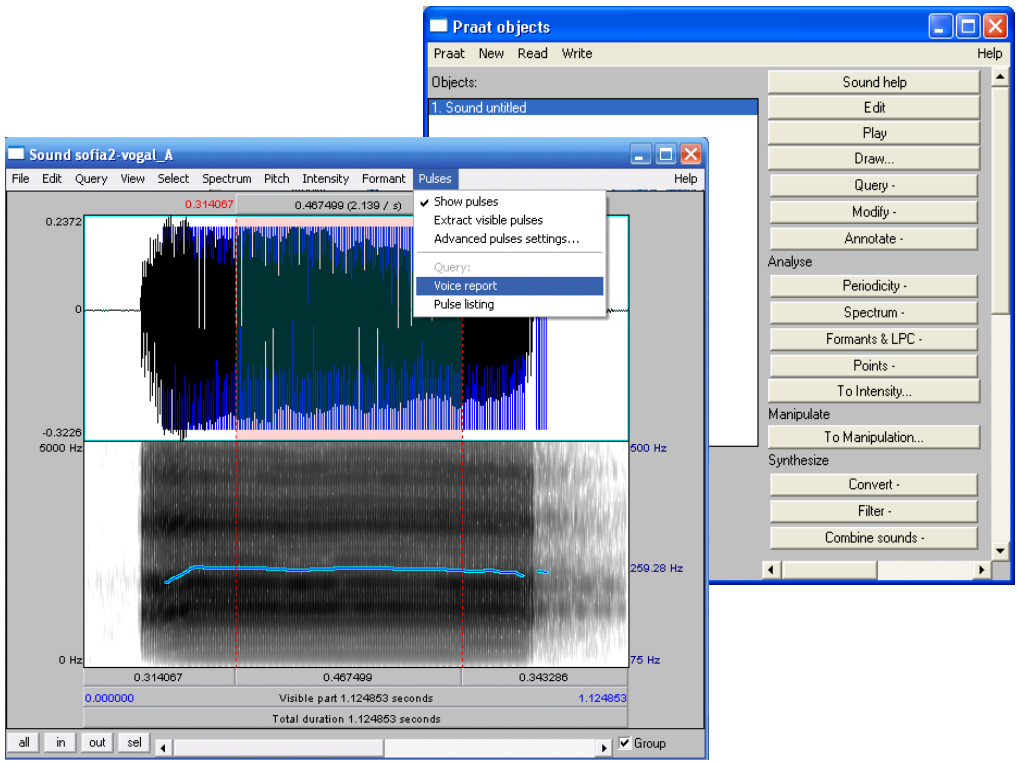


Figura 2.5: Imagem da aplicação Praat

O Praat [11] é uma aplicação de código livre e gratuito para análise e síntese de voz, que foi desenvolvida por Paul Boersma e David Weenink da Universidade de Amesterdão, ilustrada na Figura 2.5. Relativamente à análise acústica esta aplicação possibilita análise espectral (espectrografia), análise dos parâmetros qualitativos: *pitch*, *jitter*, *shimmer* e HNR, análise dos formantes e análise da intensidade do sinal.

O Praat é uma aplicação sobretudo muito utilizada em meio académico e de investigação.

2.3.2 Dr. Speech

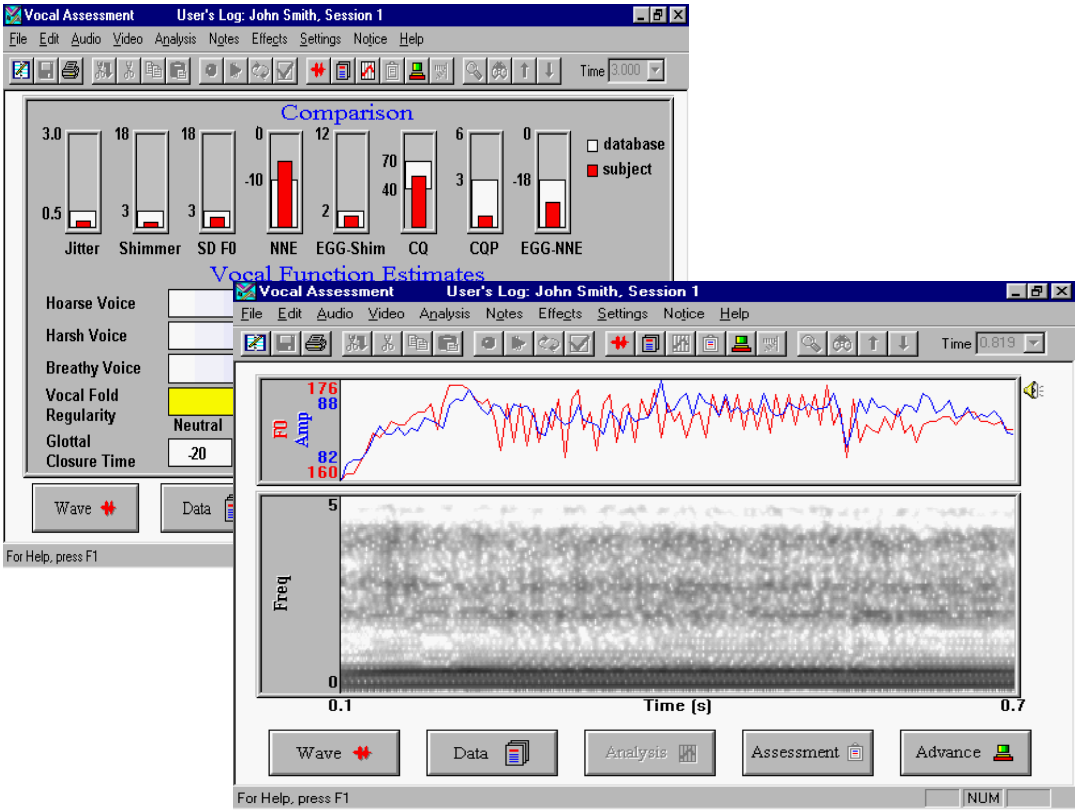


Figura 2.6: Imagem da ferramenta Vocal Assessment da aplicação Dr. Speech

O Dr. Speech [9][10] é uma aplicação proprietária para análise de voz desenvolvida pela empresa Tiger Electronics. O Dr. Speech é composto por um conjunto de ferramentas (na Figura 2.6 está apresentada a ferramenta *Vocal Assessment* que possibilita extracção dos parâmetros qualitativos) que permitem análise em tempo real do valor da frequência fundamental da voz, visualização de espectrogramas, análise da intensidade da voz, detecção de vogais, análise de formantes, visualização de fonetograma, cálculo dos parâmetros qualitativos *jitter*, *shimmer* e energia normalizada de ruído (NNE - *Normalized Noise Energy*); análise de electroglotografia (EGG) e terapia da voz mediante de jogos ou aplicações de análise de comparação com padrão em tempo real.

O Dr. Speech é uma aplicação frequentemente usada em meio clínico ou académico [25].

2.3.3 SFS

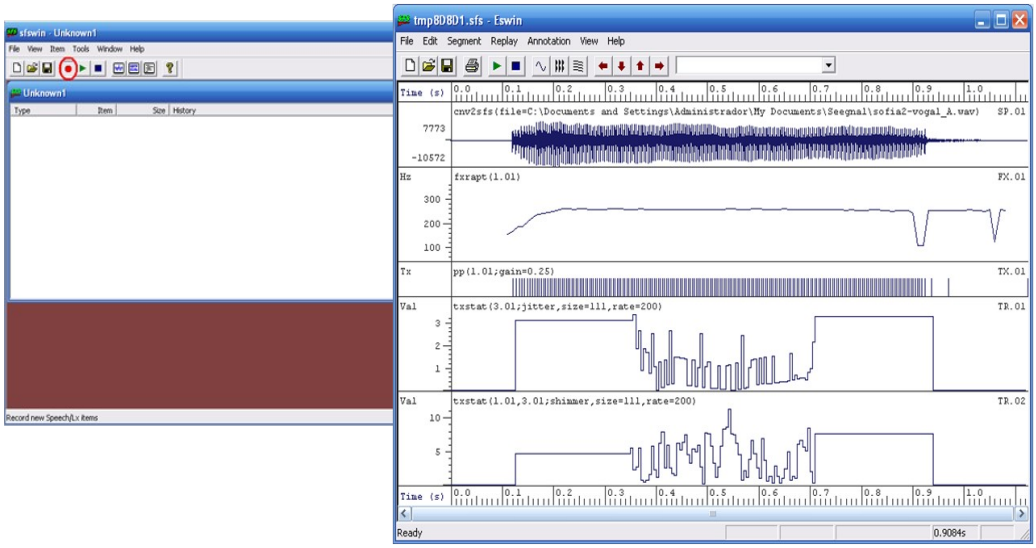


Figura 2.7: Imagem da aplicação SFS

O SFS (Speech Filing System) [13] é um ambiente de livre computação para investigação. Esta aplicação tem ferramentas de software, formatos de ficheiros e dados, bibliotecas para desenvolvimento, gráficos, linguagens especiais para programação e documentação. Apresenta a possibilidade de aquisição, reprodução e visualização de áudio, espectrografia, análise de formantes e estimação da frequência fundamental.

Este software apresenta várias ferramentas para processamento de sinal, síntese e reconhecimento para desenvolvimento de aplicações.

Na Figura 2.7 está representado a interface principal do SFS e ao lado está representada a extracção dos parâmetros qualitativos.

2.3.4 CSL

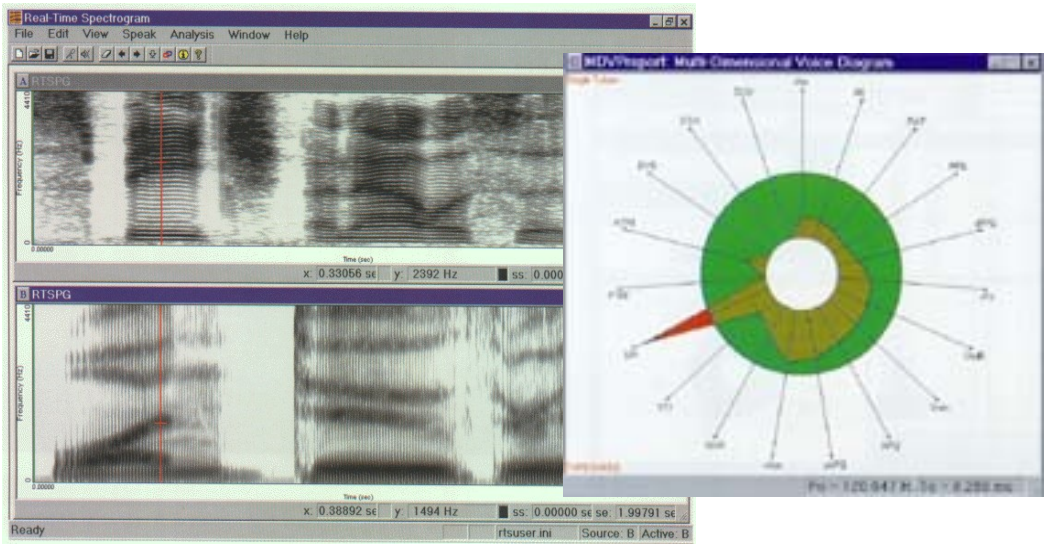


Figura 2.8: Imagens de algumas aplicações do CSL

O CSL (Computer Speech Lab) [14] é o sistema de hardware e software mais avançado de análise de fala da empresa KayPentax. A nível de hardware o CSL vem equipado com uma placa de captura e reprodução de som externa de baixa latência para garantir uma boa gravação.

A nível de software, o CSL apresenta um conjunto de várias aplicações de análise e terapia da voz, assim como uma base de dados de vozes patológicas com 1400 vozes de 700 indivíduos. De entre as aplicações de análise destacam-se o Real-Time Pitch, Real-Time Spectrogram e MDVP (MultiDimensional Voice Program). A aplicação Real-Time Pitch permite análise em tempo real ou diferido da estimação do valor de pitch. O Real-Time Spectrogram apresenta 2 espectrogramas que podem ser actualizados em tempo real, e que podem ser definidos com configurações diferentes, sendo possível ter espectrograma de banda larga e banda estreita. O MDVP é uma das aplicações mais referenciados de extracção de parâmetros qualitativos da voz, que extrai 22 parâmetros acústicos, incluindo *jitter*, *shimmer* e HNR.

Na Figura 2.8 é apresentado o Real-Time Spectrogram e o MDVP do CSL.

2.3.5 VoxMetria

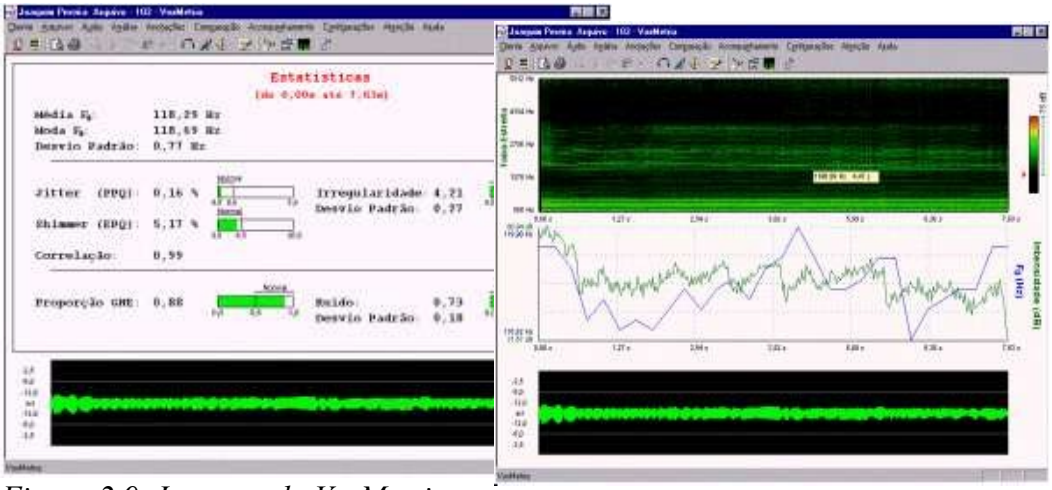


Figura 2.9: Imagens do VoxMetria

O VoxMetria [15] é uma aplicação desenvolvida pela empresa CTS Informática.

Esta aplicação permite a edição de áudio, possibilitando a análise da frequência fundamental e intensidade, análise de formantes, espectrografia (espectrograma de banda larga e estreita), extracção de parâmetros acústicos, entre os quais *jitter* e *shimmer* e diagrama de desvio fonatório.

Esta aplicação suporta também gestão de clientes, sistema de anotações e impressão de todos os gráficos que aparecem na aplicação.

Na Figura 2.9 está apresentada 2 representações possíveis do VoxMetria.

Capítulo 3 Sistema desenvolvido

Neste capitulo serão descritas as ferramentas utilizadas para implementação da aplicação, isto é, a linguagem de programação, o compilador e as bibliotecas utilizadas. A arquitectura do sistema indicará os módulos existentes assim como a interacção entre eles. Em seguida serão descritos todos os módulos da aplicação.

3.1 Ferramentas de desenvolvimento

Para o desenvolvimento do sistema foi utilizada a linguagem de programação C++, que se trata de uma linguagem de programação de alto nível com facilidades para uso em baixo nível de uso geral. Tem a característica de ser uma linguagem de programação procedimental e orientada a objectos, e que apresenta um bom desempenho.

Para o desenvolvimento do interface gráfico recorreu-se à biblioteca de código livre e multiplataforma Qt [16], que apresenta uma API para criação da interface gráfica assim como disponibiliza outras funcionalidades, entre as quais, interacção com a rede IP, interacção com base de dados, interface para OpenGL, entre outras. Também foi utilizada a biblioteca de código livre e multiplataforma Qwt [17] que facilita do desenho de gráficos 2D, utilizando a biblioteca Qt como sua base.

Para efectuar a interacção com a placa de som foi utilizada a biblioteca de código livre e multiplataforma RtAudio [18] que disponibiliza uma API para interacção em tempo real com a entrada e saída da placa de som. Possibilita também a identificação dos dispositivos de entrada e saída de áudio, assim como a sua escolha para processamento.

A ferramenta de edição de código e compilação utilizada foi o Microsoft Visual C++ Express 2005 [19], orientada para o desenvolvimento de aplicações no sistema operativo Windows.

3.2 Arquitectura

O ambiente de desenvolvido integra várias funcionalidades, pelo que foi necessário particionar o sistema em módulos funcionais. Na Figura 3.1 estão apresentados os módulos existentes na aplicação e interacção entre eles.

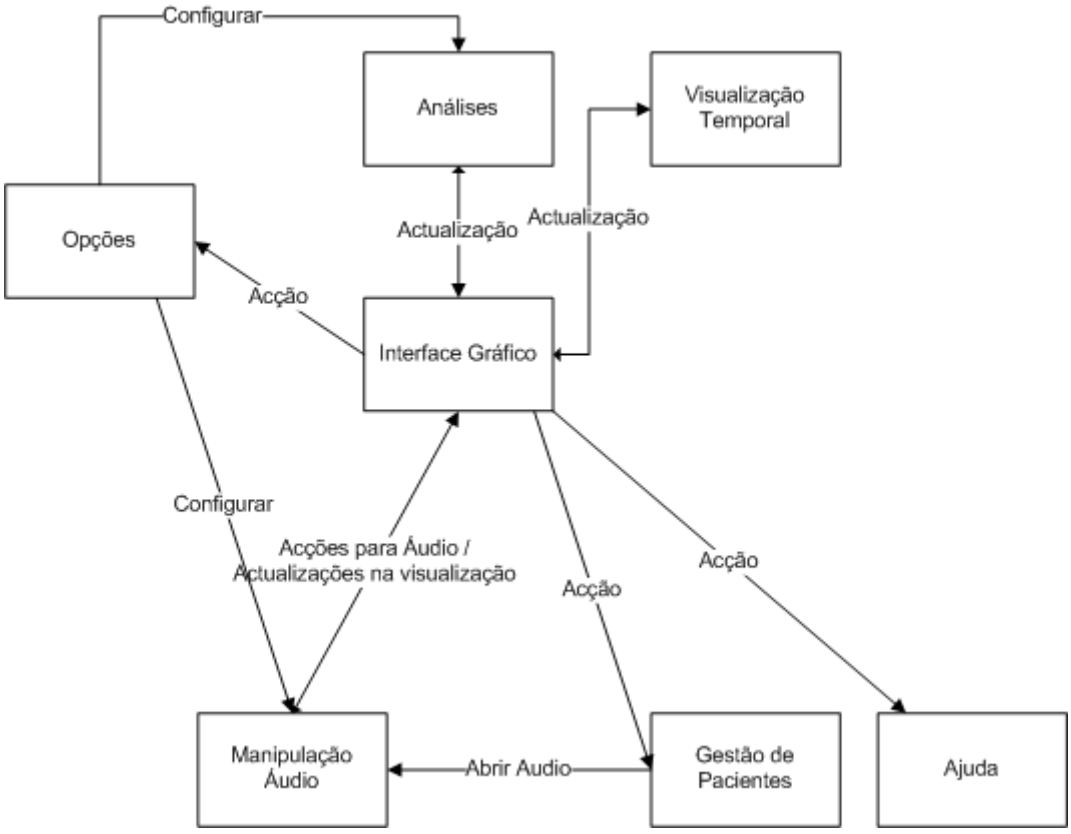


Figura 3.1: Arquitectura geral da aplicação desenvolvida

A comunicação entre os módulos é efectuada utilizando o sistema de eventos da biblioteca Qt.

3.3 Interface Gráfica

A interacção com o utilizador é um dos aspectos mais importantes em qualquer aplicação devido à necessidade de facilitar a utilização da aplicação.

Na Figura 3.2 é apresentada a interface da aplicação, estando esta seccionada e legendada para possibilitar a descrição de cada componente funcional.

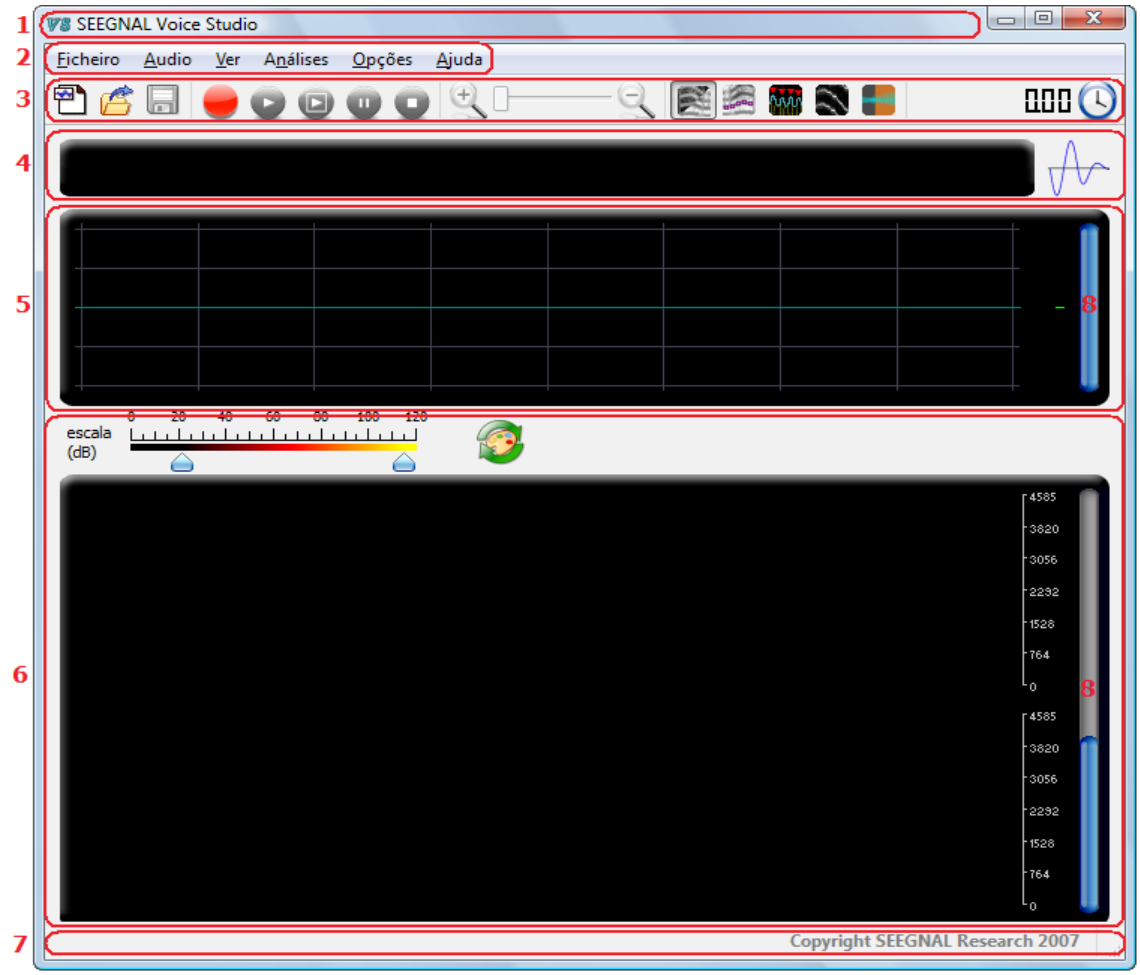


Figura 3.2: Interface Gráfica da aplicação desenvolvida

1. Barra de título – Apresenta o nome da aplicação, bem como, caso se aplique, do paciente sob análise e do ficheiro de áudio aberto.
2. Menu da aplicação – Apresenta os vários menus para controlo da aplicação.
3. Barra de ferramentas – Apresenta os botões de acção rápida.
4. Representação global do sinal – Permite obter uma visão global do sinal áudio.
5. Representação do sinal seleccionado – Apresenta a secção do sinal áudio seleccionada na "Representação global".
6. Secção de Análises– Esta zona apresenta a imagem relativa à análise actualmente seleccionada, bem como botões de acesso rápido específicos de cada tipo de análise.
7. Barra de estado – A barra de estado apresenta informações extra, como por exemplo as coordenadas de um gráfico com base na posição actual do rato, ou parâmetros importantes do áudio representado como por exemplo a frequência de amostragem.
8. Barras de deslocamento vertical / Ampliação vertical – As barras de

deslocamento vertical têm uma função dupla: por um lado, permitem uma deslocação vertical da parte visível do sinal em análise, por outro lado é a partir destas que se pode aumentar ou diminuir a ampliação vertical. Para tal, deve-se clicar com o botão esquerdo do rato num dos extremos da barra e arrastá-lo até se atingir o nível de ampliação vertical desejado. Clicando na parte central da barra e arrastando-a, está-se a usá-la para mover verticalmente a zona visível do sinal em análise. Ao efectuar um duplo-clique sobre a barra, esta regressa ao seu estado de abrangência máxima.

3.4 Modulo de Manipulação de áudio

O modulo de manipulação de áudio é o componente do ambiente desenvolvido que interage com áudio, em que estão inseridos os processos de gravação e reprodução, assim como a abertura e gravação de ficheiros áudio.

O ambiente desenvolvido permite a utilização de ficheiros de áudio com o formato WAV e o formato ACC.

Também permite uma interacção em tempo real. O ambiente possibilita a gravação de áudio e sua visualização e análise (como será descrito nas secções seguintes) em tempo real. Também é possível a reprodução total ou de segmentos de áudio de um ficheiro de áudio ou de uma gravação efectuada.

A funcionalidade de reprodução e gravação estão implementadas recorrendo à biblioteca *RtAudio* que interage com a placa de som. As definições de configuração da placa de som estão disponíveis nas opções da aplicação (ver secção Opções).

Para utilização de ficheiros WAV e ACC foi necessário implementar a leitura de ficheiros WAV e ACC, sendo necessário passá-los para a memória, ficando também num ficheiro temporário para ser utilizado na aplicação.

3.5 Visualização Temporal do sinal de áudio

A visualização temporal do sinal permite a visualização do sinal de áudio ao longo do tempo. No ambiente desenvolvido existem duas representações temporais com funcionalidades distintas: “Representação global do sinal” e a “Representação do sinal seleccionado“.

A “Representação global do sinal”, representado na Figura 3.3, tem sempre presente a representação total do sinal de áudio ao longo do tempo, permitindo definir os limites da “Representação do sinal seleccionado“. Assim, nesta representação, estarão sempre indicados os limites da “Representação do sinal seleccionado“. Para tal, basta seleccionar a secção desejada sobre a "Representação global", deslocando horizontalmente o rectângulo de selecção, ou então movendo as margens esquerda ou direita do rectângulo de selecção.

Figura 3.3: Representação global do sinal

A "Representação do sinal seleccionado", ilustrado na Figura 3.4, permite visualizar em maior detalhe o sinal de áudio, permitindo detalhe ao longo do tempo assim como na amplitude do sinal de áudio. Os limites temporais de representação desta visualização determinam os limites de visualização para as diferentes formas de análises que serão tratadas em seguida.

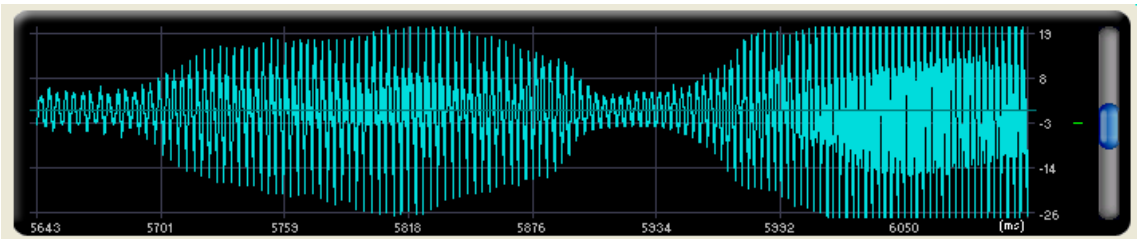


Figura 3.4: Representação do sinal seleccionado (sinal ampliado)

Nesta representação é ainda possível redefinir a secção seleccionada do sinal de forma expedita com o rato, no sentido da ampliação ou da redução. Para ampliar deve-se premir o botão esquerdo do rato, arrastar e largar de modo a enquadrar a região desejada. Para diminuir a ampliação, basta clicar com o botão direito do rato.

Na implementação destas representações utilizam-se as classes *QPainter* (biblioteca *Qt*) e *QwtPlot* (biblioteca *Qwt*). A interacção com o utilizador a partir do rato deve-se ao atendimento dos eventos de movimento do rato, clique e largar.

3.6 Análises

Uma das grandes vantagens da utilização do formato digital de áudio deve-se à possibilidade do processamento desse sinal digital para análise das características do sinal (ver capítulo Análise Acústica).

Assim as amostras de áudio após serem obtidas mediante gravação ou abertura de um ficheiro de áudio são representadas na visualização temporal. A secção de análises transforma o sinal de áudio permitindo analisar as suas características. Esta secção está composta por várias análises com o intuito de permitir diferentes iterações com espectrografia, assim como com os parâmetros qualitativos da voz.

As análises existentes na aplicação desenvolvida são:

- Dois Espectrogramas, que contém dois espectrogramas: espectrograma de banda larga e banda estreita.
- Espectrograma com traçado de *pitch*, que para além de possuir um espectrograma de banda estreita, é representado o traçado de *pitch* sobre o espectrograma.
- Espectro e Espectrograma, que contém a representação do espectrograma de banda estreita, a representação do espectro de determinado índice do espectrograma, e a representação da localização dos harmónicos.

- Cepstrograma, que caracteriza a periodicidade existente no espectro ao longo do tempo.
- Vozeamento, com a representação temporal do sinal de áudio segmentada em regiões vozeadas, não vozeadas e de silêncio permite obter medição dos parâmetros de qualitativos da voz.

3.6.1 Dois Espectrogramas

Na análise do sinal acústico, no domínio das frequências, é vulgar utilizarem-se espectrogramas de banda larga e banda estreita. Com o espectrograma de banda estreita obtém-se uma melhor resolução a nível das frequências evidenciando a existência de harmónicos. Com o espectrograma de banda larga consegue-se uma melhor resolução temporal, evidenciando a existência dos formantes.

Nesta análise, o espectrograma de banda larga utiliza 256 amostras para obtenção do espectro, enquanto que o espectrograma de banda estreita utiliza 1024 amostras. Para estes valores, usando uma frequência de amostragem de 22050Hz obtém-se as resoluções temporais e de frequência indicadas na Tabela 1.

Tipo de Espectrograma	Amostras	Resolução Frequência	Resolução Temporal
Banda Larga	256	86,13 Hz	11,6 ms
Banda Estreita	1024	21,53 Hz	46,2 ms

Tabela 1: Resolução temporal e no domínio das frequências

O processamento do sinal nesta análise descreve-se na utilização sequencial de 256 ou 1024 amostras do sinal com sobreposição de 50% aplicando uma das possíveis janelas: rectangular, triangular, seno, hamming, hanning e blackman [12] (a definição da janela a utilizar está descrita na secção de Opções). Em seguida é aplicada a transformada de Fourier (FFT), obtendo-se o espectro para os 11,6 ms ou 46,2 ms de tempo, respectivamente. A sequência de espectros constitui o espectrograma.

A representação gráfica da análise Dois Espectrogramas foi implementada utilizando a classe *QwtSpectrogram* disponibilizada pela biblioteca *Qwt*, facilitando a implementação. Nesta representação a magnitude do sinal está associada a um mapa de cores, o que possibilita a diferenciação das magnitudes ao longo do espectro. Este mapa de cores pode ser alterado permitindo utilizar o que melhor evidencie os parâmetros pretendidos ou o com melhor aspecto. É também possível definir os limites do mapa de cores, evidenciando melhor uma certa gama de magnitudes. A ampliação do espectrograma é permitida para facilitar a visualização em detalhe duma determinada gama de frequências.

Na Figura 3.5 encontra-se representada a análise Dois Espectrogramas com a representação do espectrograma de banda estreita e larga.

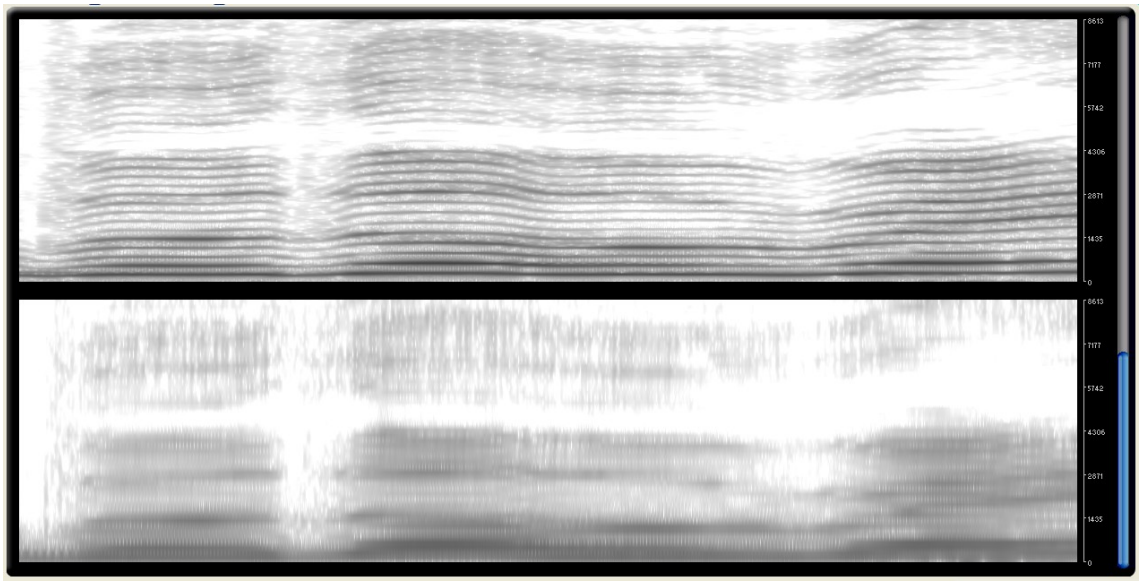


Figura 3.5: Representação da análise Dois Espectrogramas

3.6.2 Espectrograma com Traçado de Pitch

A representação do traçado de *pitch* sobre o espectrograma informa sobre o valor de *pitch* na análise do espectrograma. O tipo de espectrograma utilizado é o espectrograma de banda estreita (ver secção Dois Espectrogramas), sendo utilizado apenas a janela de seno, devido à utilização para algoritmo de cálculo do *pitch*. A representação resultante está exemplificada na Figura 3.6. O traçado de *pitch* é calculado tendo em conta o espectro obtido para cada índice da resolução temporal do espectrograma.

Esta representação à semelhança da representação do espectrograma de banda larga e banda estreita recorre ao uso da classe *QwtSpectrogram* da biblioteca *Qwt*, permitindo as mesmas funcionalidades.

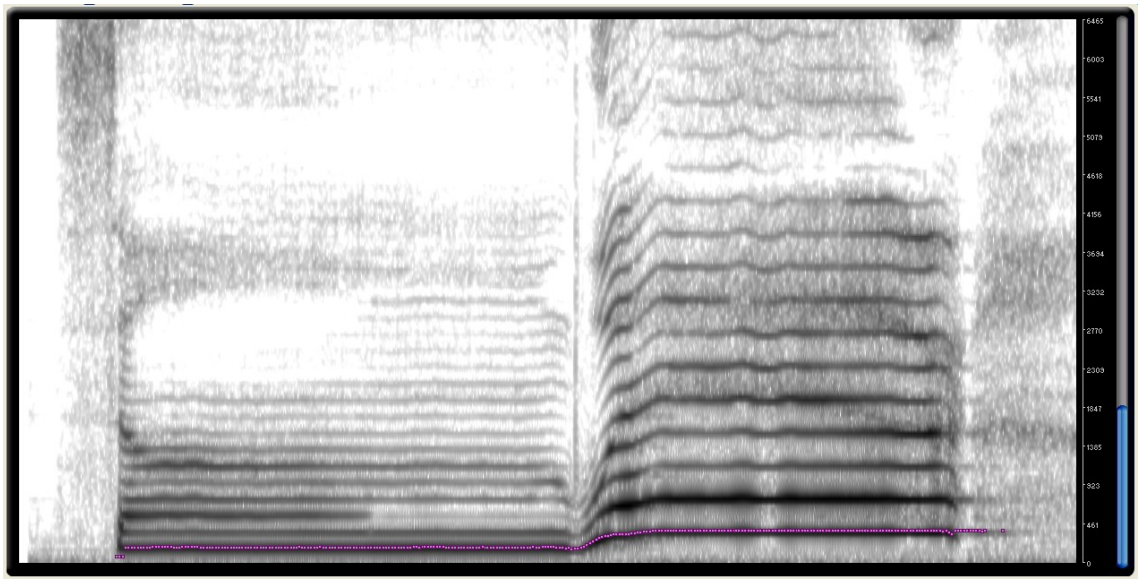


Figura 3.6: Representação de Espectrograma com traçado de pitch

3.6.3 Espectro e Espectrograma

A representação do espectro e espectrograma, exemplificada na Figura 3.7. Permite uma análise mais detalhada do espectro de frequências, sendo identificados os harmónicos do sinal acústico, representado o ruído do sinal (parte não harmónica) e sintetizada a voz correspondente à parte harmónica e de ruído separadamente.

O tipo de espectrograma utilizado nesta análise é o espectrograma de banda estreita (melhor resolução de frequência) sendo utilizada a janela de seno. A utilização da janela de seno garante a compatibilidade com o processamento de identificação de harmónicos e separação física da componente harmónica da componente de ruído.

A identificação dos harmónicos está relacionado com o algoritmo de cálculo de *pitch*, no qual a frequência de cada harmónico é múltiplo inteiro do valor de *pitch*. Estando identificados os harmónicos, é efectuada a síntese do sinal a partir da magnitude e fase dos harmónicos, resultando um sinal de voz com a componente harmónica do sinal. A síntese da componente não harmónica provém da diferença do espectro do sinal com o espectro da componente harmónica.

Na representação desta análise foi utilizada a biblioteca Qwt, usando a classe *QwtPlot* para o desenho do espectro e utilizando a classe *QwtSpectrogram* para o desenho do espectrograma. O espectro representado corresponde a um instante no espectrograma que é indicado utilizando uma barra de apontador com deslocamento horizontal (classe QSlider do Qt).

Na representação do espectro para além do desenho do espectro assinalam-se também os harmónicos, utilizando setas vermelhas, e representa-se a componente espectral de ruído. A separação da componente harmónica da componente não harmónica verifica-se na reprodução destas componentes ou na gravação em ficheiros de áudio, utilizando-se para tal os botões da barra de ferramentas da análise.

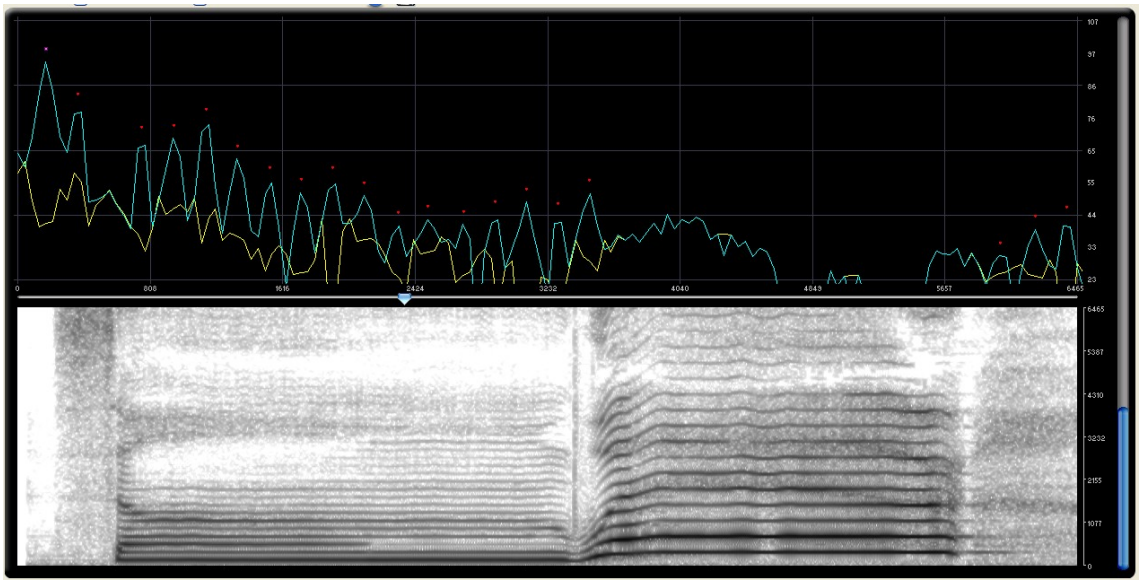


Figura 3.7: Representação do Espectro e Espectrograma

3.6.4 Cepstrograma

O cepstrum (cepstrum real) consiste na transformada de Fourier inversa do logaritmo do espectro. Remete portanto para um domínio do tempo que caracteriza a periodicidade existente no espectro. Assim o cepstrograma, exemplificado na Figura 3.8, permite para avaliar a evolução ao longo do tempo do período fundamental.

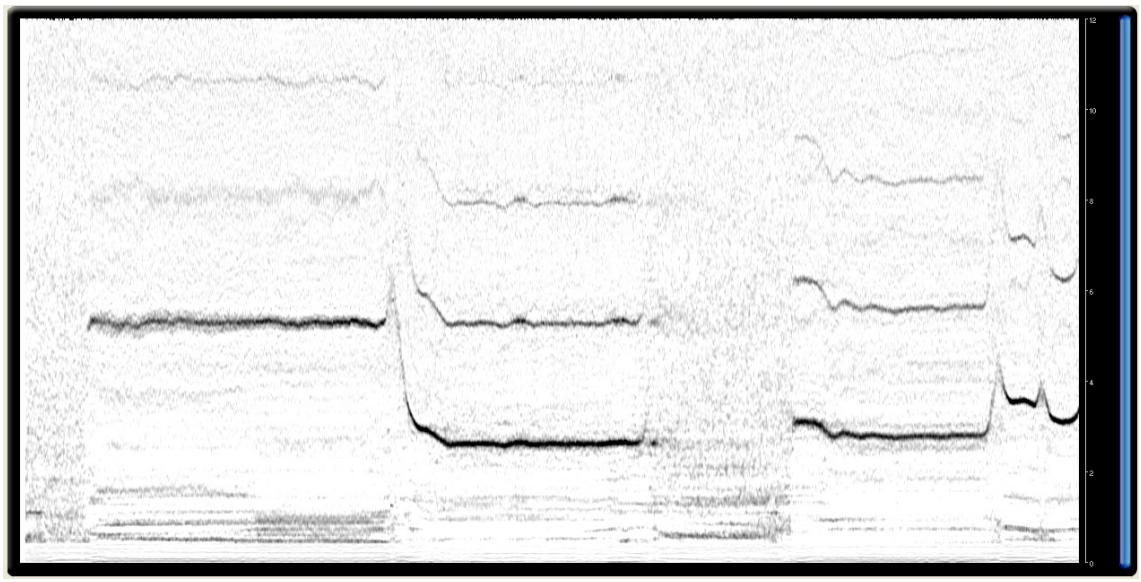


Figura 3.8: Representação do Cepstrograma

3.6.5 Vozeamento

O desenvolvimento da análise vozeamento teve uma participação mais intensiva nesta dissertação, pois aqui é possível a obtenção dos principais parâmetros qualitativos da voz.

Esta análise caracteriza-se pela representação temporal do sinal de áudio, sendo assinaladas, sobre a representação, indicações sobre a segmentação de regiões vozeadas, não vozeadas e de silêncio, indicação das marcas de *pitch*, envolvente de tempo longo e envolvente de tempo curto. Sobre as zonas vozeadas é possível serem obtidas estatísticas, tais como os principais parâmetros qualitativos.

Para a implementação gráfica desta análise utilizou-se a classe *QPainter* da biblioteca Qt, que permite um grande controle em tudo que é desenhado.

3.6.5.1 Segmentação de regiões

A diferenciação entre regiões permite identificar rapidamente as regiões vozeadas, não vozeadas e de silêncio. A diferenciação entre estas regiões é realizada automaticamente e apresentada utilizando cores que preenchem as zonas distintas, como representado na Figura 3.9.

As regiões de silêncio são caracterizadas por serem regiões do sinal de áudio onde a energia do sinal é baixa, sendo utilizado um limiar máximo abaixo do qual é considerada região de silêncio.

As regiões vozeadas são caracterizadas por possuírem uma estrutura harmónica, que é identificada pelo algoritmo de detecção de *pitch*.

As regiões não vozeadas são regiões do sinal de áudio com um nível de

energia superior ao limite máximo utilizado pelas regiões de silêncio, e regiões em que não se encontra uma estrutura harmónica.

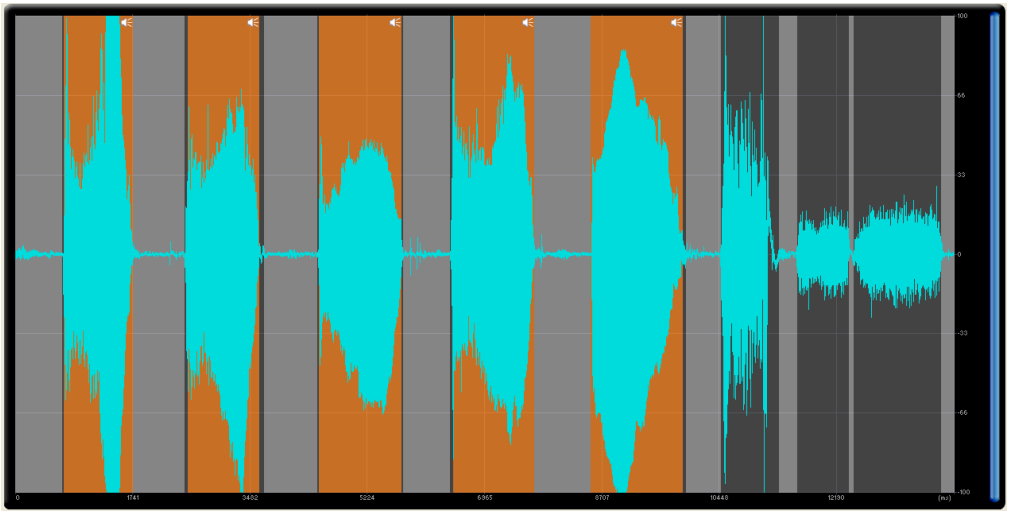


Figura 3.9: Representação de segmentação de regiões

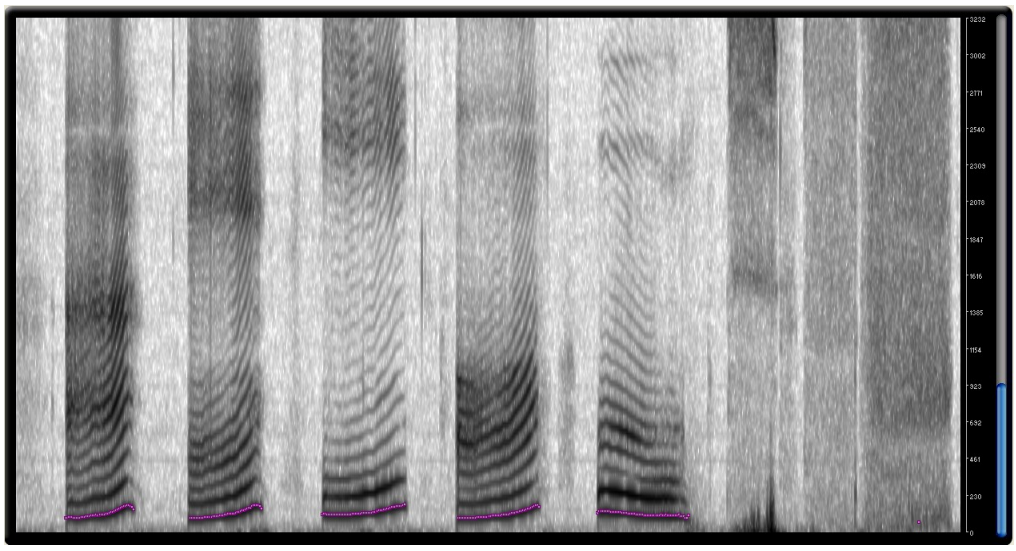


Figura 3.10: Representação do Espectrograma do ficheiro de áudio

Como se pode apreciar a partir das Figuras 3.9 e 3.10 em que é utilizada o mesmo registo de áudio, consegue-se visualizar a existência de regiões vozeadas não vozeadas e de silêncio. Verifica-se que apenas as zonas com uma estrutura harmónica são consideradas como zonas vozeadas.

3.6.5.2 Envolvente de energia de tempo longo

A envolvente de energia de tempo longo corresponde à energia média do sinal numa zona de sinal de áudio de 1024 amostras com sobreposição de 50%. O valor da energia média resulta do valor médio dos valores da magnitude do

espectro do sinal nas 1024 amostras. Assim o gráfico da envolvente energia de tempo longo corresponde a um sinal com resolução de 512 amostras. Na Figura 3.11 está destacada uma linha descontinua sobre o sinal temporal, que representa a envolvente de energia de tempo curto desse sinal.

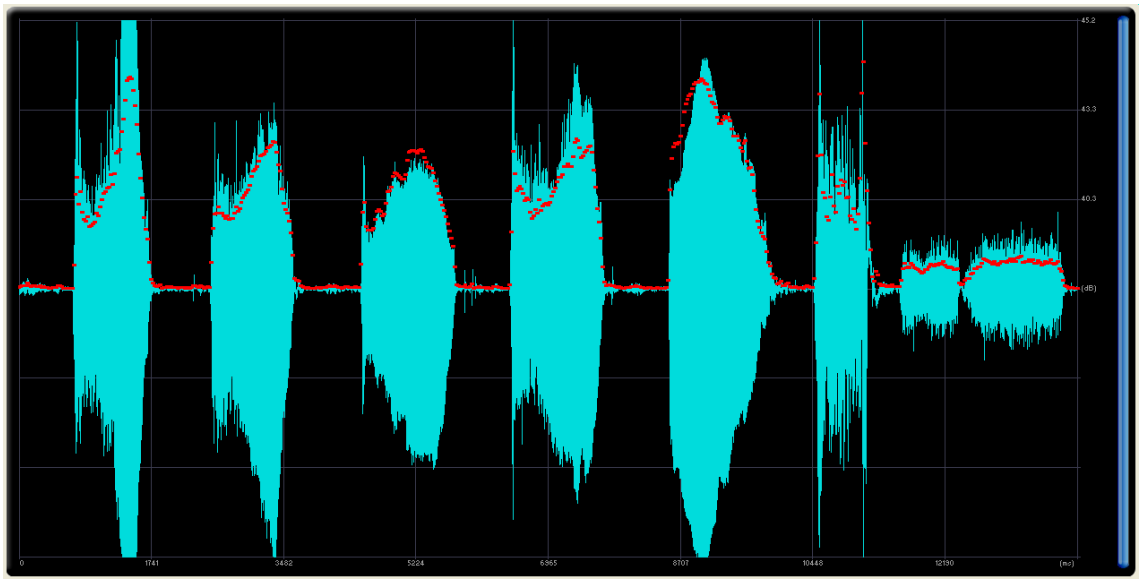


Figura 3.11: Representação da envolvente de energia de tempo longo

3.6.5.3 Envolvente de energia de tempo curto

A envolvente de tempo curto é resultado da utilização da transformada de Hilbert, que permite obter a componente imaginária de um sinal (com uma rotação ideal de fase de π/2 [12]), sendo efectuado posteriormente o cálculo de energia através das componentes real e imaginária.

A transformada de Hilbert de uma função *f* é descrita pela equação (3.1).

$$F(t)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{f(x)}{t-x}dx \Leftrightarrow F(t)=\frac{1}{\pi t}*f(t) \tag{3.1}$$

A transformada de Hilbert é descrito pela equação (3.2), estando ilustrado na Figura 3.12 um exemplo de concretização da equação (3.2) para um comprimento de 127 amostras.

$$h[n]=\begin{cases} 0, \textit{para n par} \\ \dfrac{2}{\pi n}, \textit{para n impar} \end{cases} \tag{3.2}$$

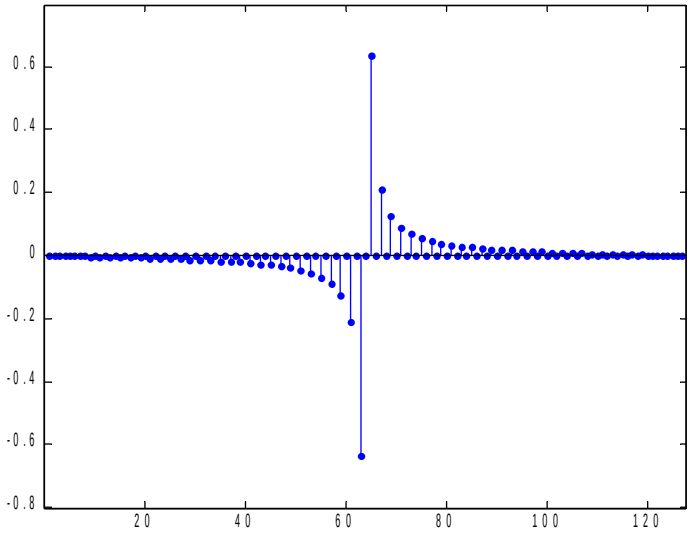


Figura 3.12: Representação discreta da transformada de hilbert

A transformada de Hilbert não afecta a magnitude espectral do sinal, afectando apenas a fase com 90º. Assim a transformada de um sinal é ortogonal ao sinal, possuindo a mesma energia.

Para calcular a envolvente de energia do sinal basta calcular o valor absoluto do sinal complexo composto pelo sinal real (sinal original) e a transformada de Hilbert desse sinal (componente imaginária), descrita na equação (3.3).

$$Envolvente = \sqrt{(h[n] * f(t))^2 + f(t)^2} \tag{3.3}$$

Na Figura 3.13 está representada uma imagem obtida a partir da análise Vozeamento, em que a onda mais clara corresponde ao sinal de voz e a onda mais escura corresponde à envolvente de energia de tempo curto.

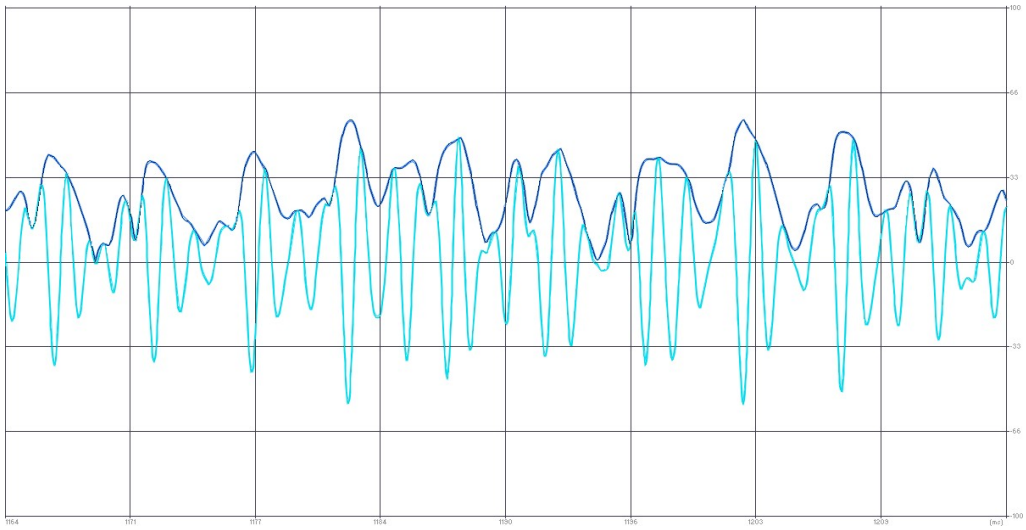


Figura 3.13: Representação da envolvente de energia de tempo curto

3.6.5.4 Indicação das Marcas de Pitch

Para o cálculo dos parâmetros acústicos de curto termo há a necessidade de identificar os ciclos do sinal de voz. A identificação das marcas de *pitch* corresponde à indicação da ocorrência de cada ciclo glotal, isto é, a indicação da ocorrência do impulso glotal. Estas marcas são indicadas sobre a representação temporal do vozeamento.

O processo de detecção de marcas de *pitch* relaciona-se com a detecção do valor de *pitch* para um frame de 1024 amostras. Nesta frame é procurada a primeira marca de *pitch*, que caso seja a primeira frame de uma região vozeada, é determinada usando um algoritmo que determina a primeira marca de *pitch* mais provável, identificando a sequência de marcas mais provável. Caso seja não seja a primeira frame, é utilizada a ultima marca de *pitch* da frame anterior para determinar a seguinte.

O inverso valor de *pitch* (período fundamental) é utilizado para determinar a posição das restantes marcas de *pitch*, que estão mais ou menos distanciadas pelo valor do período fundamental. Para cada marca é efectuada uma sintonia relativamente ao sinal para se encontrar na posição correcta.

No ambiente desenvolvido, esta procura de marcas de *pitch* no sinal está implementada utilizando dois métodos. O primeiro método consiste na utilização dos máximos do sinal de áudio para a indicação das marcas de *pitch*. No segundo método são utilizados os máximos da envolvente de energia de tempo curto.

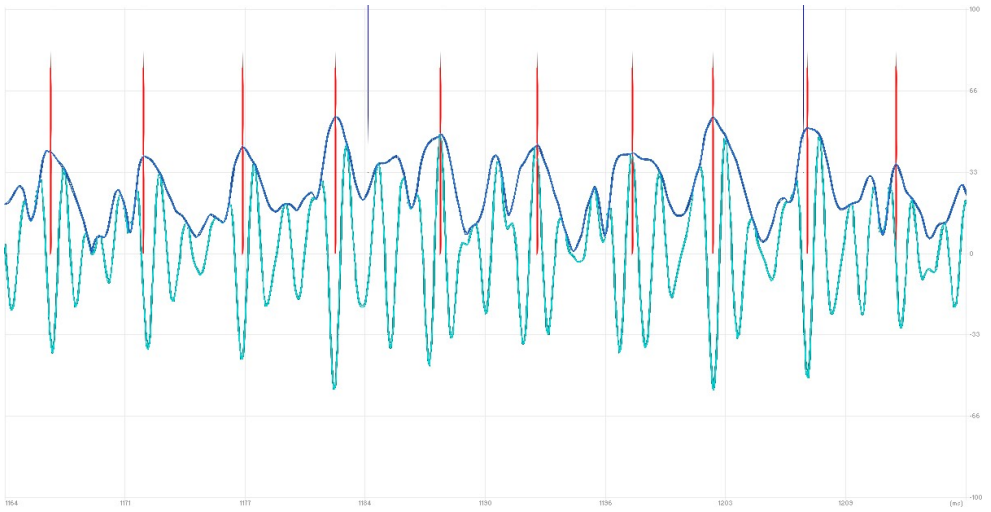


Figura 3.14: Representação das marcas de pitch

Comparando os métodos, verifica-se que o método que usa os máximos da envolvente de tempo de curto é mais robusto do que o método que utiliza os máximos do sinal porque não é vulnerável a erros introduzidos devido à inversão do sinal, isto é, em certas ocasiões (no método que usa máximos do sinal) verifica-se uma inversão no sinal, pelo que o impulso mais saliente é negativo, não sendo detectado marca de *pitch*, mas sim o impulso mais próximo positivo (na sua utilização para medição do *jitter*, o erro propaga-se). No método que usa os máximos da envolvente de energia de tempo curto do sinal isto não acontece, verificando-se na Figura 3.14. Nesta figura é apresentado o método que utiliza os máximos da envolvente de energia de tempo curto, e na 5ª marca de *pitch* visualizada (as marcas de *pitch* correspondem às linhas verticais destacadas sobre o sinal), verifica-se a robustez do algoritmo face à inversão do sinal.

3.6.5.5 Estatísticas

As estatísticas sobre um sinal vozeado são importantes na obtenção dos parâmetros acústicos determinísticos. No ambiente desenvolvido estão disponíveis medidas de F0 (valor médio, máximo e mínimo, e desvio padrão de F0), medidas de energia (valor médio, máximo e mínimo e desvio padrão de energia), parâmetros de qualidade (*jitter*, *shimmer* e HNR), tempo de fonação e extensão vocal.

As estatísticas das medidas de F0 são determinadas a partir do valor de *pitch* calculado pelo algoritmo desenvolvido pela empresa SEEGNAL Research. Nas medidas de energia é utilizada a envolvente de energia tempo longo. Os parâmetros de qualidade servem-se das marcas de *pitch* para calcular o valor de *jitter* e *shimmer*. Quanto ao valor de HNR é obtido utilizando a separação entre componentes harmónica e de ruído (ver secção HNR no capítulo de Análise Acústica).

A obtenção de estatísticas relativas a uma região vozeada é efectuada clicando com o botão direito do rato sobre a região e escolhendo a opção "Estatísticas". A representação das estatísticas está dividida em três partes: representação gráfica dos valores mais importantes, listagem das medidas, botões de acção.

A representação gráfica das estatísticas, ilustrada na Figura 3.15, contém um gráfico para representar visualmente todas as medidas de F0. Esse gráfico tem ao lado duas escalas, a escala de frequência que permite situar as medidas de F0 e a escala de indicação do género mais provável ou grupo etário (entre criança e adulto). Esta última escala baseia-se no facto do valor de F0 ser variável com o sexo e idade (ver capitulo Análise Acústica), em que se utilizam os valores predominantes: 100Hz para o homem adulto, 200Hz para a mulher adulta e 300Hz para crianças. Estes valores estão definidos por omissão, pelo que podem ser alterados para uma melhor sintonização do utilizador. A gradação de cores é implementada usando a classe *QLinearGradient* da biblioteca *Qt*.

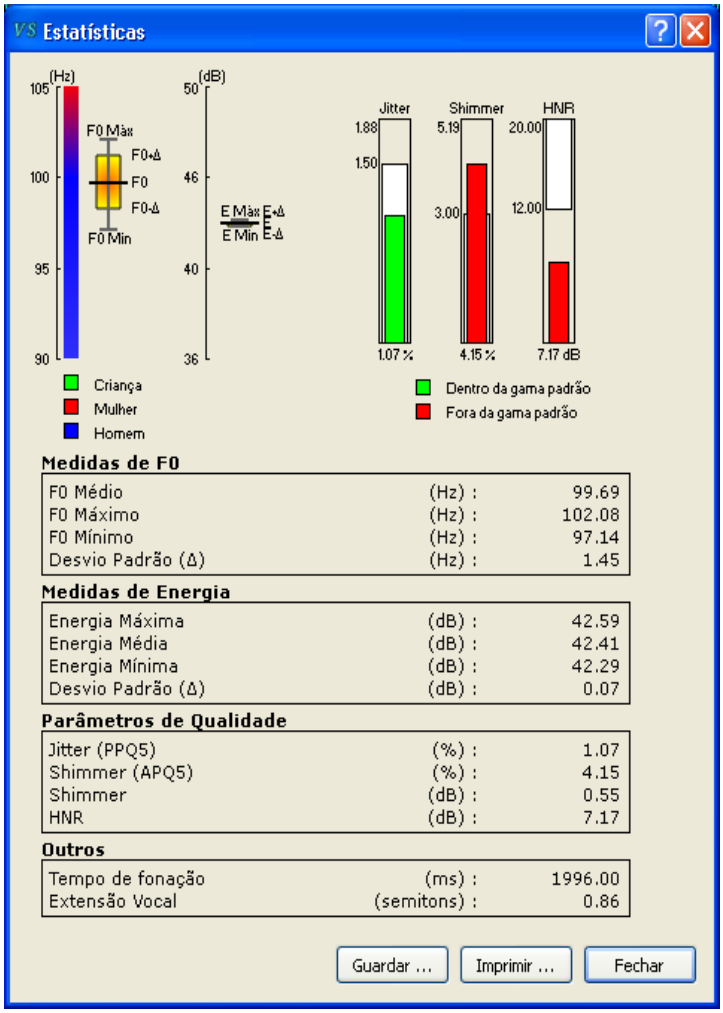


Figura 3.15: Caixa de diálogo com estatísticas de segmento vozeado

Na representação gráfica das estatísticas encontram-se também os valores de medição de energia no mesmo esquema de representação das medidas de F0. Também são apresentados os parâmetros de medidas de qualidade utilizando gráficos de barras. Estes gráficos de barras têm a indicação das gamas padrão foras das quais é provável a existência de alguma patologia vocal.

Na listagem das estatísticas está representada toda a informação determinística sobre as medidas. Nos parâmetros *jitter* e *shimmer* é indicado o método de cálculo utilizado.

Os botões de acção permitem exportar os dados obtidos para um imagem, usando o botão "Guardar..." ou imprimindo a informação, usando o botão "Imprimir...". A imagem guardada a partir das estatísticas é exactamente igual à apresentada na caixa de diálogo. Quanto à impressão é gerada uma imagem com informação do utilizador, representação da análise vozeamento e as estatísticas apresentadas (ver anexo 1). A impressão é efectuada a 600dpi permitindo uma boa resolução na página impressa.

Na implementação da impressão foi utilizada a classe *QPainter* da biblioteca *Qt* para permitir a criação da imagem de impressão, e a classe *QPrinter* da biblioteca *Qt* para lidar com a interface de impressão.

3.7 Opções

A possibilidade de configuração de qualquer aplicação é necessária para suportar diversas condições que os utilizadores possam desejar, tornando a aplicação personalizável.

No ambiente desenvolvido são disponibilizadas as opções de configuração da placa de som, dos espectrogramas, da análise vozeamento e da impressão.

A implementação das opções teve como base a utilização da classe *QSettings* da biblioteca *Qt*, que permite guardar as configurações (ficheiro de configurações INI ou registo do Windows) e ser utilizado nos restantes módulos da aplicação.

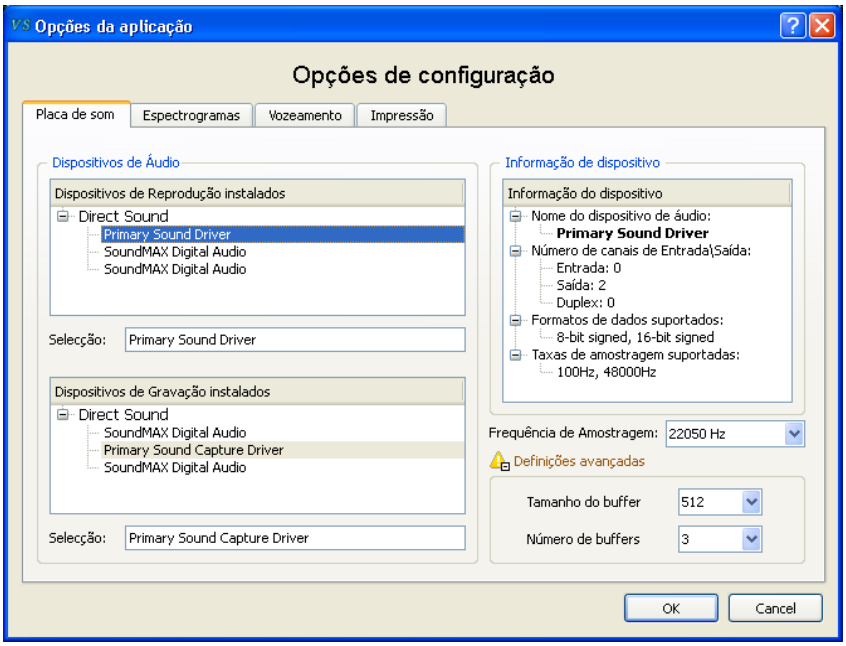


Figura 3.16: Caixa de diálogo de Opções, separador Placa de som

Nas opções da configuração da placa de som (ver Figura 3.16) é possível seleccionar o dispositivo de entrada e de saída de áudio, permitindo a visualização de informação relativamente a cada dispositivo de áudio. É também possível seleccionar a frequência de amostragem utilizada para gravação de áudio, em que as frequências de amostragem disponíveis resulta das frequências de amostragem suportadas entre o dispositivo de entrada e de saída. As opções avançadas permitem especificar o tamanho do buffer e o número de *buffers*, o que se torna útil para configuração quando se verificar a existência de anomalias na gravação ou reprodução (existência de intermitências ou cliques).

Na implementação das opções da placa de som utilizou-se a biblioteca *RtAudio*, que permite obter informação acerca dos dispositivos de áudio, assim como visualizar as *API* (*Application Programming Interface*) disponíveis na placa de som. Com a utilização desta biblioteca é bastante fácil a utilização da aplicação em diferentes plataformas, no entanto a aplicação apenas foi desenvolvida para o sistema operativo Windows e portanto apenas suporta as *APIs DirectSound* e *ASIO* (Audio Stream Input/Output).

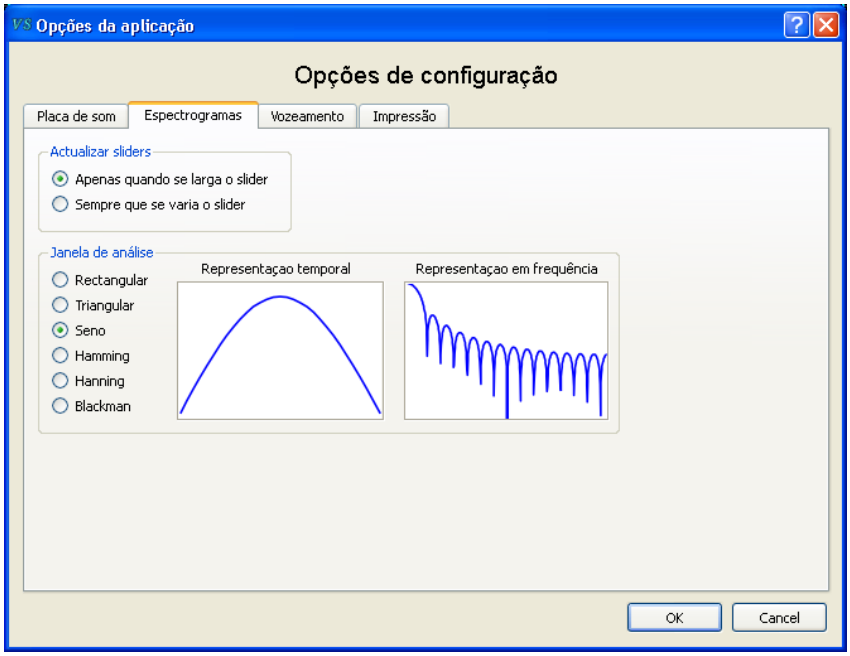


Figura 3.17: Caixa de diálogo Opções, separador Espectrogramas

Nas opções de configuração dos espectrogramas (ver Figura 3.17) é possível seleccionar a janela de análise utilizada para o cálculo dos espectrogramas, na análise de Espectrograma de banda larga e banda estreita. Ainda é possível definir a actualização dos espectrogramas pela activação da barra vertical de selecção (*sliders)*, isto é, no processo de ampliação ou redução da gama de frequências visível de qualquer espectrograma, permitir ou não a sua actualização sempre que variar o *slider* ou apenas quando se larga o *slider.* Esta configuração melhora a resposta de aplicação quando seleccionada para actualização quando se larga o slider, em situações que o computador esteja em sobrecarga.

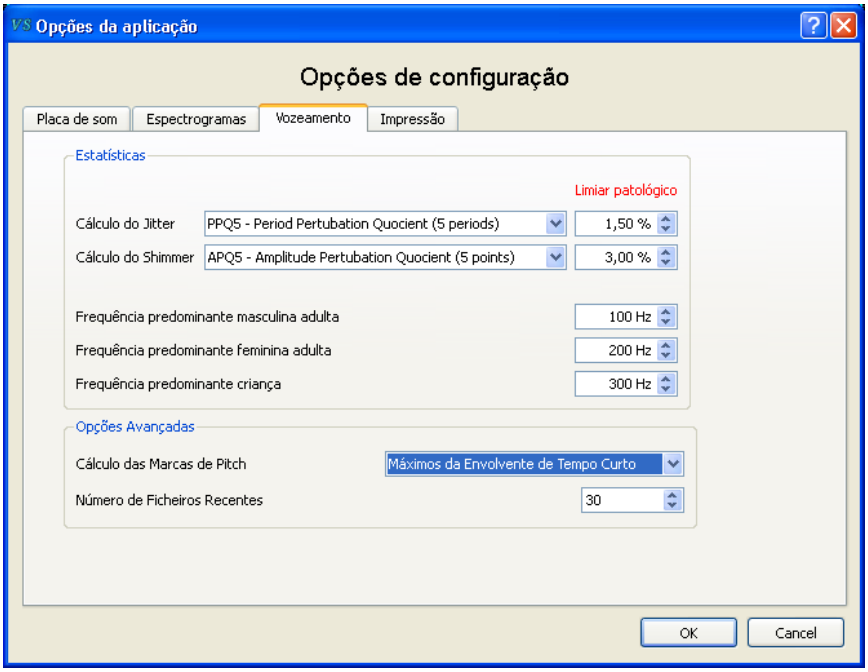


Figura 3.18: Caixa de diálogo Opções, separador Vozeamento

As opções de configurações de vozeamento (ver Figura 3.18) são constituídas pelas opções das estatísticas e pelas opções avançadas.

Nas opções das estatísticas é possível indicar qual o método de cálculo para o *jitter* e *shimmer*. Para o *jitter* estão disponíveis os métodos: *jitter* local, RAP, PPQ5, PPQ11. Para o *shimmer* estão disponíveis: *shimmer* local, APQ3, APQ5, APQ11. Para cada um destes métodos é possível definir o limiar patológico. Esta solução é necessária devido à inexistência de consenso relativamente ao limiar normativo. Por omissão estão definidos os valores os métodos seleccionados são o PPQ5 e APQ5 por apresentar em boa imunidade face a pequenas irregularidades. Os limiares patológicos são 1.5% e 3.0% respectivamente.

As frequências predominantes para classificação da voz segundo o género mais provável e idade podem de ser definidas nas opções das estatísticas.

Nas opções avançadas é possível definir o método de cálculo das marcas de *pitch*, podendo-se optar pelos “Máximos da envolvente de Tempo Curto” ou pelos “Máximos do sinal”.

Outra opção avançada é a definição do número de ficheiros recentes, isto é, a análise vozeamento guarda a informação de segmentação de zonas vozeadas, não vozeadas e de silêncio, permitindo a manutenção da segmentação (caso seja alterada) na próxima execução . Portanto o número de ficheiros recentes corresponde ao número de ficheiros recentemente abertos em que a aplicação ainda está a suportar a manutenção da segmentação. Por omissão o número de ficheiros recentes é 30.

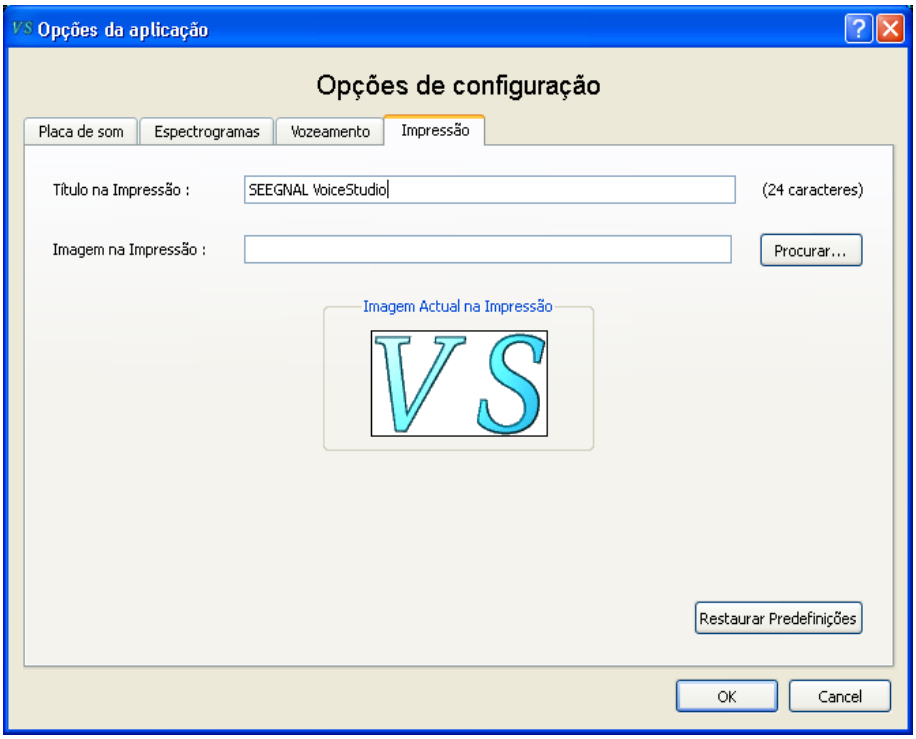


Figura 3.19: Caixa de diálogo Opções, separador Impressão

Nas opções de impressão (ver Figura 3.19) é possível personalizar a folha impressa de modo a aparecer a indicação (nome e logotipo) da clínica ou centro académico que utilize a aplicação.

3.8 Gestão de Pacientes

Em ambiente clínico existe a necessidade de organização da informação acerca das sessões dos pacientes. Com esse intuito, a gestão de pacientes permite gerir a informação do paciente, assim como agrupar os ficheiros de áudio em sessões, disponibilizando a criação de notas de sessão.

Na implementação da gestão de pacientes, a informação inserida é guardada numa base de dados do tipo SQLite, que permite através de comunicação SQL guardar informação de forma organizada num ficheiro. Para interagir com o ficheiro SQLite utilizou-se a biblioteca *Qt*, que utiliza o *plugin sqlite*.

A interface gráfica da gestão de pacientes está representada na Figura 3.20. Na interface é visível uma listagem dos pacientes, sendo possível adicionar, editar, remover e pesquisar pacientes. Cada paciente pode ter várias sessões e cada sessão poderá ter vários ficheiros de áudio. Para cada sessão é possível adicionar notas de informação, podendo-se utilizar

formatação para realçar partes do texto.

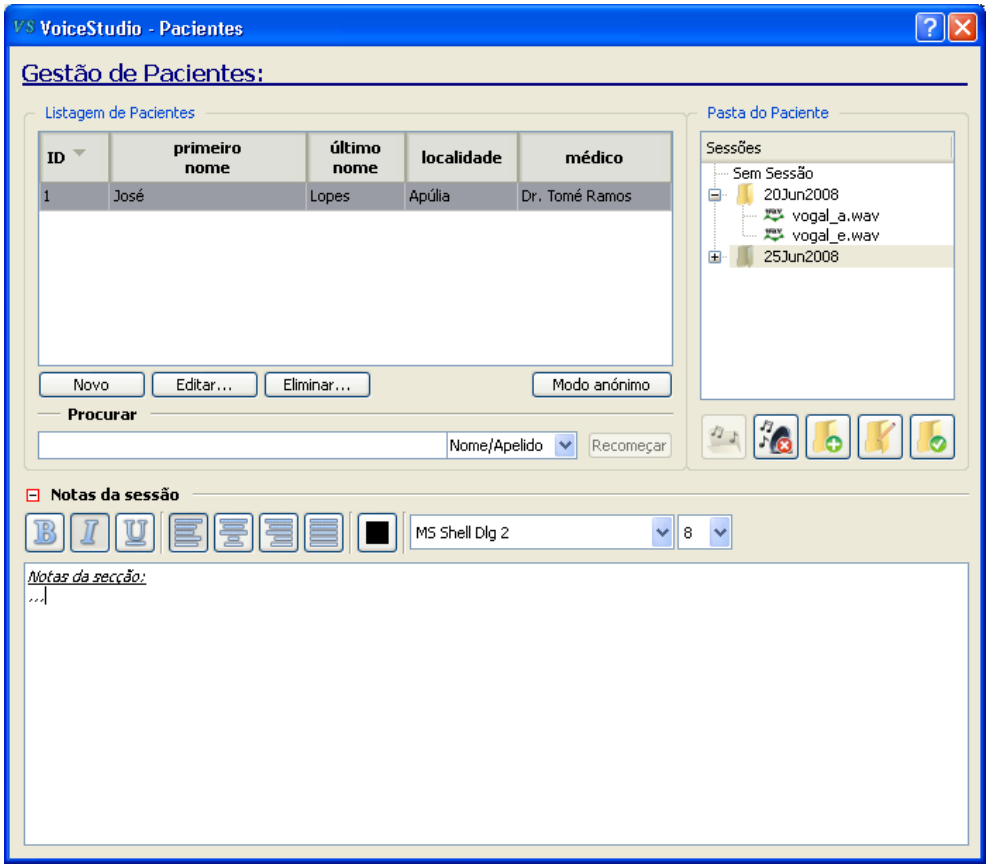


Figura 3.20: Caixa de diálogo de gestão de pacientes

A gestão de pacientes permite utilizar uma determinada sessão de um determinado utilizador para utilização na aplicação, em que todos os ficheiros gravados ficam incluídos na sessão seleccionada. Para não utilizar o modo sessão, basta utilizar o modo anónimo.

3.9 Ajuda

O sistema de ajuda é um meio fundamental em qualquer aplicação permitindo ao utilizador a auto-aprendizagem de todas as potencialidades da aplicação, assim como ser o primeiro passo de interacção em caso de dúvidas.

Na implementação do sistema de ajuda utilizou-se o código fonte do assistente de ajuda do *Qt* (*Assistant*) para criar um sistema de ajuda baseado em páginas HTML (*HyperText Markup Language*), dispondo de um índice de conteúdos, assim como a possibilidade de pesquisa e impressão, entre outras funcionalidades provenientes da ferramenta do *Qt*. Na Figura 3.21 está representada a interface gráfica do sistema de ajuda.

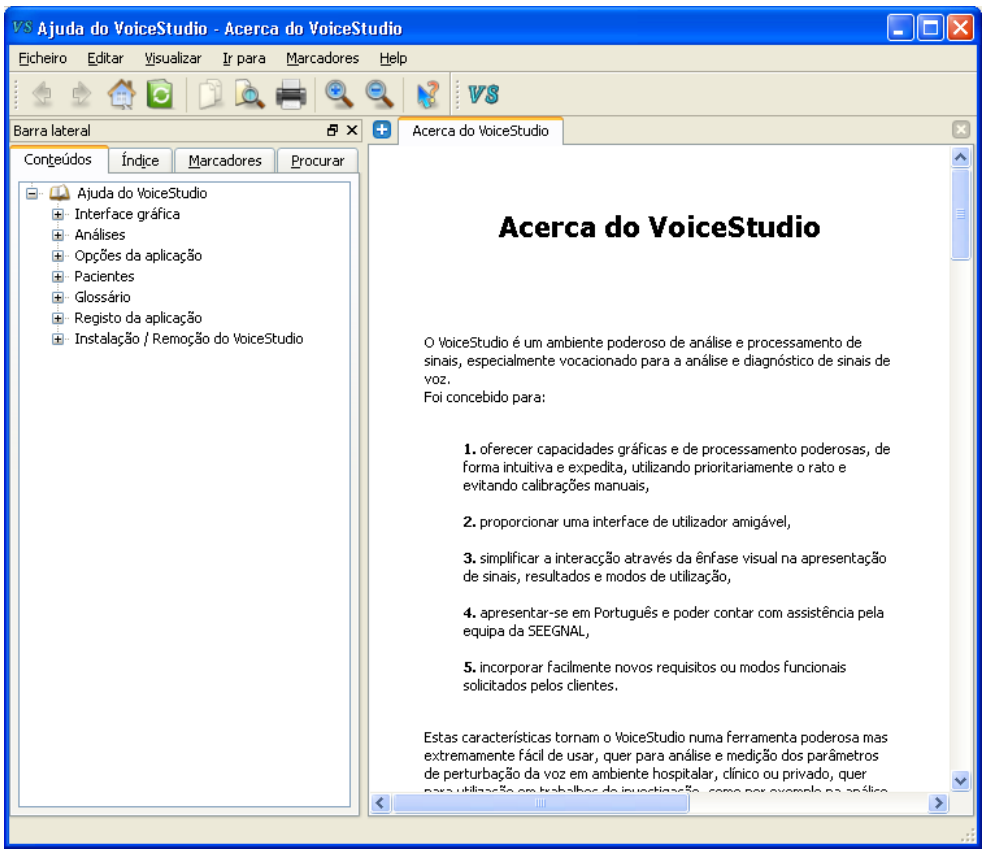


Figura 3.21: Sistema de Ajuda

Capítulo 4 Testes aos parâmetros qualitativos da voz

Para avaliar os algoritmos de cálculo de *jitter*, *shimmer* e HNR foi utilizado um modelo de síntese de voz, desenvolvido na empresa SEEGNAL Research, sendo utilizadas duas frequências fundamentais, F0=100Hz e F0=200Hz, de modo a simular uma voz masculina e outra feminina. Para cada género geraram-se 15 versões da vogal /a/ com valores de *jitter* ou *shimmer* ou HNR. Portanto 5 versões correspondem a valores predefinidos de *jitter*: 0%; 0,45%; 0,9%; 1,35%; 1,8%. Os valores predefinidos de *shimmer* são 0%; 0,9%; 1,8%; 2,7%; 3,6%. Relativamente ao HNR os valores são 5dB, 10dB, 15dB, 20dB, 25dB. Nas gamas de valores apresentadas estão incluídos valores patológicos e normais. Na avaliação dos algoritmos, utilizaram-se para referência duas aplicações conhecidas (Praat e Dr. Speech) para efectuar as mesmas medições.

4.1 Modelo de síntese de voz

Para garantir a robustez da medição dos parâmetros qualitativos foi implementado um modelo de síntese de voz com perturbações de *jitter*, *shimmer* e HNR predefinidas. Em seguida está apresentado o modelo fonte-filtro implementado:

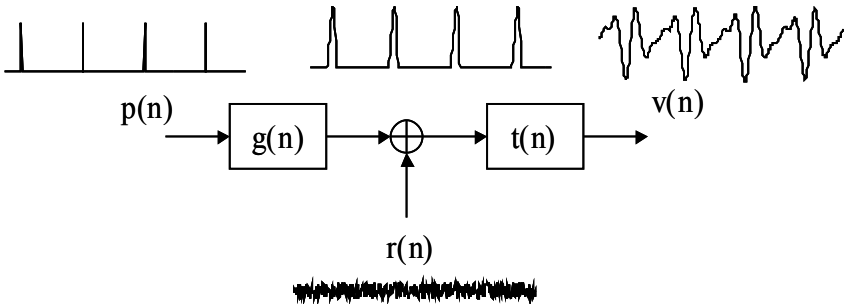


Figura 4.1: Modelo fonte-filtro de produção de fala vozeada.

A fonte ideal é representada pela sequência de impulsos p(n) de acordo com a equação (4.1).

$$p(n)=\sum_k a_k \delta(n-kT-\Delta T_k) \tag{4.1}$$

Um impulso de índice k tem amplitude a_k e é colocado no instante $kT+\Delta T_k$ em que ΔT denota um pequeno desvio em relação a T. T representa o período fundamental da vibração das pregas vocais e o seu recíproco é $F_0=1/T$ (frequência fundamental). Se a sequência de impulsos for perturbada em amplitude, a_k será distinto para cada valor de k. Por outras palavras, a sequência será afectada de *shimmer*. O *shimmer* será nulo se a_k não depender de k, ou seja, se a_k for constante.

Se a sequência de impulsos for perturbada por uma irregularidade na colocação temporal dos impulsos, o valor de ΔT_k será distinto para cada valor de k. Por outras palavras, a sequência será afectada de *jitter*. O *jitter* será nulo se ΔT_k for constante para todos os valores de k. Idealmente, $\Delta T_k=\Delta T=0$.

De acordo com o modelo fonte-filtro, a sequência ideal de impulsos p(n) é filtrada por g(n) que representa a forma de onda de um único impulso glotal. Esta operação é equivalente à convolução entre p(n) e g(n) e o resultado é uma sequência de impulsos glotais. Esta sequência é adicionada a ruído estacionário, representado por r(n), e o conjunto é filtrado por t(n) que modeliza as ressonâncias do tracto vocal. O resultado, v(n), representado na equação (4.2), consiste no sinal de voz.

$$v(n)=[p(n)*g(n)+r(n)]*t(n) \tag{4.2}$$

A vantagem de se utilizar vozes sintéticas é a possibilidade de provocar as alterações pretendidas na voz. Assim os valores de jitter e shimmer são directamente aplicados na fonte de sinal. Quanto a valor de HNR é necessário determinar o valor das componentes harmónicas e ruído.

Passando a equação (4.2) para o domínio de Fourier obtém-se a equação (4.3).

$$V(w)=[P(w)\times G(w)+R(w)]\times T(w)=H(w)+N(w) \tag{4.3}$$

A partir da representação no domínio da frequência do sinal de voz, V(w), tem descritas as duas componentes: H(w) e N(w), tal como detalhado na equação (4.4). Agora basta aplicar na equação do HNR, resultando o valor teórico de HNR.

$$\begin{aligned} H(w)&=P(w)\times G(w)\times T(w) \\ N(w)&=R(w)\times T(w) \end{aligned} \tag{4.4}$$

4.2 Resultados utilizando ficheiros de vozes sintéticas

Em seguida são apresentadas figuras contendo os resultados das medições de *jitter, shimmer* e HNR para vozes sintéticas, utilizando os programas VoiceStudio, Dr Speech e Praat.

Avaliando a medição do *jitter* (Figura 4.2, Figura 4.3, Figura 4.4) verifica-se que de forma global, o VoiceStudio apresenta um menor erro no valor do *jitter*, aproximando-se mais do valor teórico.

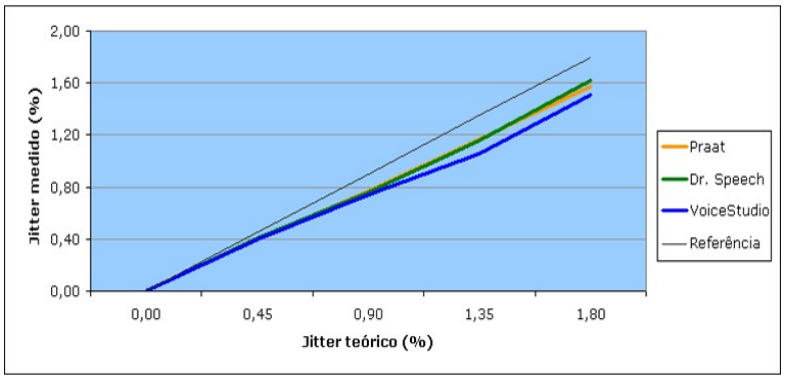


Figura 4.2: Resultado da medição de jitter em voz sintética masculina

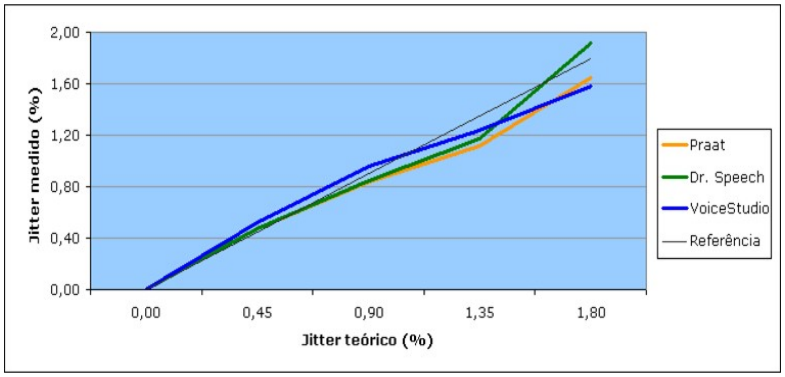


Figura 4.3: Resultado da medição de jitter em voz sintética feminina

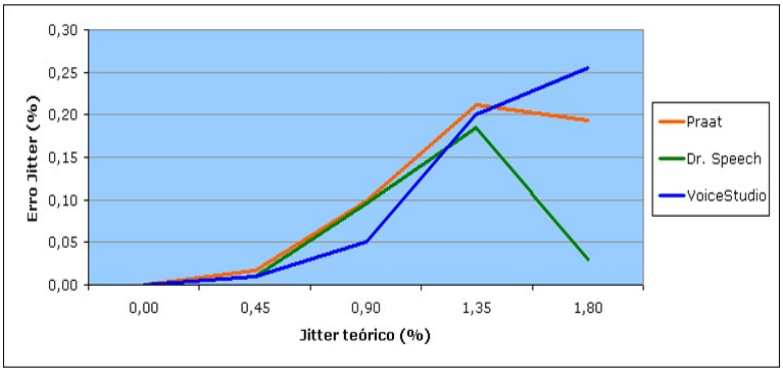


Figura 4.4 : Resultado do erro jitter em voz sintética masculina/feminina

Avaliando a medição do *shimmer* (Figura 4.5, Figura 4.6, Figura 4.7) verifica-se que o VoiceStudio apresenta um bom comportamento em relação às outras aplicações, somente para 0,9% apresenta um valor ligeiramente superior à aplicação Praat.

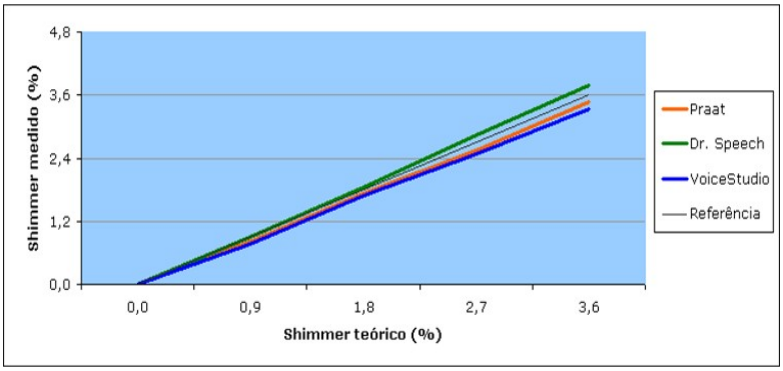


Figura 4.5: Resultado da medição de shimmer em voz sintética masculina

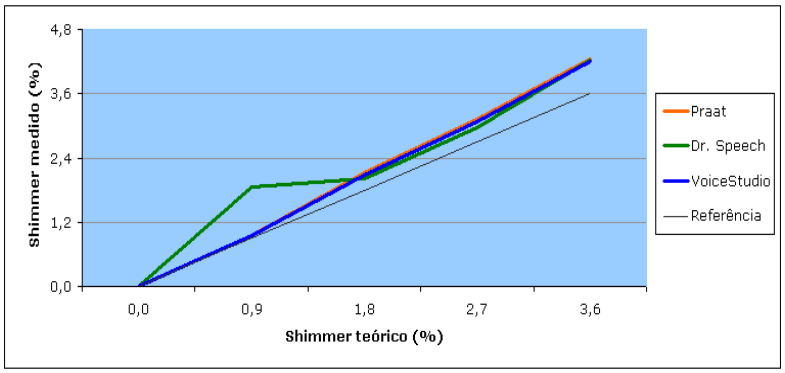


Figura 4.6: Resultado da medição de shimmer em voz sintética feminina

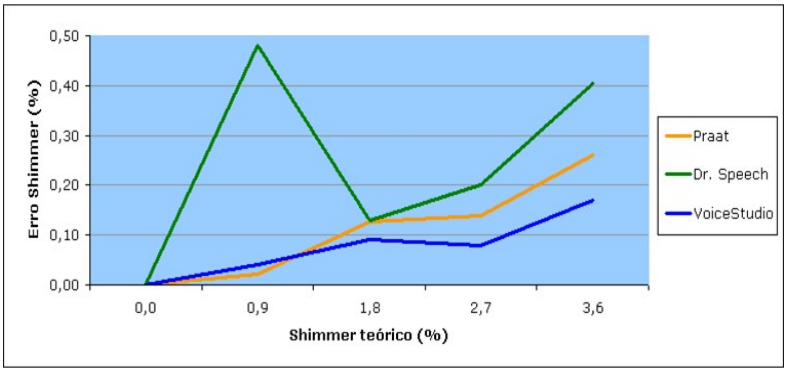


Figura 4.7: Resultado do erro shimmer em voz sintética masculina/feminina

Avaliando a medição do HNR (Figura 4.8, Figura 4.9, Figura 4.10) verifica-se um desvio acentuado das medições pela aplicação Dr. Speech, em que as medições efectuadas pelo VoiceStudio estão mais próximas do valor teórico.

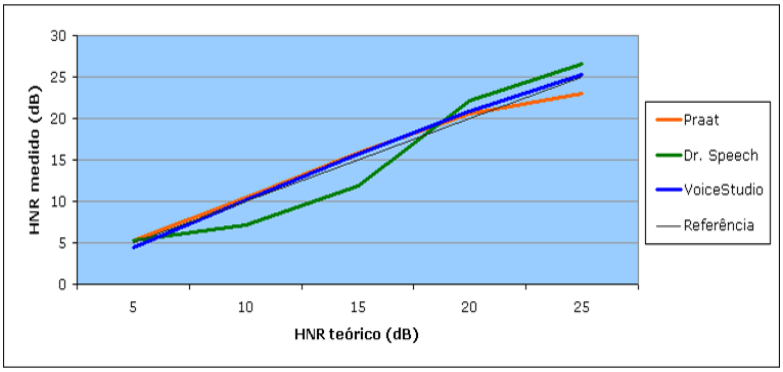


Figura 4.8: Resultado da medição de HNR em voz sintética masculina

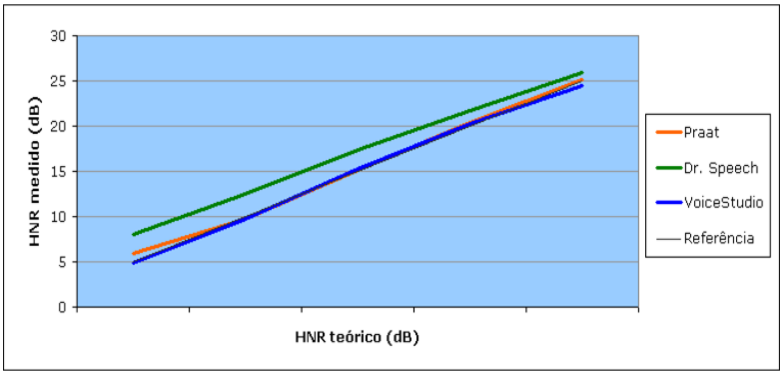


Figura 4.9: Resultado da medição de HNR em voz sintética feminina

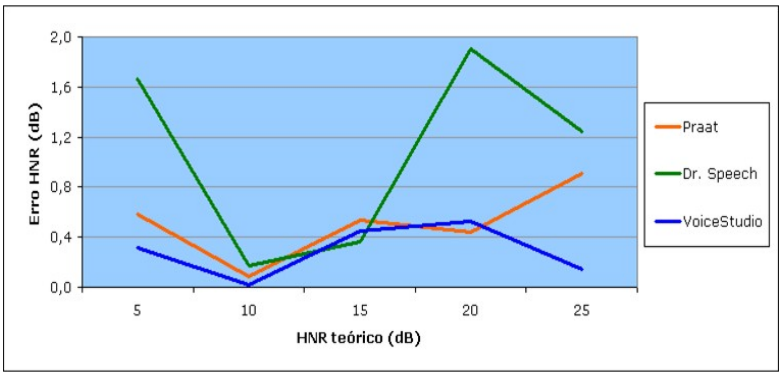


Figura 4.10: Resultado do erro HNR em voz sintética masculina/feminina

Em termos gerais para todas as medições (*jitter*, *shimmer*, HNR) o VoiceStudio apresenta um bom comportamento.

Capítulo 5 Conclusões

O ambiente de análise robusta dos principais parâmetros qualitativos da voz foi desenvolvido, demonstrando apresentar grande funcionalidade e fiabilidade.

Na literatura e nas soluções tecnológicas de análise acústica, o *pitch*, o *jitter*, o *shimmer* e o HNR são utilizados como os principais parâmetros qualitativos da voz. Estes parâmetros são obtidos a partir do sinal de áudio de uma vogal sustentada.

A utilização da envolvente de energia de tempo curto (obtida utilizando a transformada de Hilbert) possibilita uma boa detecção das marcas de pitch, melhorando assim os valores obtidos de *jitter* e *shimmer*.

Nesta dissertação foi apresentado uma nova abordagem de cálculo do HNR cujo desempenho foi avaliado positivamente usando vozes sintéticas.

Proveniente do trabalho desenvolvido em relação à medida de HNR, resultou um artigo cientifico para as I Jornadas sobre Tecnologia e Saúde organizado pelo Instituto Politécnico da Guarda em Abril de 2008.

O ambiente de análise robusta dos principais parâmetros qualitativos da voz foi desenvolvido e incorporado na aplicação SEEGNAL VoiceStudio. Esta encontra-se disponível no endereço http://www.seegnal.pt e já em fase de comercialização.

5.1 Perspectivas de evolução

A análise dos principais parâmetros qualitativos da voz foi implementada no ambiente, porém existem outros parâmetros acústicos que seria importante terem sido tomados em consideração. Por exemplo, uma avaliação do tremor da voz (amplitude e frequência do tremor) e da continuidade harmónica (ao longo do tempo e frequência). O tremor da voz já é um parâmetro muito utilizado para diagnóstico da voz. A continuidade harmónica é importante pois é um parâmetro que pretende avaliar a qualidade vocal de uma forma mais restrita do que o valor de HNR. O parâmetro da continuidade harmónica também seria um parâmetro interessante para avaliar a qualidade

do canto.

Um trabalho de investigação interessante na área de diagnóstico da voz seria criar uma grande base de dados de vozes patológicas e normais para obter conclusões relativamente aos limiares patológicos do j*itter*, *shimmer* e HNR.

Capítulo 6 Referências

[1] Isabel Guimarães, “A Ciência e a Arte da Voz Humana”, Escola Superior de Saúde de Alcoitão, Alcabideche, 2007.

[2] José Lopes, Susana Freitas, Ricardo Sousa, Joaquim Matos, Filipe Abreu, Aníbal Ferreira, “A medida HNR: a sua relevância na análise acústica da voz e sua estimação precisa”, nas I Jornadas sobre Tecnologia e Saúde, Instituto Politécnico da Guarda, Abril 2008.

[3] Luciana Andrade, “Determinação dos Limiares de normalidade dos parâmetros acústicos da voz,” Universidade de São Paulo, 2003.

[4] Isabel Guimarães, “An electrolaryngographic study of dysphonic Portuguese speakers”, University of London, 2002.

[5] Leila Horta e Shiro Tomita, “Um método de investigação dos distúrbios da fala e voz: a espectrografia vocal”, 2001.

[6] Thais Vanzella, “Normatização dos parâmetros acústicos vocais em crianças em idade escolar”, 2006

[7] Ingo Titze, “Workshop on Acoustic Voice Analysis”, National Center for Voice and Speech, 1994 .

[8] Domingos Oliveira, “Distúrbios vocais e problemas de entoação: estratégias articulatórias e correlatos acústicos”. Disponível em http://www.clinvoz.com.br/tese.htm .

[9] “Voice Lab in Clinical Practice”. Disponível em http://www.drspeech.com/Paper.html .

[10] “Dr. Speech”. Disponível em http://www.drspeech.com .

[11] “Praat”. Disponível em http://www.fon.hum.uva.nl/praat/

[12] Alan V. Oppenheim, Ronald W. Schafer e John R. Buck, “Discrete-Time Signal Processing, 2nd edition”, Prentice-Hall, 1999.

[13] “Speech Filing System”, Disponível em http://www.phon.ucl.ac.uk/resource/sfs/ .

[14] “Computer Speech System”. Disponível em http://www.kayelemetrics.com/Product%20Info/CSL%20Family/4500/4500.htm .

[15] “VoxMetria”. Disponível em **http://www.ctsinformatica.com.br/#voxMetria.html**

[16] “Qt Cross-Platform Application Framework”, Disponível em **http://**

trolltech.com/products/qt/ .
[17] "Qwt - Qt Widgets for Technical Applications". Disponível em http://qwt.sourceforge.net/ .
[18] "The RTAudio Home Page". Disponível em http://www.music.mcgill.ca/~gary/rtaudio .
[19] "Microsoft Visual C++ Express Edition". Disponível em http://www.microsoft.com/express/vc/ .
[20] Aníbal Ferreira, "Accurate Estimation in the ODFT Domain of the Frequency, Phase and Magnitude of Stationary Sinusoids", IEEE Worksop on Applications of Signal Processing to Audio and Acoustics, New Platz, E.U.A., Outubro de 2001.
[21] Aníbal Ferreira, "Combined Spectral Envelope Normalization and Subtraction of Sinusoidal Components in the ODFT and MDCT Frequency Domains", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Platz, E:U:A, Outubro 2001.
[22] Aníbal Ferreira e Deepen Sinha, "Accurate and Robust Frequency Estimation in the ODFT Domain". IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, E. U. A., Outubro de 2005.
[23] Paul Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", Proceedings of the Institute of Phonetic Sciences, 1993.
[24] Eiji Yumoto, "Harmonic-to-noise ratio as index of a degree of hoarseness", Journal of the Acoustic Society of America, Junho de 1982.
[25] Ana Mendes, "Voice acoustic patterns of patients diagnosed with vibroacoustic disease", Revista Portuguesa de Pneumologia, 2006
[26] Peter Murphy e Olatunji Akande, "Noise estimation in voice signals using short-term cepstral analysis", Journal of the Acoustic Society of America, Março de 2007.
[27] Guus Krom, "A cepstrum based technique for determining a harmonics-to-noise ratio in speech signals", Journal of the Hearing Research, 1993.
[28] Yinyoung Qi e Robert E. Hillman, "Temporal and spectral estimations of harmonics-to-noise ratio in voice signals", Journal of the Acoustic Society of America, Julho de 1997.

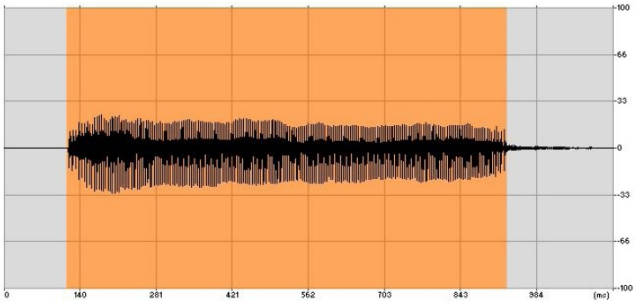
Anexo 1

SEEGNAL VoiceStudio

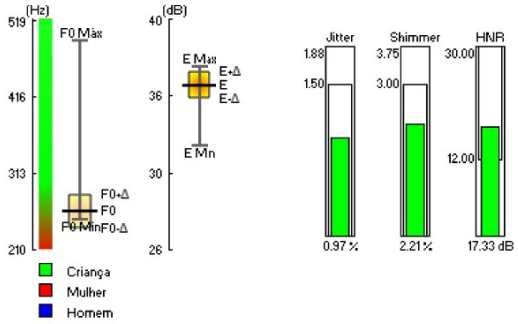
Descrição

Data e Hora	:	20 Dezembro 2007, 13:49
Nome do Ficheiro	:	C:\vogal_A_crianca.wav
Freq. Amostragem	(Hz) :	22050
Duração	(ms) :	1.12

Análise - Vozeamento



Estatísticas



Medidas de F0

F0 Médio	(Hz) :	260.59
F0 Máximo	(Hz) :	490.00
F0 Mínimo	(Hz) :	250.57
Desvio Padrão (Δ)	(Hz) :	22.32

Medidas de Energia

Energia Máxima	(dB) :	36.86
Energia Média	(dB) :	35.64
Energia Mínima	(dB) :	31.72
Desvio Padrão (Δ)	(dB) :	0.85

Parâmetros de Qualidade

Jitter (PPQ5)	(%) :	0.97
Shimmer (APQ5)	(%) :	2.21
Shimmer	(dB) :	0.32
HNR	(dB) :	17.33

Outros

Tempo de fonação	(ms) :	812.00
Extensão Vocal	(semitons) :	11.61

SEEGNAL Research, Lda
TECMAIA MaiaDigital, sala 3, Rua Frederico Ulrich Nº 2650, 4470-605 Moreira da Maia
Tlf: +351 229 489 030 http://www.seegnal.pt seegnal@seegnal.pt

Figura 1: Imagem de impressão das estatísticas da análise Vozeamento