Faculdade de Engenharia da Universidade do Porto

Modelização harmónica precisa de sons vozeados por humanos

Francisca Vieira de Brito

VERSÃO FINAL

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores
Major Telecomunicações, Eletrónica e Computadores

Orientador: Prof. Dr. Aníbal João de Sousa Ferreira

17 de julho de 2019

Resumo

A produção da fala resulta da passagem do fluxo de ar que sai dos pulmões pela laringe, o qual depois é modulado pela vibração das pregas vocais e convertido em impulsos acústicos que excitam o trato vocal. A vibração das pregas vocais ocorre de forma periódica com uma frequência designada de frequência fundamental, sendo esta a componente de um sinal de fala que contém mais informação acústica.

A frequência fundamental, a qual depende da vibração das pregas vocais, é responsável pela tonalidade ou *pitch* de um som. O *pitch* é a sensação subjetiva que cada pessoa possui em relação à frequência fundamental de um som, permitindo a distinção entre sons graves e agudos e informação sobre a fonte sonora. Deste modo, a estimação robusta e precisa da frequência fundamental de um sinal de voz falada, bem como das suas microvariações, tem interesse em diversas aplicações de processamento de fala como codificação, análise, síntese e reconhecimento da mesma.

Uma primeira fase de estimação da frequência fundamental passa por uma análise microscópica de estimação precisa da frequência de componentes sinusoidais individuais, já que um sinal de fala é constituído por várias sinusoides que, frequentemente, seguem uma estrutura harmónica. Neste âmbito, desenvolveu-se um novo estimador que, após diferentes testes realizados, mostrou ter o melhor desempenho entre os estimadores que utilizam a janela sinusoidal. A utilização desta janela prende-se com o facto de ser uma escolha conveniente para aplicações áudio baseadas na transformada real *Modified Discrete Cosine Transform* (MDCT).

Uma segunda fase de análise macroscópica diz respeito à estimação da frequência fundamental de uma estrutura harmónica. Foram testados diferentes algoritmos, uns de análise no domínio dos tempos e outros de análise no domínio das frequências, tendo todos apresentado um bom desempenho. O algoritmo SearchTonal, de análise no domínio das frequências, estando já preparado para realizar a modelização e síntese harmónica, foi utilizado no desenvolvimento desta dissertação.

A modelização harmónica precisa de sons vozeados por humanos reflete o funcionamento das pregas vocais, designadamente as microvariações da frequência fundamental. O objetivo é

obter modelos de magnitude e de fase espectrais totalmente paramétricos de modo a serem utilizáveis de forma flexível na representação e síntese de sinais de voz falada, preservando as suas características sonoras naturais. Após testes para diferentes cenários de síntese harmónica, concluiu-se que a utilização dos modelos paramétricos de magnitude e fase, modelos LPC e NRD, respetivamente, não compromete a qualidade auditiva do sinal processado.

A partir da modelização harmónica caracterizou-se a importância das microvariações da frequência fundamental de um sinal de voz falada, quer ao nível da identidade sonora de um orador, quer o impacto sonoro que decorre se estas forem alteradas de forma intencional por aplanamento ou por implantação dos padrões extraídos de um outro orador. Foram realizados testes de perceção auditiva, os quais permitiram concluir que as microvariações conferem naturalidade à voz. No entanto, a frequência fundamental média da voz é uma característica percetual predominante na captação da atenção do ouvido humano, pelo que as microvariações da frequência fundamental não são capazes de modificar a assinatura sonora de um dado orador quando implantadas sobre o valor médio da frequência fundamental de um outro orador.

Abstract

The production of speech results from the passage of airflow from the lungs through the larynx, which is then modulated by the vibration of the vocal folds and converted into acoustic impulses that excite the vocal tract. The vibration of the vocal folds occurs periodically with a frequency called as fundamental frequency, which is the component of a speech signal that contains more acoustic information.

The fundamental frequency, which depends on the vibration of the vocal folds, is responsible for the tone or pitch of a sound. Pitch is the subjective sensation that each person has in relation to the fundamental frequency of a sound, allowing the distinction between low and high sound and information about the sound source. Thus, robust and accurate estimation of the fundamental frequency of a spoken speech signal, as well as its microvariations, has a high relevance in various speech processing applications such as coding, analysis, synthesis and recognition.

A first phase of fundamental frequency estimation passes through a microscopic analysis of the precise estimation of the frequency of individual sinusoidal components, since a speech signal consists of several sinusoids that often follow a harmonic structure. In this context, a new estimator was developed that, after different tests, presented the best performance among the estimators that use the sinusoidal window. The use of this window is because it is a convenient choice for MDCT-based audio applications.

A second phase of macroscopic analysis concerns the estimation of the fundamental frequency of a harmonic structure. Different algorithms were tested, some of analysis in time domain and others in analysis in frequency domain, having all of them presented a good performance. The algorithm SearchTonal, which performs analysis in the domain of the frequencies, being already prepared for harmonic modeling and synthesis, was used in the development of this dissertation.

The accurate harmonic modeling of human voiced sounds reflects the functioning of the vocal folds, namely microvariations of the fundamental frequency. The goal is to obtain fully

parametric models of magnitude and spectral phase to be used in a flexible way in the representation and synthesis of speech signals, preserving their natural sound characteristics. After testing for different scenarios of harmonic synthesis, it was concluded that the use of parametric models of magnitude and phase, LPC and NRD models, respectively, does not compromise the auditory quality of the processed signal.

From the harmonic modeling, the importance of microvariations of the fundamental frequency of a spoken voice signal has been characterized, either at the sound identity level of a speaker or the sound impact that occurs if they are intentionally altered by flattening or by implanting models drawn from another speaker. Auditory perception tests were performed, which allowed the conclusion that microvariations confer naturality to the voice. However, the mean fundamental frequency of the voice is a predominant perceptual characteristic in capturing the attention of the human ear, reason why the microvariations of the fundamental frequency are not able to modify the sound signature of a given speaker when implanted on the average value of the fundamental frequency of another speaker.

Agradecimentos

Começo por agradecer ao Professor Aníbal Ferreira pela sua orientação e pela oportunidade de trabalhar neste projeto.

Ao João Silva, com quem trabalhei todos os dias, um imenso obrigada pelo bom ambiente, pela disponibilidade, pela paciência, pela ajuda no desenvolvimento desta dissertação.

Agradeço do fundo do coração à minha mãe, ao meu pai e ao meu irmão por todo o apoio ao longo destes anos, por todos os momentos de lazer e descontração "obrigados", mas que tão bem fizeram, por estarem sempre ao meu lado nos momentos bons e menos bons. Um obrigada muito especial à minha mãe pela sua ajuda, compreensão, opiniões, sugestões, por nunca desistires de mim e pela verificação ortográfica desta dissertação.

Por último, mas não menos importante, agradeço ao meu melhor amigo e namorado, André. Obrigada por todo o apoio, por me fazeres sempre rir, em qualquer momento, por me ouvires e estares sempre ao meu lado.

Índice

Lista de figuras

Lista de tabelas

Abreviaturas e Símbolos

Lista de abreviaturas (ordenadas por ordem alfabética)

AM	*Amplitude Modulation*
AWGN	*Additive White Gaussian Noise*
CRLB	*Cramér-Rao Lower Bound*
DFT	*Discrete Fourier Transform*
FCT	Fundação para a Ciência e a Tecnologia
FM	*Frequency Modulation*
IL	*Intensity Level*
LPC	*Linear Predictive Model*
MDCT	*Modified Discrete Cosine Transform*
MIDI	*Musical Instrument Digital Interface*
NRD	*Normalized Relative Delay*
ODFT	*Odd-frequency Discrete Fourier Transform*
PCM	*Pulse Code Modulation*
PSD	*Power Spectral Density*
RMSE	*Root Mean Square Error*
SNR	*Signal to Noise Ratio*
SPF	*Speaker* Feminino
SPM	*Speaker* Masculino
SWIPE	*Sawtooth Waveform Inspired Pitch Estimation*

Lista de símbolos

dB	deciBel
F0	Frequência fundamental
Hz	Hertz
ms	Milissegundo
T0	Período fundamental

Capítulo 1

Introdução

Neste capítulo apresenta-se o enquadramento e a motivação do trabalho realizado. De seguida, apresentam-se os principais objetivos e a estrutura deste documento.

1.1. Enquadramento e motivação

O tema da dissertação, apresentado como "Modelização harmónica precisa de sons vozeados por humanos", enquadra-se num projeto FCT cujo objetivo é a reconstrução de voz disfónica.

A voz normal vozeada envolve a vibração das pregas vocais. No entanto, várias razões podem impedir que uma pessoa não produza vozeamento ao falar, como por exemplo, o ambiente envolvente ser privado ou silencioso, a afonia ou até mesmo algumas condições de saúde. Nestes casos, o regime de comunicação oral é dito de fala sussurrada. A fala sussurrada é problemática pois tipicamente a projeção da voz é fraca, a inteligibilidade da fala é mais afetada por ruído e a identificação do orador torna-se mais difícil ou até mesmo impossível.

Uma solução típica para as pessoas laringectomizadas, isto é, a quem foi retirada a laringe devido a alguma condição grave de saúde, é a utilização de um aparelho denominado eletrolaringe. Este aparelho vibratório gera um sinal artificial periódico que permite a recuperação da voz, mas que soa algo robótico. O que se pretende com este projeto é ajudar as pessoas afetadas por disfonia de voz, temporária ou permanente, a comunicar de forma efetiva e mais confortável.

O facto de o alcance dos objetivos deste projeto poderem vir a ter um grande impacto na qualidade de vida das pessoas laringectomizadas é uma motivação. A conversão de voz sussurrada em voz artificial com a identidade sonora do próprio orador facilitará a comunicação humano-com-humano e humano-com-máquina, pois o discurso será mais inteligível, natural e característico.

1.2. Objetivos

O principal objetivo do trabalho desenvolvido ao longo desta dissertação é realizar uma modelização precisa das características da voz vozeada que refletem o funcionamento das pregas vocais, designadamente ao nível das microvariações da frequência fundamental (F0), de modo a poder caracterizar-se, quer a sua importância ao nível da identidade sonora de um orador, quer o impacto sonoro que decorre se esta for alterada de forma intencional, por exemplo, por aplanamento ou por implantação das características de um outro orador.

Deste modo, é necessário começar pela estimação da frequência fundamental, trabalho que está dividido em duas fases:

- uma primeira fase de perspetiva microscópica em que é realizada a melhor estimação possível da frequência de componentes sinusoidais individuais, considerando o impacto quer de ruído quer da influência de outras sinusoides;
- uma segunda fase de perspetiva macroscópica em que se utilizam os resultados obtidos na primeira fase para relacionar várias sinusoides em alinhamento harmónico e inferir a frequência fundamental adjacente, ou seja, é nesta fase que se realiza a estimação da frequência fundamental de um sinal.

Para o trabalho a realizar ao longo do desenvolvimento da dissertação foram propostos os seguintes objetivos:

- revisão da literatura e levantamento do estado da arte;
- familiarização com ambientes de *software* para análise harmónica de sinais de voz;
- estudo comparativo de estimadores de frequência de componentes sinusoidais individuais;
- estudo comparativo de algoritmos de estimação da frequência fundamental da voz (*pitch estimation*);
- seleção e aperfeiçoamento de um algoritmo de *pitch estimation* que consiga captar microvariações da frequência fundamental e caracterização do seu desempenho;
- modelização das microvariações da frequência fundamental do sinal de voz em contexto de vogais isoladas e criação de padrões de variação com potencial idiossincrático;
- transformação intencional das microvariações da frequência fundamental do sinal de voz de um dado orador, quer por critérios simples de uniformização ou aplanamento, quer por implantação dos padrões extraídos de um orador diferente, e avaliação do decorrente impacto objetivo e percetivo;
- escrita da dissertação e publicação de um artigo.

1.3. Estrutura da dissertação

Este documento está organizado em 7 capítulos:

- Capítulo 1 - apresentação do enquadramento e motivação do trabalho de investigação realizado, bem como dos seus objetivos, e organização do restante documento;
- Capítulo 2 - estado da arte relativo ao funcionamento da produção da fala e quais as principais características de um sinal de voz falada;
- Capítulo 3 - perspetiva microscópica de estimação precisa de frequência. Descrição de diferentes estimadores de frequência de componentes sinusoidais individuais e apresentação dos resultados de avaliação de desempenho destes;
- Capítulo 4 - perspetiva macroscópica de estimação precisa de frequência. Descrição de diferentes algoritmos de estimação da frequência fundamental de sinais de fala que seguem uma estrutura harmónica e apresentação dos resultados de avaliação de desempenho destes;
- Capítulo 5 - modelização de magnitude e de fase paramétrica de sinais sonoros de fala e síntese harmónica no domínio das frequências. Avaliação objetiva e subjetiva após testes para diferentes cenários de síntese;
- Capítulo 6 - transformação intencional das microvariações da frequência fundamental de sinais de voz falada. Avaliação dos resultados por meio de testes de perceção auditiva;
- Capítulo 7 - conclusões do trabalho e sugestões para trabalho futuro.

Capítulo 2

Produção e caracterização de um sinal de voz falada

Neste capítulo descreve-se o funcionamento da produção da fala e quais as principais características de um sinal sonoro de fala.

2.1. Mecanismos de produção da fala

Nesta secção apresenta-se o modelo que traduz o mecanismo de produção da fala, como é que, fisiologicamente, esta ocorre e a importância das pregas vocais na fonação.

2.1.1. Modelo fonte-filtro

O modelo fonte-filtro de produção da voz traduz-se num sistema simplificado que admite um sinal à sua entrada e que produz um sinal modificado à saída. Na figura 2.1 está representado o modelo fonte-filtro.

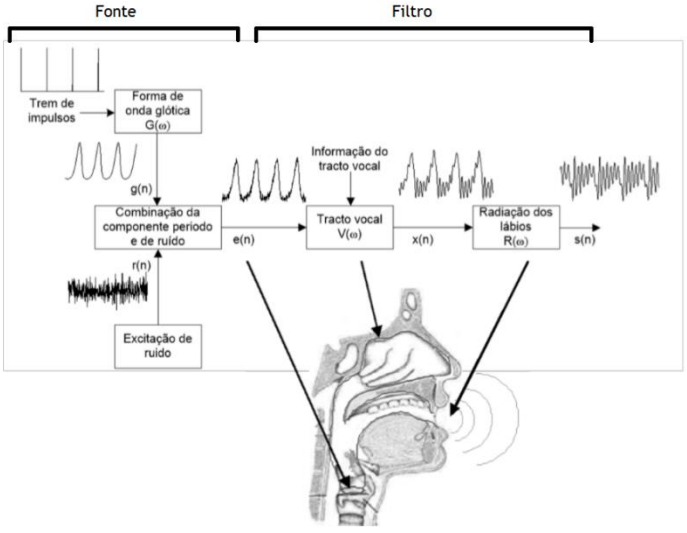

Figura 2.1 - Modelo fonte-filtro da produção de fala [1]. *(adaptada)*

A fonte é o sinal gerado quando o ar é expelido pelos pulmões, o qual é composto por duas componentes: i) os impulsos glotais periódicos resultantes da vibração das pregas vocais; ii) ruído de natureza aleatória. O filtro representa a influência do trato vocal e ainda da radiação labial, na modificação do sinal, obtendo-se o som vocal com o seu timbre característico [1].

É importante notar que o modelo apresentado na figura 2.1 diz respeito a um som sustentado, pois os filtros permanecem essencialmente inalterados ao longo do tempo. No entanto, na fala normal existe dinâmica temporal de produção dos diferentes fonemas com fenómenos de coarticulação, pelo que o modelo fonte-filtro só é representativo em regiões quase estacionárias do sinal de voz, com uma duração típica de 10 a 20 ms [1].

De um modo geral, o modelo fonte-filtro é importante pois permite discussão na caracterização de diferentes tipos de sons vocais, motiva a interpretação espectral da produção de fala e destaca oportunidades para utilizar o sinal de voz de modo a capturar características idiossincráticas [1].

2.1.2. Modo e lugar de articulação

O modo de produção da fala depende de como o ar expelido pelos pulmões chega ao exterior. Quando existe vibração das pregas vocais, isto é, fonação, os sons são designados de vozeados. Caso contrário, os sons são desvozeados ou não vozeados, dificultando a distinção entre alguns fonemas. Os sons vozeados podem sofrer ou não de turbulência, dividindo-se em três casos distintos: i) turbulência causada pela constrição de um ponto no trato vocal, dando-se a produção de consoantes fricativas; ii) turbulência que ocorre pela obstrução total da passagem do ar, seguida da sua libertação repentina, produzindo consoantes oclusivas; iii) a não existência de turbulência, resulta na produção de vogais [2].

O lugar de articulação, no caso das consoantes, é o sítio da constrição e, no caso das vogais, é marcado pela posição da língua, sendo afetado também pela abertura do maxilar [2].

2.1.3. Pregas vocais e fonação

A fonação ou produção da fala, de um ponto de vista fisiológico, ocorre da seguinte forma: o fluxo de ar proveniente dos pulmões é empurrado para a laringe, onde é modulado pela vibração das pregas vocais e convertido em impulsos acústicos, fornecendo um sinal de excitação ao trato vocal. O trato vocal, tal como referido anteriormente, filtra o fluxo de ar modulado, o qual depois é radiado pelos lábios. Os diferentes tipos de som produzidos ocorrem pela regulação da vibração das pregas vocais ou devido à configuração do trato vocal, articulação da língua, mandíbula, palato mole e lábios [3].

As pregas vocais são um tecido musculoso que se move, abrindo e fechando a glote. A oscilação destas, interrompe periodicamente o fluxo de ar que é expelido pelos pulmões, em razão da pressão subglótica. Caso não esteja a ser produzido som, as pregas vocais estão abertas e

caso esteja apenas a ser produzido ruído, as pregas vocais também estão separadas, mas existe uma constrição em algum ponto do trato vocal que causa turbulência. Na produção de som, as pregas vocais fecham, causando resistência ao fluxo de ar, até abrirem, deixando que o ar passe através da glote. Este processo ocorre repetidamente e a duração de cada ciclo é designada de período fundamental (T0), que é o inverso da frequência fundamental (F0) [3].

Na figura 2.2 é possível observar diferentes configurações da glote para vários tipos de fonação.

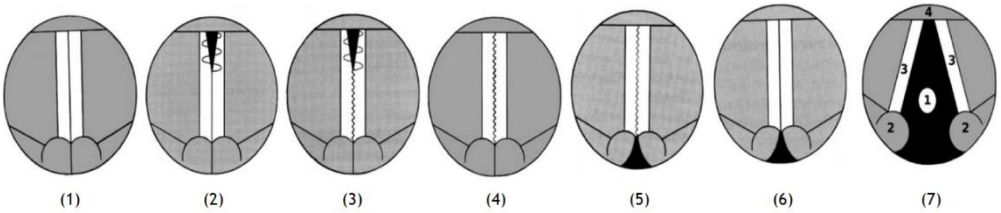

Figura 2.2 - Configurações da glote para diferentes tipos de fonação: (1) glote parada, (2) voz laringelarizada, (3) voz estridente, (4) voz modal, (5) voz ofegante, (6) sussurro, (7) sem voz (1-glote, 2-cartilagem aritenoide, 3-prega vocal, 4-epiglote) [3].

2.2. Características de um sinal sonoro de fala

Nesta secção apresentam-se os parâmetros físicos característicos dos sinais sonoros e as sensações subjetivas associadas a cada um destes parâmetros.

2.2.1. Frequência fundamental (F0) e formantes

A frequência fundamental é o termo físico associado ao *pitch*, definida como a frequência mais baixa, ou seja, o primeiro harmónico da onda sonora. Esta grandeza é definida apenas em sons periódicos ou quase periódicos, sendo que, em situações ambíguas, corresponde à perceção do *pitch*. A sua importância prende-se com o facto de ser esta a frequência responsável pela tonalidade do som, resultando numa voz mais grave ou mais aguda, que depende do elemento vibrador do corpo humano que permite a sua produção - as pregas vocais [1] [4].

Os sons produzidos na fala, na sua maioria, são constituídos por várias componentes de diferentes frequências, designadas de parciais. Neste caso, o parcial de frequência mais baixa é designado de F0 e os restantes parciais, caso sejam múltiplos inteiros de F0, são designados de harmónicos [1].

Os formantes são regiões de concentração destacada de energia presentes no espectro sonoro, a seguir a F0, sendo que a frequência do primeiro formante é menor do que a do segundo formante, e assim sucessivamente. Estas frequências são importantes na definição do timbre da voz de uma pessoa [1].

2.2.2. Intensidade sonora

A intensidade sonora é a perceção da amplitude da onda sonora. Há dois limites que definem a sensibilidade auditiva para a intensidade sonora de um som puro: o limiar da audição, referente à intensidade mínima que torna um som puro audível, e o limiar da dor, referente à intensidade sonora que provoca sofrimento no sistema auditivo. O nível da intensidade sonora, ou *Intensity Level* (IL), é medido numa escala logarítmica, sendo a sua unidade o deciBel (dB) [4].

2.2.3. Tonalidade, intensidade subjetiva e timbre

A tonalidade ou *pitch* é a sensação subjetiva que cada pessoa possui em relação à frequência fundamental de um determinado som e permite a distinção entre um som grave e um som agudo [4]. Esta característica fornece informação sobre a fonte sonora, sendo que, no que diz respeito à fala, ajuda a identificar o género do orador (a tonalidade tende a ser mais aguda para os oradores do género feminino do que para os oradores do género masculino) [5], fornece significado adicional às palavras (por exemplo, permite perceber se uma frase é dita com a intenção de interrogação ou não) e pode ajudar a identificar o estado emocional do orador (por exemplo, um estado de alegria leva a que a voz produzida seja mais aguda e que o intervalo de tons seja maior, enquanto um estado de tristeza leva à produção de uma voz normal ou com uma tonalidade mais grave e com um intervalo de tons mais pequeno) [6]. Deste modo, a tonalidade relaciona-se fisicamente com a frequência fundamental (em Hz) do som [4].

A intensidade subjetiva ou *loudness* está relacionada com a perceção qualitativa do som pelo ouvido humano, fornecendo também informações sobre a fonte sonora [4]. Moore define *loudness* como sendo "aquele atributo de sensação auditiva em que os sons podem ser ordenados numa escala extensível de silencioso a alto" [7]. Fisicamente relaciona-se com a intensidade sonora (em dB) [4].

Por último, o timbre permite distinguir dois sons com a mesma tonalidade e a mesma intensidade subjetiva. Relaciona-se com o perfil espectral do sinal sonoro, também designado de envolvente espectral [4].

2.3. Síntese

Neste capítulo foram apresentadas noções fundamentais acerca da voz falada: como é que é produzida a fala de um ponto de vista fisiológico e quais as principais características físicas e subjetivas de um sinal de voz.

Apesar do *pitch* ser uma sensação subjetiva que depende de cada indivíduo, a relação que existe entre essa sensação e a frequência fundamental demonstra a importância da estimação da frequência fundamental de um sinal de voz falada, bem como das suas microvariações, de forma robusta e precisa, de modo a obter uma caracterização objetiva da voz.

Capítulo 3

Estimação da frequência de componentes sinusoidais individuais

Neste capítulo apresenta-se a primeira fase do trabalho de estimação da frequência fundamental, na qual é realizada a análise microscópica. Descrevem-se diferentes estimadores de frequência e apresentam-se os resultados obtidos após a realização de testes.

3.1. Introdução

A estimação precisa da frequência de sinusoides reais tem sido alvo de intensas pesquisas devido à sua importância na codificação de alta qualidade de áudio e de fala [8] [9], na análise de voz cantada [10] e na transcrição automática de música *Pulse Code Modulation* (PCM) para notação musical simbólica, como por exemplo *Musical Instrument Digital Interface* (MIDI) [13]. Tal estimação é o ponto de partida para a análise e modelização sinusoidal utilizada em áreas de processamento de sinal, tais como, as telecomunicações, a medicina e a instrumentação [12].

O espectro de um sinal de fala é tipicamente constituído por várias sinusoides que, frequentemente, seguem uma estrutura harmónica [13]. A estimação exata da frequência de cada sinusoide é usualmente obtida a partir da transformada discreta de Fourier (DFT, *Discrete Fourier Transform*) de um sinal discreto. Esta estimação envolve a utilização de várias amostras do espectro da DFT, designadas de bins da DFT, já que o seu espectro é discreto nas frequências e limitado pela resolução da DFT ($2\pi/N$, onde N é o tamanho da DFT) [14].

As condições de teste dos algoritmos podem diferir significativamente, tornando o seu desempenho relativo bastante difícil. Por exemplo, existem diferentes janelas que podem ser aplicadas ao sinal antes de se realizar a DFT, as quais têm um forte impacto na seletividade da frequência da DFT e no efeito de *leakage* [15]. A sinusoide pode ser uma sinusoide complexa

ou uma sinusoide real. O procedimento de interpolação da DFT pode ser iterativo ou não iterativo. Num cenário mais realista, um sinal de fala é constituído por várias sinusoides, devido à sua natureza harmónica [14], o qual é alvo de análise no capítulo 4.

Visto o objetivo ser a análise precisa, e em tempo real, de cada componente sinusoidal individual da estrutura harmónica do sinal de fala, a estimação da frequência deve ser realizada utilizando um algoritmo que [14]:

- evite procedimentos iterativos e é computacionalmente simples;
- evite o cálculo de uma DFT maior do que o comprimento do vetor de dados;
- maximize a precisão da estimativa e a robustez ao ruído quando não só este está presente, mas também se verifica a interferência de outras sinusoides.

3.2. O problema da estimação

O sinal sinusoidal inicial considerado é dado por:

$$x_i(n) = A\ \sin\left[\frac{2\pi}{N}(l + \Delta l)n + \ \varphi\right] + r(n), \tag{3.1}$$

onde A é a magnitude da sinusoide, l e Δl são, respetivamente, a parte inteira e a parte fracionária da frequência da sinusoide na escala de frequências em bins da DFT com periodicidade N, n são valores de 0 a $N-1$, φ é a fase da sinusoide, e $r(n)$ é ruído aditivo Gaussiano branco [16].

A este sinal é aplicada uma janela $h(n)$, e só depois é aplicada a transformada para o domínio das frequências utilizando uma DFT de N pontos, tal como indica a seguinte equação [14]:

$$X_i(k) = \ \sum_{n=0}^{N-1} h(n)x_i(n)e^{-j\frac{2\pi}{N}kn}. \tag{3.2}$$

O objetivo principal do problema da estimação é encontrar um algoritmo que utilize os valores de $X(k)$ e interpole o valor de Δl uma vez que l pode ser facilmente identificado, correspondendo ao máximo local de $|X(k)|$ [14]. Esta abordagem está representada na figura 3.1. Deste modo, a partir de três amostras da DFT do sinal, X_{l-1}, X_l e X_{l+1}, pretende-se estimar a frequência do pico espectral l_{pico}, tal como indica a figura 3.2. Esta estimativa é dada pela soma do índice inteiro do pico, l, com a parte fracionária, Δl, na escala em bins da DFT:

$$l_{pico} = l + \ \Delta l, \tag{3.3}$$

onde Δl pode ser positivo ou negativo [12].

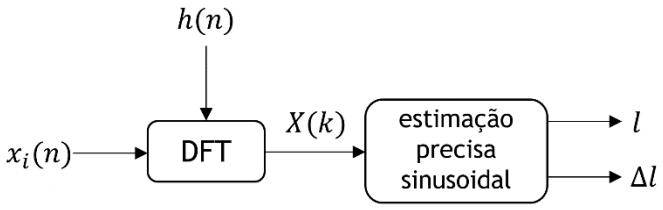

Figura 3.1 - Diagrama de blocos simplificado do algoritmo de estimação da frequência. Os índices $\boldsymbol{n}$ e $\boldsymbol{k}$ representam, respetivamente, o tempo e a frequência. A frequência exata da sinusoide na escala em bins da DFT é representada pela parte inteira $\boldsymbol{l}$ e pela parte fracionária $\Delta\boldsymbol{l}$.

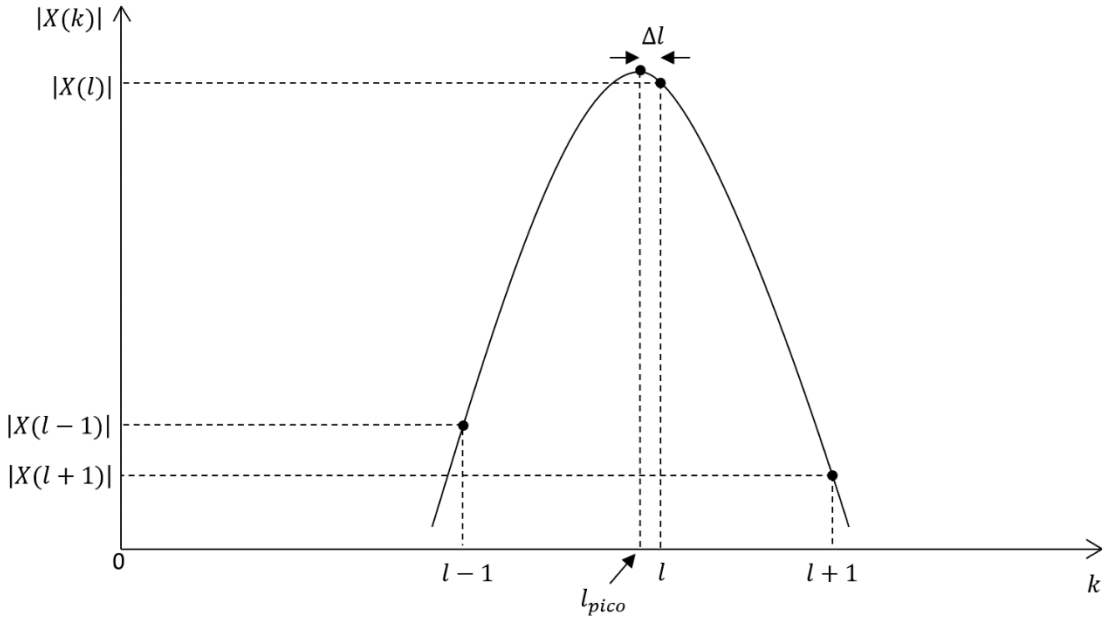

Figura 3.2 - Amostras de magnitude da DFT de uma sinusoide. Neste caso, $\boldsymbol{l_{pico} = l - \Delta l}$ [12]. *(adaptada)*

3.2.1. Janelas e as suas implicações

A aplicação de uma janela ao sinal tem como objetivo estimar com precisão a frequência de todas as sinusoides individuais existentes num sinal de fala, ou seja, de vários tons. Assim, tal como é afirmado por Harris, "o alcance dinâmico máximo na deteção multitonal (de vários tons) requer que a transformada de Fourier da janela exiba um lóbulo central altamente concentrado com estrutura de lóbulo lateral muito baixa" [17].

Os algoritmos que são apresentados utilizam diferentes janelas, num total de três, de seletividade decrescente e atenuação dos lóbulos laterais principais crescente: a janela retangular

$$h_R(n) = 1, n = 0, 1, \dots, N-1, \tag{3.4}$$

a janela sinusoidal

$$h_S(n) = \sqrt{h_H} = \sin\frac{\pi}{N}(n + 0.5)\,, n = 0, 1, \dots, N-1 \tag{3.5}$$

e a janela de Hanning

$$h_H(n) = \frac{1}{2}\left[1 - \cos\frac{2\pi}{N}(n + 0.5)\right], n = 0, 1, \ldots, N - 1. \quad (3.6)$$

Na figura 3.3 estão representadas as respostas em frequência para cada uma das três janelas descritas. Observa-se que a largura do lóbulo principal da janela retangular, sinusoidal e de Hanning é, respetivamente, $4\pi/N$, $6\pi/N$ e $8\pi/N$. Esta largura pode ser vista como a frequência mínima de separação entre duas sinusoides de modo a poderem ser resolvidas por linhas espectrais adjacentes da DFT, isto é, as sinusoides aparecem como picos individuais no espectro de magnitude da DFT. Nesta perspetiva, a janela retangular tem a melhor seletividade e a janela de Hanning tem a pior seletividade. Por outro lado, quanto mais largo for o lóbulo principal, melhor será a atenuação entre o lóbulo principal e os lóbulos laterais – esta característica é designada por *leakage* (fugas). Nesta perspetiva, a janela retangular tem o maior *leakage* e a janela de Hanning tem o menor *leakage*. Quanto menor o *leakage*, menor a influência entre duas sinusoides no espectro de magnitude da DFT. Estes são aspetos importantes que podem influenciar o desempenho do processo de estimação da frequência quando há ruído ou interferência de outras sinusoides no sinal [16].

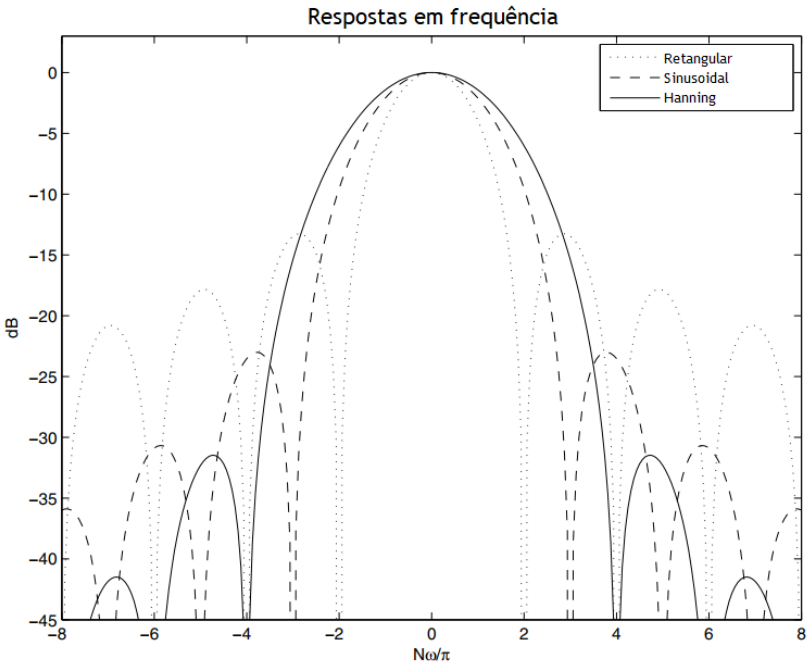

Figura 3.3 - Resposta em frequência da janela retangular, da janela sinusoidal e da janela de Hanning [16].

De modo a evitar tanto quanto possível o *leakage*, é apropriado interpolar o valor de Δl utilizando as duas, três ou quatro linhas espectrais maiores da DFT em torno de um pico espectral, quando se aplica a janela retangular, sinusoidal ou de Hanning, respetivamente [14].

3.2.2. Limite inferior de Cramér-Rao (CRLB)

Para avaliar o desempenho dos estimadores de frequência que vão ser apresentados, toma-se como referência o limite inferior de Cramér-Rao (CRLB, *Cramér-Rao Lower Bound*). Para o caso de uma sinusoide complexa, o CRLB é dado por [18] [19] [20] [21]:

$$var\{\widehat{w_l}\} \geq \frac{6\sigma^2}{A^2N(N^2-1)}, \tag{3.7}$$

onde A é a magnitude da sinusoide, N é o tamanho da DFT e σ^2 é a variância do ruído, assumindo que este é de média nula, branco, complexo e Gaussiano. De modo a simplificar a avaliação relativa do desempenho dos diferentes estimadores de frequência, ao CRLB dado por $var\{\widehat{w_l}\}$ será aplicada a raiz quadrada e esse valor é normalizado pela resolução de frequência natural da DFT ($2\pi/N$), expressando o resultado como uma percentagem da largura de banda normalizada, tal como se pode verificar na equação seguinte [14]:

$$RMSE = \frac{\sqrt{var\{\widehat{w_l}\}}}{2\pi/N} \geq \sqrt{\frac{3\sigma^2}{2A^2\pi^2N(1-1/N^2)}}. \tag{3.8}$$

3.3. Estimadores baseados na janela retangular

Nesta secção descrevem-se alguns estimadores de frequência não iterativos baseados na janela retangular, [18] e [12].

3.3.1. Estimador de Macleod

Macleod [18] desenvolveu um interpolador de três amostras de frequência que envolve uma amostra de pico no espectro da DFT e amostras dos seus dois vizinhos. Considerando que para melhorar o desempenho é necessário utilizar informação sobre a fase e a magnitude da DFT no estimador de frequência, primeiro são calculados $\alpha = \Re\{X(l)\,X^*(l)\}$, $\alpha_L = \Re\{X(l-1)\,X^*(l)\}$ e $\alpha_R = \Re\{X(l+1)\,X^*(l)\}$, que levam a [14]

$$\Upsilon = \frac{\alpha_L - \alpha_R}{2\alpha + \alpha_L + \alpha_R}, \tag{3.9}$$

e, finalmente, à estimação de Δl dada por [14]:

$$\widehat{\Delta l} = \frac{\sqrt{1+8\Upsilon^2}-1}{4\Upsilon},\ -\frac{1}{2} \leq \widehat{\Delta l} \leq \frac{1}{2}. \tag{3.10}$$

3.3.2. Estimador de Jacobsen e Kootsookos

O estimador de frequência desenvolvido por Jacobsen *et al*. [12] é definido apenas por uma regra para a estimação da parte fracionária Δl e, tal como o estimador anterior, utiliza três amostras da DFT centradas no pico espectral [14]:

$$\widehat{\Delta l} = \Re\left\{\frac{X(l+1)-X(l-1)}{2X(l)-X(l+1)-X(l-1)}\right\},\ -\frac{1}{2} \leq \widehat{\Delta l} \leq \frac{1}{2}. \tag{3.11}$$

O autor indica que este simples estimador é surpreendentemente preciso até para valores de relação sinal-ruído (SNR, *Signal to Noise Ratio*) muito baixos, o que é em parte explicado pelo facto de ter uma habilidade intrínseca de cancelar tendências estatísticas [14].

3.4. Estimadores baseados na janela de Hanning

Nesta secção descrevem-se alguns estimadores de frequência não iterativos baseados na janela de Hanning, [22], [18], [23] e [12].

3.4.1. Estimador de Grandke

Grandke [22] reconhece que o *leakage* devido a "interferência harmónica" é um problema quando se utilizam interpoladores de frequência baseados em janelas retangulares, por isso sugeriu um estimador de frequência a ser usado com a janela de Hanning, o qual providencia quase nenhum *leakage* de "longo alcance" [14]:

$$\widehat{\Delta l} = \frac{2|X(l+1)|-|X(l)|}{|X(l+1)|+|X(l)|},\ 0 \leq \widehat{\Delta l} < 1. \tag{3.12}$$

Em relação à interferência harmónica, Grandke também antecipa que "janelas mais sofisticadas" do que a janela de Hanning podem contornar as restrições existentes devido ao facto de "os tons terem de ser espaçados o suficiente" [22].

3.4.2. Estimador de Macleod

Para além do estimador de frequência proposto utilizando a janela retangular, descrito na subsecção 3.3.1, Macleod [18] também propôs um estimador de frequência utilizando a janela de Hanning e que tem "rejeição de *leakage* intrínseco". Calculando primeiro $= \Re\{X(l)\,X^*(l)\}$, $\alpha_L = \Re\{X(l-1)\,X^*(l)\}$ e $\alpha_R = \Re\{X(l+1)\,X^*(l)\}$, a estimação da frequência é dada por [14]:

$$\widehat{\Delta l} = 2\frac{\alpha_L-\alpha_R}{2\alpha-\alpha_L-\alpha_R},\ -\frac{1}{2} \leq \widehat{\Delta l} \leq \frac{1}{2}. \tag{3.13}$$

3.4.3. Estimador de Quinn

O estimador de frequência proposto por Quinn [23] utiliza as linhas espectrais da DFT de cada lado do máximo local (em $k = l$), de modo a melhorar a robustez ao ruído. Definido $\alpha_L = \frac{2\alpha+1}{1-\alpha}$, onde $\alpha = \Re\left\{\frac{X(l-1)}{X(l)}\right\}$, e $\alpha_L = \frac{2\alpha+1}{\alpha-1}$, onde $\alpha = \Re\left\{\frac{X(l+1)}{X(l)}\right\}$, a frequência é estimada por [14]:

$$\widehat{\Delta l} = \frac{\alpha_L+\alpha_R}{2} + \Upsilon(\alpha_L^2) - \Upsilon(\alpha_R^2),\ -\frac{1}{2} \leq \widehat{\Delta l} \leq \frac{1}{2}, \tag{3.14}$$

onde [14]

$$\Upsilon(x) = -\frac{5}{14}\log(35x^2 + 120x + 32) - \frac{\sqrt{155}}{140}\log\left(\frac{12+7x-4\sqrt{\frac{31}{5}}}{12+7x+4\sqrt{\frac{31}{5}}}\right). \tag{3.15}$$

Os resultados de desempenho obtidos para uma sinusoide real não são muito informativos, pois foram obtidos para um valor de SNR específico [23].

3.4.4. Estimador de Jacobsen e Kootsookos

Jacobsen *et al.* [12], para além do estimador de frequência descrito na subsecção 3.3.2, também propuseram no mesmo artigo dois interpoladores de frequência a serem utilizados com a janela de Hanning, incluindo [14]:

$$\widehat{\Delta l} = 1.36\frac{|X(l+1)|-|X(l-1)|}{|X(l)|+|X(l+1)|+|X(l-1)|},\ -\frac{1}{2} \leq \widehat{\Delta l} \leq \frac{1}{2}. \tag{3.16}$$

No entanto, os resultados de desempenho são apresentados para um único tom e para um intervalo de valores de SNR pequeno (-2 dB até 10 dB).

3.5. Estimadores baseados na janela sinusoidal

Nesta secção descrevem-se alguns estimadores de frequência não iterativos baseados na janela sinusoidal, [16], [24] e o estimador proposto.

3.5.1. Estimador ArcTan

O estimador designado ArcTan foi proposto por Aníbal Ferreira e por Ricardo Sousa, sendo que no respetivo artigo [16] são apresentados três estimadores de frequência baseados no Arc-Tan consoante a aplicação de uma janela sinusoidal, retangular ou de Hanning. No entanto, apenas será apresentado aqui o estimador utilizando a janela sinusoidal.

O estimador ArcTan faz uso dos três bins mais largos da DFT, pois, no máximo, existem três bins da DFT que pertencem ao lóbulo principal da resposta de frequência da janela sinusoidal [16]. Considerando que a transformada de Fourier do sinal sinusoidal inicial ao qual foi aplicada a janela é dada por [16]:

$$x(n) = x_i(n)h_S(n) \leftrightarrow X(e^{jw}) = \frac{1}{2\pi}X_i(e^{jw}) * H_S(e^{jw}), \tag{3.17}$$

onde $H_S(e^{jw}) = \sum_{n=0}^{N-1} h_S(n)e^{-jwn}$, $|X(l)|$ é um máximo local e, ou $|X(l+1)| > |X(l-1)|$, o que denota que a frequência exata pode ser estimada como $l + \Delta l$, ou $|X(l+1)| < |X(l-1)|$, pelo que a frequência exata pode ser estimada como $l - \Delta l$ [16]. A estimação de Δl deve ser então um valor entre 0.0 e 0.5.

No primeiro caso, em que $|X(l+1)| > |X(l-1)|$, quando $0.0 \leq \Delta l \leq \gamma$, onde γ é um parâmetro de otimização, a melhor estimativa de frequência é dada por:

$$\widehat{\Delta l} \approx \frac{3}{\pi}\arctan\frac{1}{\sqrt{3}}\frac{1-Q^G}{1+Q^G}, \tag{3.18}$$

onde $Q = \frac{|X(l-1)|}{|X(l+1)|} \approx \left[\frac{\cos\frac{\pi}{3}(\Delta l+1)}{\cos\frac{\pi}{3}(1-\Delta l)}\right]^{\frac{1}{G}}$ [16]. Quando $\gamma \leq \Delta l \leq 0.5$, a melhor estimativa de frequência é dada por:

$$\widehat{\Delta l} \approx \frac{3}{\pi}\arctan\frac{2S^F-1}{\sqrt{3}}, \tag{3.19}$$

onde $S = \frac{|X(l+1)|}{|X(l)|} \approx \left[\frac{\cos\frac{\pi}{3}(1-\Delta l)}{\cos\frac{\pi}{3}\Delta l}\right]^{\frac{1}{F}}$ [16].

A otimização do erro de estimação no sentido da teoria de decisão *minmax* é realizada tal como descrita em [25], onde são especificados os valores dos parâmetros γ, G e F.

No segundo caso, em que $|X(l+1)| < |X(l-1)|$, as mesmas expressões são obtidas após redefinir $Q = \frac{|X(l+1)|}{|X(l-1)|}$ e $S = \frac{|X(l-1)|}{|X(l)|}$ [16].

Tal como demonstrado em [25], o erro máximo absoluto de estimação relativo à largura do bin é consistentemente inferior a 0.1 %.

3.5.2. Estimador de Dun e Liu

O estimador de frequência proposto por Dun e Liu [24] também utiliza três amostras de frequência do espectro - o pico e os seus dois vizinhos - e define três regras de estimação diferentes para a estimação da frequência quando $0.0 \leq \Delta l \leq 1.0$.

No artigo [24], Dun e Liu obtêm uma relação entre $\alpha_0 = \frac{|X(l-1)|}{|X(l+1)|}$ e Δl dada por:

$$\alpha_0 = \frac{(1-\Delta l)(2-\Delta l)}{\Delta l(1+\Delta l)}, \tag{3.20}$$

a partir da qual é definido o estimador.

Deste modo, o estimador proposto é a combinação das três equações seguintes [9]:

$$\widehat{\Delta l} = \frac{|X(l)|-|X(l-1)|}{|X(l)|+|X(l-1)|}, \text{ se } \widehat{\alpha_0} > \alpha_0 \mid \Delta l = 0.5 - \gamma/2; \tag{3.21}$$

$$\widehat{\Delta l} = \frac{3+\alpha_0-\sqrt{\alpha_0^2+14\alpha_0+1}}{2(1-\alpha_0)}, \text{ se } \alpha_0 \neq 1 \text{ e } \alpha_0 \mid \Delta l = 0.5+\gamma/2 \leq \widehat{\alpha_0} \leq \alpha_0 \mid \Delta l = 0.5-\gamma/2; \quad (3.22)$$

$$\widehat{\Delta l} = \frac{2|X(l+1)|}{|X(l)|+|X(l+1)|}, \text{ se } \widehat{\alpha_0} < \alpha_0 \mid \Delta l = 0.5+\gamma/2. \quad (3.23)$$

Pode ocorrer o caso em que $\alpha_0 = 1$, sendo que o valor da estimação será $\widehat{\Delta l} = 0.5$.

O facto de a janela utilizada ser sinusoidal, juntamente com a fina resolução e a baixa complexidade, tornam este algoritmo uma escolha conveniente para aplicações áudio baseadas na transformada real *Modified Discrete Cosine Transform* (MDCT) [24].

3.5.3. Estimador proposto

O estimador proposto baseia-se nas regras de estimação propostas por Dun e Liu. A diferença está na aplicação destas, pois para cada condição, são combinadas duas ou mais regras.

Deste modo, o estimador proposto é definido pelas seguintes equações: se $\widehat{\alpha_0} > \alpha_0 \mid \Delta l = 0.5-\gamma/2$, o valor da estimação é dado por

$$\widehat{\Delta l} = \frac{1}{2}\left(\frac{|X(l)|-|X(l-1)|}{|X(l)|+|X(l-1)|} + \frac{3+\alpha_0-\sqrt{\alpha_0^2+14\alpha_0+1}}{2(1-\alpha_0)}\right); \quad (3.24)$$

se $\alpha_0 \neq 1$ e $\alpha_0 \mid \Delta l = 0.5+\gamma/2 \leq \widehat{\alpha_0} \leq \alpha_0 \mid \Delta l = 0.5-\gamma/2$, o valor da estimação é dado por

$$\widehat{\Delta l} = \frac{1}{3}\left(\frac{|X(l)|-|X(l-1)|}{|X(l)|+|X(l-1)|} + \frac{3+\alpha_0-\sqrt{\alpha_0^2+14\alpha_0+1}}{2(1-\alpha_0)} + \frac{2|X(l+1)|}{|X(l)|+|X(l+1)|}\right); \quad (3.25)$$

se $\widehat{\alpha_0} < \alpha_0 \mid \Delta l = 0.5+\gamma/2$, o valor da estimação é dado por

$$\widehat{\Delta l} = \frac{1}{2}\left(\frac{2|X(l+1)|}{|X(l)|+|X(l+1)|} + \frac{3+\alpha_0-\sqrt{\alpha_0^2+14\alpha_0+1}}{2(1-\alpha_0)}\right). \quad (3.26)$$

Também aqui pode ocorrer o caso em que $\alpha_0 = 1$, sendo que o valor da estimação será $\widehat{\Delta l} = 0.5$.

3.6. Testes e resultados

3.6.1. Condições de teste

O sinal sinusoidal de entrada utilizado para obter o desempenho dos diferentes estimadores de frequência anteriormente apresentado é definido como:

$$x_i(n) = A\ \sin\left[\frac{2\pi}{N}(l+\Delta l)n+\ \varphi\right] + r(n), \tag{3.27}$$

onde $A = 1$, $l = 20$, Δl varia de 0.0 a 1.0 em passos de 0.01, $N = 512$, φ tem valor aleatório e $r(n)$ é o ruído, gerado para cada valor específico de Δl, de acordo com o SNR desejado, num intervalo de -10 dB a 70 dB, com passos de 2.5 dB, em que $SNR = 10\log_{10}(A^2/\sigma^2)$.

De modo a avaliar o desempenho dos estimadores foram realizados testes para o sinal indicado na equação (3.27), sem e com interferência harmónica, a qual é sujeita a uma modulação sinusoidal combinada: modulação em amplitude (AM, *Amplitude Modulation*) e modulação em frequência (FM, *Frequency Modulation)*. Com interferência harmónica, o sinal é dado pela seguinte equação:

$$x_m(n) = A\ \sin\left[\frac{2\pi}{N}(l+\Delta l)n+\ \varphi\right] + r(n) +\ A(1+a_m(n))\sum_{k=1,k\neq 2}^{\left\lfloor\frac{N/2}{F0}\right\rfloor}\sin\left[k\ F0\ \left(\frac{2\pi}{N}n + f_m(n)\right)\right], \tag{3.28}$$

onde $\lfloor . \rfloor$ é o maior inteiro, m é o índice do segmento (cada segmento é composto por N amostras), $A = 1$, $l = 20$, Δl varia de 0.0 a 1.0 em passos de 0.01, $N = 512$, φ tem valor aleatório e $r(n)$ é o ruído gerado para cada valor específico de Δl de acordo com o SNR desejado, num intervalo de -10 dB a 70 dB, em passos de 2.5 dB. A interferência harmónica é modulada em AM e FM utilizando:

$$a_m(n) = \alpha\sin\left(\frac{2\pi}{N}\theta n + \varphi_{am}\right) \tag{3.29}$$

e

$$f_m(n) = \beta\sin\left(\frac{2\pi}{N}\theta n + \varphi_{fm}\right), \tag{3.30}$$

assumindo uma frequência de amostragem de 22050 Hz e um desvio de frequência de 6.5 Hz e, por isso, $\theta = \frac{6.5x512}{22050} \approx 0.151$, $\alpha = 0.045$ e $\beta = \frac{\alpha}{\theta} \approx 0.298$.

O primeiro teste realizado com interferência harmónica, de acordo com a equação (3.28), tem uma frequência fundamental $F0$ de 10.25 bins da DFT. Tal valor foi escolhido de modo a que a frequência da sinusoide alvo seja aproximadamente um harmónico de $F0$, isto é, $3F0 - (l + 1.0) = l - F0$. Deste modo, o sinal é afetado por interferência de dois harmónicos.

O segundo teste realizado com interferência harmónica tem uma frequência fundamental $F0$ de 5.125 bins da DFT, pelo que neste caso a expressão aplicada é $5F0 - (l + 1.0) = l - 3F0$.

Deste modo, o sinal é afetado por interferência de quatro harmónicos, o que implica que, na equação (3.28), k seja diferente de 4, e não diferente de 2 como no caso de teste anterior.

Em suma, os testes realizados, cujos resultados são apresentados de seguida, foram os seguintes:

- o sinal de entrada, de acordo com a equação (3.27), não é afetado por interferência harmónica;
- o sinal de entrada, de acordo com a equação (3.28), é afetado por interferência de dois harmónicos, utilizando $F0 = 10.25$ bins da DFT;
- o sinal de entrada, de acordo com a equação (3.28), é afetado por interferência de quatro harmónicos, utilizando $F0 = 5.125$ bins da DFT.

Os resultados obtidos para cada estimador de frequência são representados como o erro quadrático médio (RMSE, *Root Mean Square Error*) (em % de largura de bin normalizada) em função do SNR, tendo como referência o CRLB normalizado dado pela equação (3.8).

Todos os testes foram realizados utilizando o *software* MATLAB.

3.6.2. Discussão de resultados

Para o primeiro teste, em que o sinal sinusoidal não é afetado por interferência harmónica, os resultados obtidos com os diferentes estimadores de frequência são ilustrados na figura 3.4.

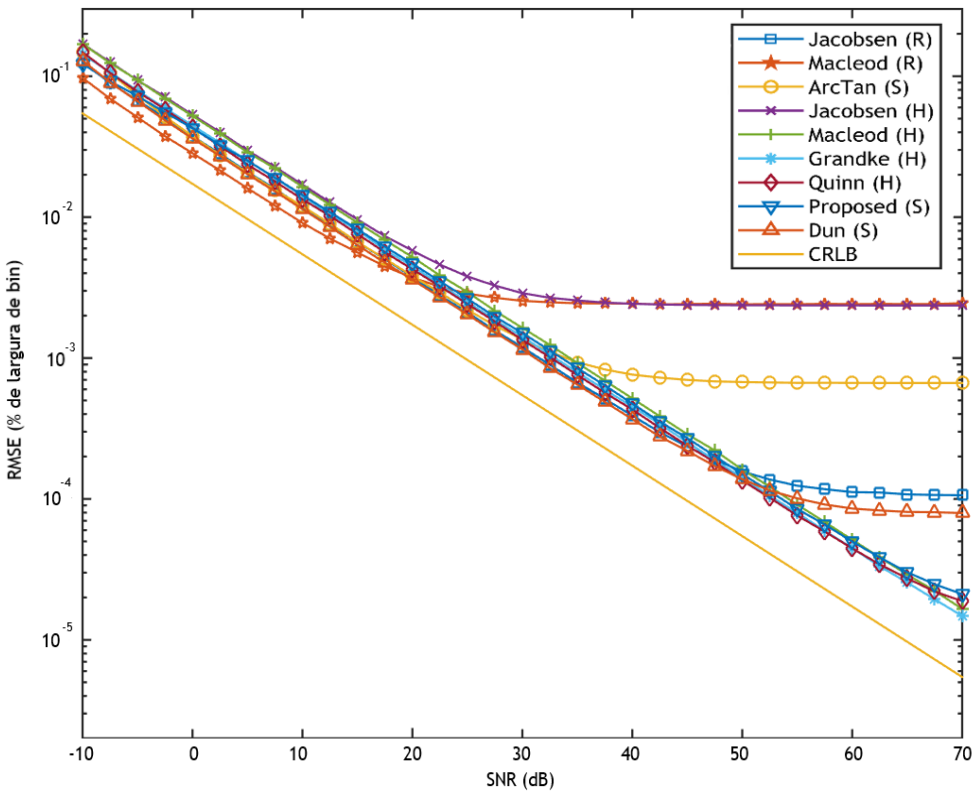

Figura 3.4 - RMSE (em % de largura de bin normalizada) dos oito estimadores de frequência abordados em função do SNR, quando o sinal sinusoidal não é afetado por interferência harmónica. O CRLB está também representado como referência.

Após a análise dos resultados, são identificáveis três regiões onde certos estimadores são mais vantajosos, considerando o CRLB como referência. De forma a facilitar a análise gráfica, utilizou-se a letra H para indicar a utilização da janela de Hanning, a letra R para indicar a utilização da janela retangular e a letra S para indicar a utilização da janela sinusoidal. As regiões identificadas são:

- uma região correspondente aos valores de SNR entre 15 dB e 30 dB onde os estimadores de Jacobsen (H) e de Macleod (R) saturam;
- uma região correspondente aos valores de SNR entre 30 dB e 40 dB onde o estimador ArcTan (S) satura;
- e uma região correspondente aos valores de SNR entre 45 dB e 60 dB onde os estimadores de Jacobsen (R) e de Dun (S) saturam.

Os restantes estimadores, de Grandke (H), proposto (S), de Quinn (H) e de Macleod (H), mostram um desempenho notável para valores de SNR acima de 50 dB, não acusando saturação, sendo a distância à referência CRLB quase constante.

Para o segundo teste, com o sinal sinusoidal afetado por interferência de dois harmónicos, os resultados obtidos são ilustrados na figura 3.5.

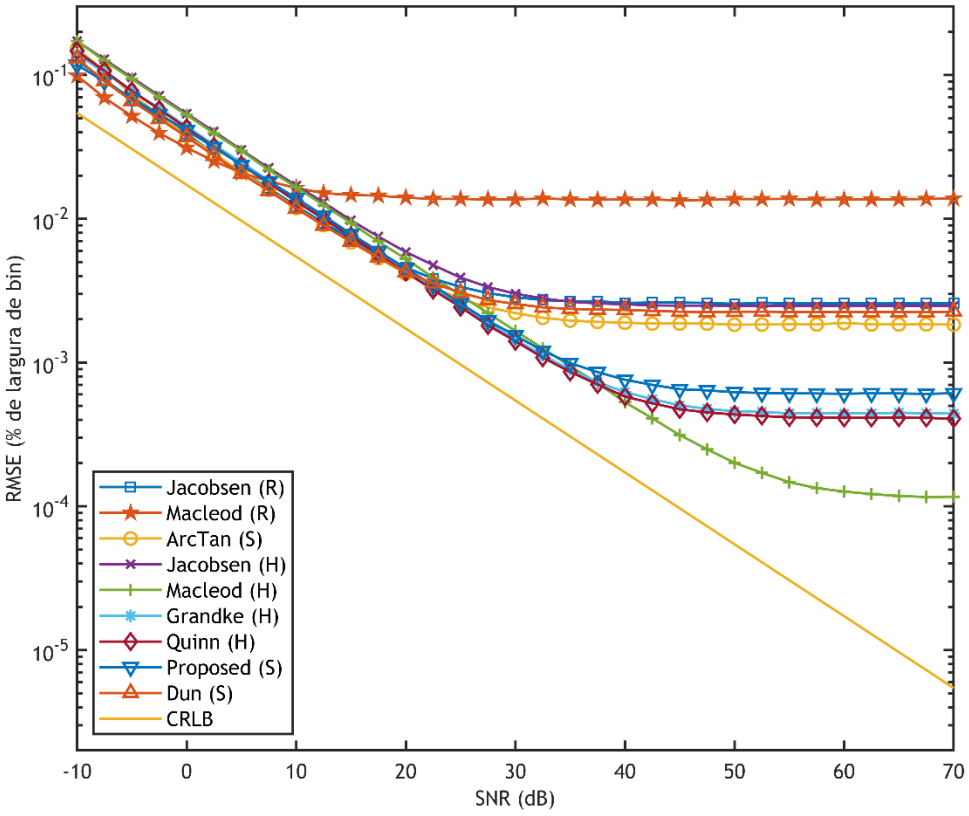

Figura 3.5 - RMSE (em % de largura de bin normalizada) dos oito estimadores de frequência abordados em função do SNR, quando o sinal sinusoidal é afetado por interferência de dois harmónicos. O CRLB está também representado como referência.

Após a análise dos resultados são identificadas quatro regiões onde certos estimadores são mais vantajosos, considerando o CRLB como referência:

- uma região correspondente aos valores de SNR entre -10 dB e 5 dB onde o estimador de Macleod (R) apresenta uma pequena vantagem em relação aos restantes estimadores, apesar de, a partir dos 5 dB, este estimador saturar;
- uma região correspondente aos valores de SNR entre 15 dB e 35 dB onde os estimadores de Jacobsen (H), de Jacobsen (R), de Dun (S) e ArcTan (S) saturam;
- uma região correspondente aos valores de SNR entre 30 dB e 50 dB onde os estimadores de Grandke (H), proposto (S) e de Quinn (H) saturam;
- e uma região correspondente aos valores de SNR acima dos 55 dB onde o estimador de Macleod (H) satura.

Para o terceiro teste, com o sinal sinusoidal afetado por uma interferência mais significativa de quatro harmónicos, os resultados obtidos são ilustrados na figura 3.6.

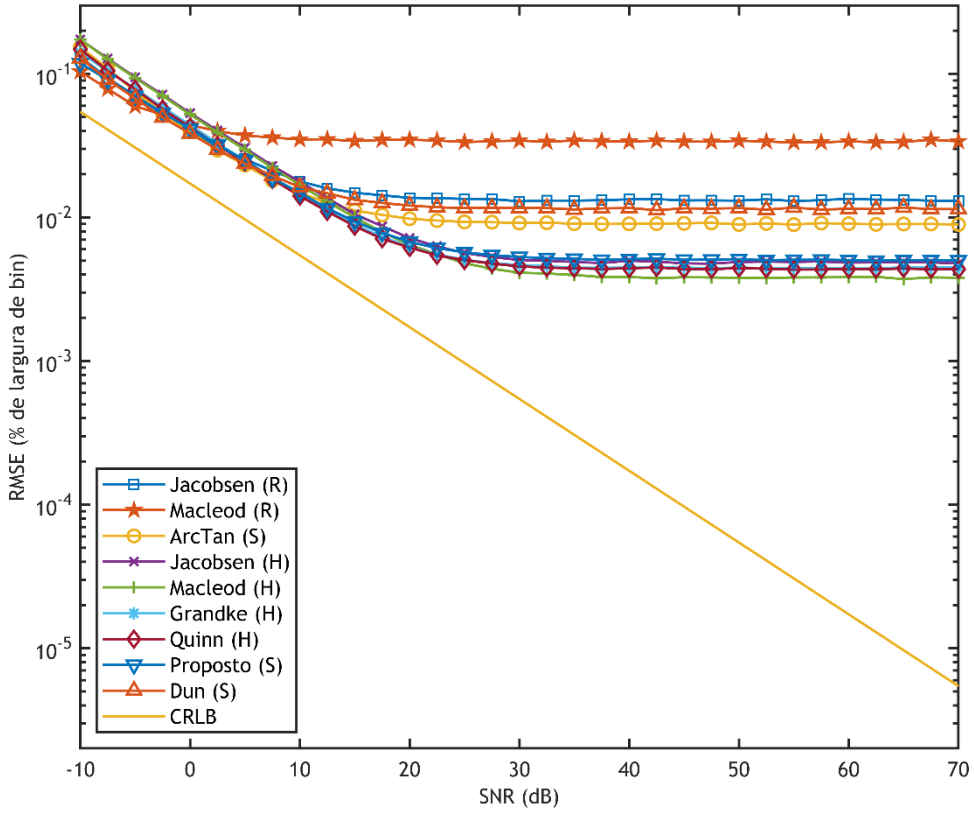

Figura 3.6 - RMSE (em % de largura de bin normalizada) dos oito estimadores de frequência abordados em função do SNR, quando o sinal sinusoidal é afetado por interferência de quatro harmónicos. O CRLB está também representado como referência.

Nesta situação verificam-se três regiões onde certos estimadores são mais vantajosos, considerando o CRLB como referência:

- uma região correspondente aos valores de SNR entre -5 dB e 5 dB onde o estimador de Macleod (R) satura;
- uma região correspondente aos valores de SNR entre 10 dB e 20 dB onde os estimadores de Jacobsen (R), de Dun (S) e ArcTan (S) saturam;
- e uma região correspondente aos valores de SNR entre 15 dB e 30 dB onde os estimadores de Grandke (H), proposto (S), de Jacobsen (H) de Quinn (H) saturam.

De acordo com estes resultados, é de salientar que o estimador de frequência proposto, apresenta um bom desempenho em todos os testes. Porém, é de destacar que, quando comparado com os estimadores que utilizam a janela sinusoidal, apresenta os melhores resultados. Isto mostra que a combinação das três regras de estimação propostas por Dun e Liu [24] possibilita uma melhoria significativa na estimação precisa da frequência (em bins).

3.7. Síntese

Neste capítulo foi abordada uma visão microscópica da estimação precisa da frequência de componentes sinusoidais individuais. De modo a analisar o desempenho de cada um dos estimadores, efetuaram-se testes em diferentes cenários, utilizando-se sinais sintéticos sem e com interferência harmónica. Concluiu-se que o estimador proposto tem um bom desempenho em todos os cenários de testes e o melhor desempenho entre todos os estimadores que utilizam a janela sinusoidal, a qual é uma escolha conveniente para aplicações áudio baseadas em MDCT [24].

Deste modo, o estimador proposto demonstra ser propício ao sucesso da estimação precisa da frequência de componentes sinusoidais individuais, desempenhando um papel fulcral na estimação da frequência fundamental em ambiente macroscópico, descrito no capítulo 4.

Capítulo 4

Estimação da frequência fundamental de uma estrutura harmónica

Neste capítulo apresenta-se a segunda fase do trabalho de estimação da frequência fundamental, na qual é realizada a análise macroscópica. Descrevem-se diferentes algoritmos de estimação da frequência fundamental de sinais de fala e apresentam-se os resultados obtidos após a realização de testes.

4.1. Introdução

A estimação da frequência fundamental em sinais sonoros tem aplicação em diversas áreas de processamento de áudio, tais como na música, comunicações, linguística e patologia da fala [27]. Quando os sinais sonoros são de fala, a estimação da frequência fundamental é importante para aplicações de análise e codificação de fala, reconhecimento de orador e determinação de distúrbios da fala.

Existe uma grande variedade de algoritmos de estimação da frequência fundamental, os quais podem ser divididos em três grupos principais, dependendo da abordagem utilizada: i) métodos no domínio dos tempos; ii) métodos no domínio das frequências; iii) métodos híbridos, recorrendo aos dois domínios anteriores.

4.2. Análise no domínio dos tempos

Os métodos temporais procuram determinar a existência de periocidade no sinal, determinando, desta forma, o seu período fundamental. A estimação do *pitch* consiste em maximizar a correspondência entre uma janela do sinal com uma versão deslocada dessa janela [28]. Assim, o objetivo é favorecer o aparecimento dos melhores candidatos, por exemplo, levando

em consideração as variações de amplitude do sinal ou normalizando a correlação, de modo a evitar a redução para metade ou a duplicação do *pitch* (erros grosseiros) [28].

Os algoritmos de estimação da frequência fundamental no domínio dos tempos abordados são o algoritmo Boersma [29], desenvolvido por Paul Boersma, e ainda o algoritmo YIN [30], desenvolvido por Alain de Cheveigné e Hideki Kawahara.

Outros métodos baseados na mesma metodologia de análise no domínio dos tempos podem ser consultados em [31] e [32].

4.2.1. Algoritmo Boersma

Em 1993, Paul Boersma desenvolveu um algoritmo robusto para deteção de periodicidade baseado na autocorrelação. Este algoritmo é utilizado como um estimador de frequência fundamental numa ferramenta chamada Praat, a qual é uma referência mundial na área de processamento de áudio, nomeadamente de fala. O Praat foi desenvolvido por Paul Boersma e David Weenink no departamento de Ciências Fonéticas da Universidade de Amesterdão [33].

O melhor candidato para o período fundamental de um sinal pode ser encontrado a partir da posição do máximo da função de autocorrelação do sinal sonoro, enquanto o grau de periodicidade, isto é, a relação harmónicos-ruído do sinal sonoro pode ser encontrada a partir da altura relativa deste máximo [29].

Um resumo dos passos executados pelo algoritmo Boersma é dado aqui:

1. ligeiro aumento da taxa de amostragem de modo a remover os lóbulos laterais da janela de Hanning, no domínio das frequências;
2. cálculo do pico global;
3. por segmento, seleção dos pares que são bons candidatos para a periodicidade;
4. definição do caminho através dos candidatos selecionados e associação do custo de cada caminho;
5. escolha do caminho com menos custo, com o auxílio de programação dinâmica.

Segundo o autor, Boersma é um algoritmo que, quando testado com sinais periódicos e com sinais com ruído aditivo ou *jitter*, permite resultados muito mais precisos do que os restantes métodos existentes, sendo capaz de medir relações harmónicos-ruído no domínio dos tempos com precisão e confiabilidade [29].

No artigo [29] e no website [33] é possível encontrar a implementação computacional, os resultados e o *software* desenvolvidos, permitindo uma discussão mais completa e detalhada deste método.

4.2.2. Algoritmo YIN

Alain de Cheveigné e Hideki Kawahara desenvolveram, em 2001, um método de estimação da frequência fundamental, no domínio dos tempos, baseado na função de autocorrelação. Este método designa-se YIN [30] (nome proveniente da filosofia oriental *yin* e *yang*) e propõe-se a corrigir alguns problemas que surgem da autocorrelação e melhorar o seu desempenho.

A autocorrelação faz com que os picos também ocorram nos sub-harmónicos, originando problemas na seleção de qual pico corresponde à frequência fundamental, bem como quais representam harmónicos ou parciais. Em [30], os autores tentam resolver estes problemas.

Os passos executados ao longo do algoritmo YIN são os seguintes [30]:

1. método da autocorrelação;
2. função das diferenças;
3. função *cumulative mean normalized difference*;
4. *threshold* absoluto;
5. interpolação parabólica;
6. estimação local do melhor valor.

Segundo os autores, os dois passos que melhoram substancialmente o desempenho deste algoritmo são os passos 3 e 5. A utilização da função *cumulative mean normalized difference*, em vez de só utilizar a função das diferenças, reduz a ocorrência de erros sub-harmónicos. A função *cumulative mean normalized difference* é dada por [30]:

$$d'_t(\tau) = \begin{cases} 1, se\ \tau = 0 \\ \frac{d_t(\tau)}{\left[(1/\tau)\sum_{j=1}^{\tau} d_t(j)\right]}, se\ \tau \neq 0 \end{cases}. \quad (4.1)$$

A função $d_t(\tau)$ é a função das diferenças simples definida como [30]:

$$d_t(\tau) = \sum_{j=1}^{W}(x_j - x_{j+\tau})^2, \quad (4.2)$$

onde x é uma função periódica e W é o tamanho da janela. A execução de uma interpolação parabólica permite a redução de erros quando a estimativa do período não é um fator do tamanho da janela utilizada [30].

A tabela 4.1 apresenta os passos do algoritmo e o erro associado à medida que estes são implementados. Estes resultados foram retirados do artigo [30] desenvolvido pelos autores do algoritmo YIN, que apresenta os passos do algoritmo e resultados obtidos quando aplicado a uma determinada base de dados.

O algoritmo YIN é relativamente simples, eficiente e pode ser implementado com uma baixa latência. Tem a vantagem de poder ser estendido de diversas maneiras, para que possa ser aplicável a sinais aperiódicos, com frequência fundamental variável e ruído aditivo [30].

Tabela 4.1 - Passos executados pelo algoritmo YIN e erros associados a cada um [30].

Passo do algoritmo YIN	Erro
1. método da autocorrelação	10.0 %
2. função das diferenças	1.95 %
3. função *cumulative mean normalized difference*	1.69 %
4. *threshold* absoluto	0.78 %
5. interpolação parabólica	0.77 %
6. estimação local do melhor valor	0.50 %

No artigo [30] pode encontrar-se uma descrição mais completa e detalhada do algoritmo YIN.

4.3. Análise no domínio das frequências

Os métodos de análise nas frequências permitem a extração de mais informação, já que um sinal no domínio das frequências permite uma melhor visualização da sua estrutura fina, designadamente os harmónicos, a partir dos quais se pode inferir a frequência fundamental. Neste domínio, a determinação do *pitch* depende do surgimento de harmónicos no espectro, sendo que o objetivo é minimizar o risco de redução para metade do *pitch*, por exemplo, explorando a diferença de amplitude entre os harmónicos e os vales entre harmónicos [30].

Os algoritmos de estimação da frequência fundamental no domínio das frequências abordados são o algoritmo *Sawtooth Waveform Inspired Pitch Estimator* (SWIPE) [31], desenvolvido por Arturo Camacho, e o algoritmo SearchTonal, desenvolvido por Aníbal Ferreira.

Outros métodos baseados na mesma metodologia de análise no domínio das frequências podem ser consultados em [34] e [35].

4.3.1. Algoritmo SWIPE

Em 2007, Arturo Camacho desenvolveu um método de estimação da frequência fundamental designado SWIPE, no domínio das frequências, para processamento de voz e música. O SWIPE estima o *pitch* como sendo a frequência fundamental de uma onda dente de serra cujo espectro é mais idêntico ao espectro do sinal de entrada [27].

O teorema de Wiener-Khinchin mostra que a autocorrelação também pode ser calculada como a transformada cosseno inversa de Fourier do quadrado do espectro do sinal; portanto, o algoritmo SWIPE é comparável à autocorrelação, no sentido em que realiza uma transformada integral do espectro utilizando um cosseno como *kernel*. No entanto, o método utiliza a raiz quadrada da magnitude do espectro em vez do seu quadrado [27].

Camacho introduz três modificações principais no *kernel* cosseno de modo a evitar problemas que surjam da autocorrelação [27]: i) coloca a zeros o primeiro quarto do primeiro ciclo

do cosseno; ii) multiplica o *kernel* por uma envolvente que decai numa razão de $1/f$, evitando problemas de periodicidade da função de autocorrelação para sinais periódicos; iii) normaliza o *kernel* de forma a que a largura dos lóbulos espectrais principais coincida com a largura dos lóbulos positivos do cosseno, utilizando uma janela com tamanho dependente do *pitch*, evitando a tendência que a autocorrelação tem de dar valores mais elevados para sinais periódicos com elevada frequência fundamental do que para sinais periódicos com baixa frequência fundamental. Os tipos de sinais que maximizam o produto interno entre o espectro e o *kernel* são sinais periódicos cuja envolvente espectral decai numa razão de $1/f$, motivo pelo qual o algoritmo SWIPE utiliza um sinal com forma de onda em dente de serra [27].

Se um sinal é periódico com frequência fundamental f, o seu espectro deve conter picos múltiplos de f e vales entre eles [27]. A média da distância média de um pico a um vale para os primeiros n picos do espectro é dada por [27]:

$$D_n(f) = \frac{1}{n}\sum_{k=1}^{n} d_k(f) = \frac{1}{n}\left[\frac{1}{2}\left|X\left(\frac{f}{2}\right)\right| - \frac{1}{2}\left|X\left(\left(n+\frac{1}{2}\right)f\right)\right| + \sum_{k=1}^{n}|X(kf)| - \left|X\left(\left(k-\frac{1}{2}\right)f\right)\right|\right], \quad (4.3)$$

onde X é o sinal de entrada no domínio das frequências, n é o número de picos do espectro e f é a frequência fundamental. Um objetivo implícito do algoritmo SWIPE é encontrar a frequência para a qual a distância média de um pico a um vale nos seus harmónicos é maximizada [27].

Os principais passos executados ao longo do algoritmo SWIPE são os seguintes [27]:

1. medição da distância média de um pico a um vale;
2. desfocagem dos harmónicos de modo a permitir inarmonicidade;
3. distorção do espectro;
4. ponderação dos harmónicos;
5. determinação do número de harmónicos a serem utilizados para analisar o *pitch*;
6. distorção da escala de frequência;
7. seleção do tipo e do tamanho da janela, permitindo uma correspondência perfeita.

Uma melhoria do algoritmo é obtida analisando o espectro apenas nos seus primeiros harmónicos primos. Em [27], Camacho apresenta também o método SWIPE', uma modificação do SWIPE que restringe a análise do espectro aos primeiros harmónicos primos, removendo do *kernel* os picos não primos (lóbulos positivos do cosseno) e os seus vales vizinhos.

Os algoritmos SWIPE e SWIPE' foram testados pelo autor utilizando uma base de dados constituída por sinais sonoros de instrumentos musicais e de fala (normal e com anomalias), e o desempenho de cada algoritmo foi comparado com outros doze algoritmos, cuja descrição mais completa pode ser consultada em [27].

Camacho concluiu que o algoritmo SWIPE' tem um desempenho melhor do que todos os outros algoritmos para toda a base de dados e que o algoritmo SWIPE tem o segundo melhor

desempenho para os sinais sonoros de instrumentos musicais e de fala normal, ficando em terceiro lugar quando aplicado a sinais sonoros de fala com anomalias [27].

4.3.2. Algoritmo SearchTonal

O método SearchTonal tem origem num algoritmo desenvolvido por Aníbal Ferreira. Este é um método de estimação da frequência fundamental que implementa a análise *cepstral* [36] e um banco de regras heurísticas de seleção dos candidatos a frequência fundamental e respetiva estrutura harmónica [37].

Os principais passos executados pelo algoritmo SearchTonal são [38]:

1. identificação dos oito candidatos a frequência fundamental mais prováveis através de uma análise *cepstral*;
2. análise detalhada do espectro de magnitude de cada candidato, de modo a calcular a probabilidade daquele candidato considerar aspetos como descontinuidades harmónicas, número total e potência dos parciais harmónicos existentes;
3. por último, todos os candidatos são classificados e o candidato que atingir a maior pontuação de probabilidade é selecionado.

O processo de escolha do valor da frequência fundamental pode ser representado pelo fluxograma representado na figura 4.1.

Nos capítulos 5 e 6 dedicados à análise, modelização e síntese harmónica, o algoritmo SearchTonal é a base para a extração da estrutura harmónica de um sinal de voz humana.

O algoritmo SearchTonal tem vindo a ser utilizado e melhorado de forma a tornar esta estimação precisa e robusta, tal como se pode consultar em [37], [38], [39] e [40].

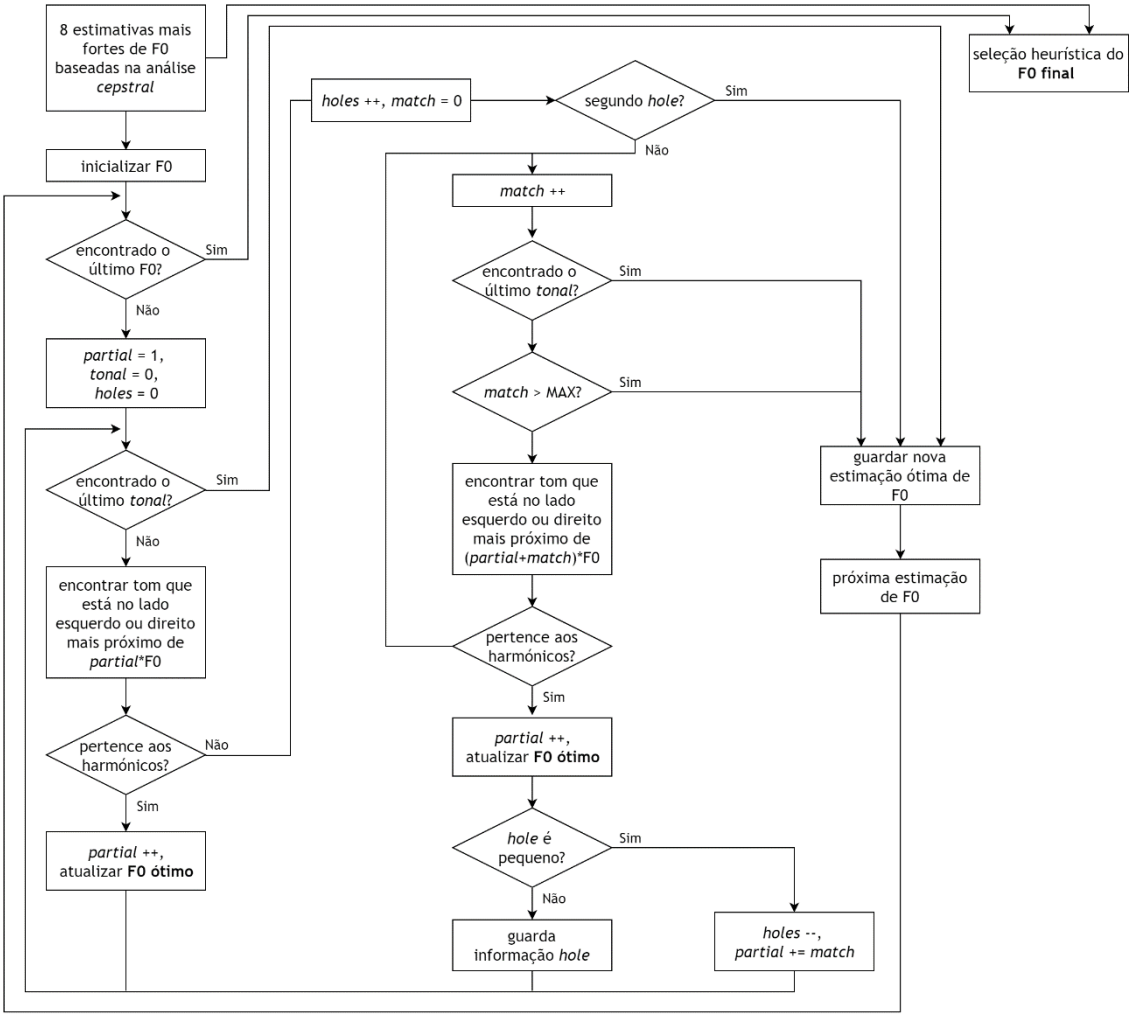

Figura 4.1 - Fluxograma do método de seleção de F0 realizado pelo algoritmo SearchTonal [37]. *(adaptada)*

4.4. Testes e resultados

Nesta secção descrevem-se as condições em que foram realizados os testes dos diferentes algoritmos e apresentam-se também os resultados obtidos.

4.4.1. Condições de teste

Os algoritmos que se pretendem testar e avaliar o seu desempenho na estimação da frequência fundamental são os seguintes, já referidos anteriormente:

- algoritmo Boersma;
- algoritmo YIN;
- algoritmo SWIPE';
- algoritmo SearchTonal.

De modo a avaliar o desempenho dos diferentes algoritmos de estimação da frequência fundamental, foram considerados os seguintes sinais:

- sinais FM sintéticos sem e com interferências harmónicas, não afetados por ruído;
- sinais FM sintéticos sem e com interferências harmónicas, afetados por ruído.

O uso de sinais sintéticos permite uma avaliação fidedigna, pois temos conhecimento do valor real a ser estimado (*ground truth)*.

Cada sinal FM sintético é composto por quatro harmónicos, sendo a frequência fundamental de 270 Hz. A sua duração é de 2 segundos e foi gravado a uma frequência de amostragem de 22050 Hz. Existem as versões sem e com interferências harmónicas, sendo que o sinal sem interferências harmónicas, em que a amplitude de cada harmónico decai com a ordem do mesmo, é definido pela seguinte equação:

$$wave_{FM}(t) = \sum_{k=1}^{4} \frac{A}{k} \sin(2\pi t k F0 + 6\sin(6\pi t)), \tag{4.4}$$

onde $t = 2\ s$ é a duração do sinal, $A = 1$ é a amplitude, $F0 = 270\ Hz$ é a frequência fundamental e k é a ordem do harmónico.

As interferências harmónicas são constituídas por uma oscilação FM e por dois *chirps*, um ascendente e um descendente, em que as amplitudes decaem. Estas interferências foram definidas de modo a não se correlacionarem com a estrutura harmónica definida pela frequência fundamental, tal como indica a seguinte equação:

$$interf(t) = \frac{A}{8}\sin(2\pi t 1.6794 F0 + 2\sin(2\pi 9.743 t)) + \frac{A}{12}\sin(\varphi_0 + 2\pi f(t) t) + \frac{A}{16}\sin(\varphi_0 + 2\pi f(t) t), \tag{4.5}$$

onde $t = 2\ s$ é a duração do sinal, $A = 1$ é a amplitude, $F0 = 270\ Hz$ é a frequência fundamental, $\varphi_0 = 0$ é a fase inicial do sinal *chirp* e $f(t)$ é a função que descreve a variação da frequência instantânea do sinal *chirp*. Neste caso, $f(t)$ é uma função linear dada por $f(t) = \frac{f_f - f_i}{2(t_f - t_i)} t + F0$, onde f_i é a frequência inicial, f_f é a frequência final, t_i é o instante de tempo inicial e t_f é o instante de tempo final, cujos valores são diferentes para os dois *chirps*.

O sinal FM sintético com interferências harmónicas e sem ruído resulta da soma das equações (4.4) e (4.5):

$$wave(t) = wave_{FM}(t) + interf(t). \tag{4.6}$$

Na figura 4.2 está representado o espectrograma do sinal FM sintético sem interferências harmónicas e sem ruído, definido pela equação (4.4). Um espectrograma é uma representação tridimensional da dinâmica espectral de um sinal, incluindo um eixo para as frequências, outro para a magnitude da densidade espectral e um terceiro para os tempos. Visualmente, a figura obtida ao representar um espectrograma é composta por um eixo horizontal que representa o

tempo, um vertical que representa a frequência e as variações na cor indicam a magnitude da densidade espectral a uma dada frequência.

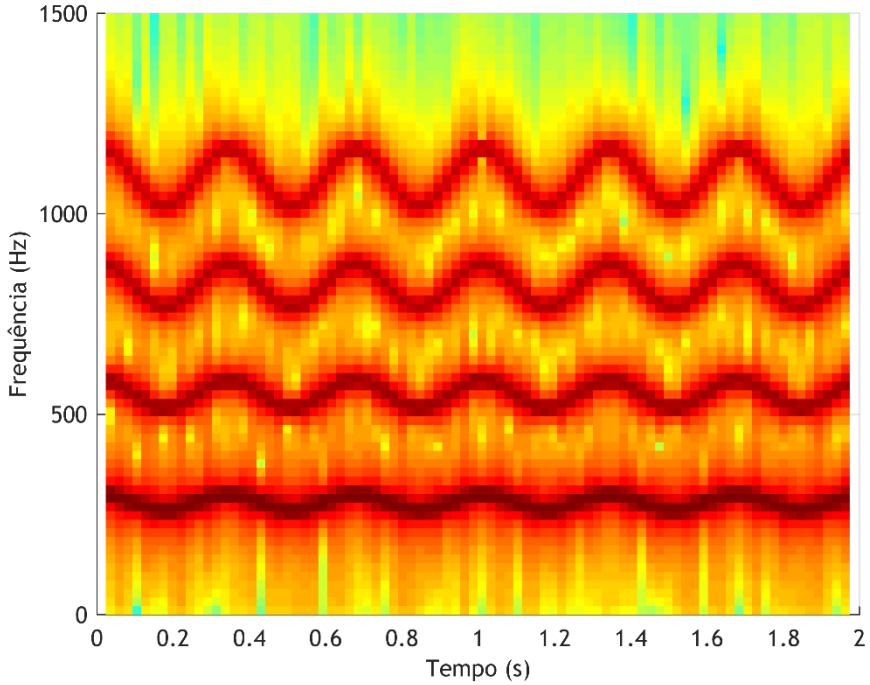

Figura 4.2 - Espectrograma do sinal FM sintético sem interferências harmónicas e sem ruído.

Na figura 4.3 está representado o espectrograma do sinal FM sintético com interferências harmónicas e sem ruído, definido pela equação (4.6).

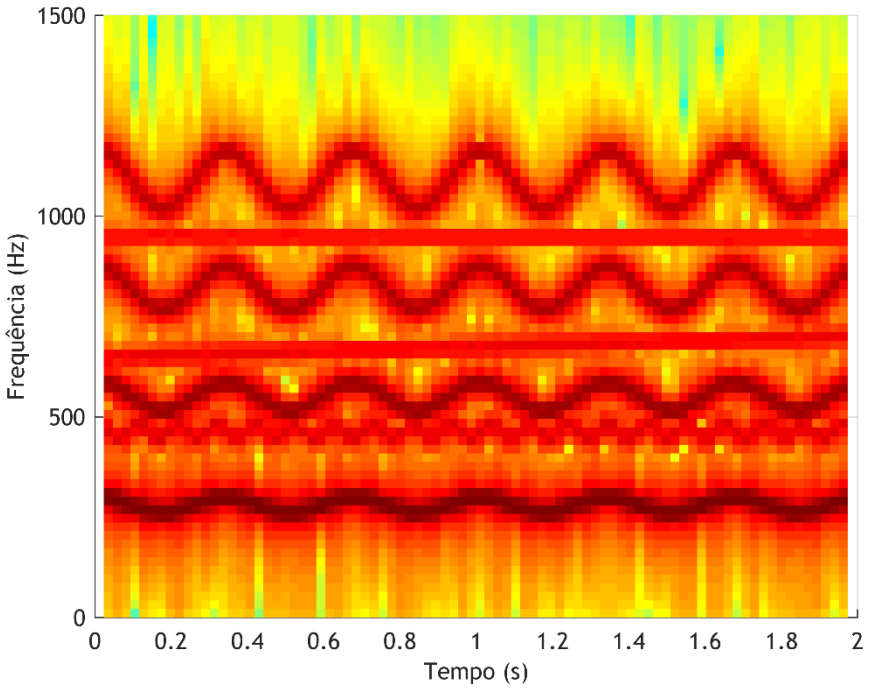

Figura 4.3 - Espectrograma do sinal FM sintético com interferências harmónicas e sem ruído.

De forma a avaliar a robustez dos algoritmos, adiciona-se ao sinal FM, com e sem interferências, ruído gaussiano branco (AWGN, *Additive White Gaussian Noise*) aleatório com valor de SNR de 50 dB. O ruído é definido pela seguinte equação:

$$ruído = \frac{1}{\sqrt{2}\,10^{\frac{SNR}{20}}} randn(), \tag{4.7}$$

onde SNR é dado em dB, $randn()$ é a função que gera valores aleatórios com distribuição gaussiana.

A geração dos sinais FM sintéticos bem como o teste dos diferentes algoritmos de estimação da frequência fundamental foram realizados utilizando o *software* MATLAB.

4.4.2. Discussão de resultados

Para a avaliação do desempenho dos diferentes métodos calculou-se o erro relativo entre a estimação do *pitch* dada por cada um dos respetivos algoritmos e a sinusoide que representa a frequência fundamental do sinal FM sintético.

Nos testes realizados com sinais FM sintéticos, sem e com interferências harmónicas, não afetados por ruído, os resultados dos erros relativos obtidos para cada método de estimação são apresentados na tabela 4.2.

Tabela 4.2 - Erros relativos de estimação de F0 obtidos com os diferentes métodos para sinais FM sintéticos sem e com interferências harmónicas, não afetados por ruído.

Algoritmo de estimação da frequência fundamental	Erro relativo (%) para sinal FM sintético sem interferências harmónicas	Erro relativo (%) para sinal FM sintético com interferências harmónicas
Boersma	0.12682 %	0.054429 %
YIN	0.14608 %	0.081841 %
SWIPE'	0.26859 %	0.74468 %
SearchTonal	0.066656 %	0.064038 %

Na figura 4.4 estão representadas as sinusoides do sinal FM sintético com frequência fundamental de 270 Hz (linha tracejada magenta) e da estimação do *pitch* dada por cada algoritmo (linha contínua azul), para o caso de não haver interferências harmónicas. Na figura 4.5, os sinais representados consideram a presença de interferências harmónicas. Em ambas as figuras é possível observar o *pitch* médio dado por cada algoritmo.

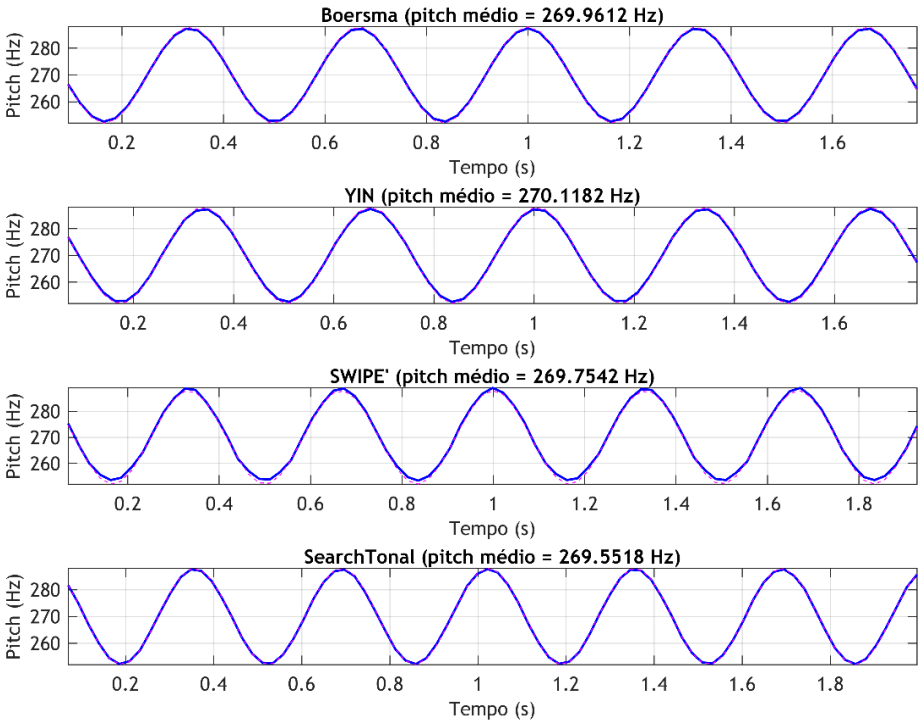

Figura 4.4 - Estimação do *pitch* dada por cada algoritmo (linha contínua azul) e sinal FM sintético sem interferências harmónicas e sem ruído, com F0 de 270 Hz (linha tracejada magenta).

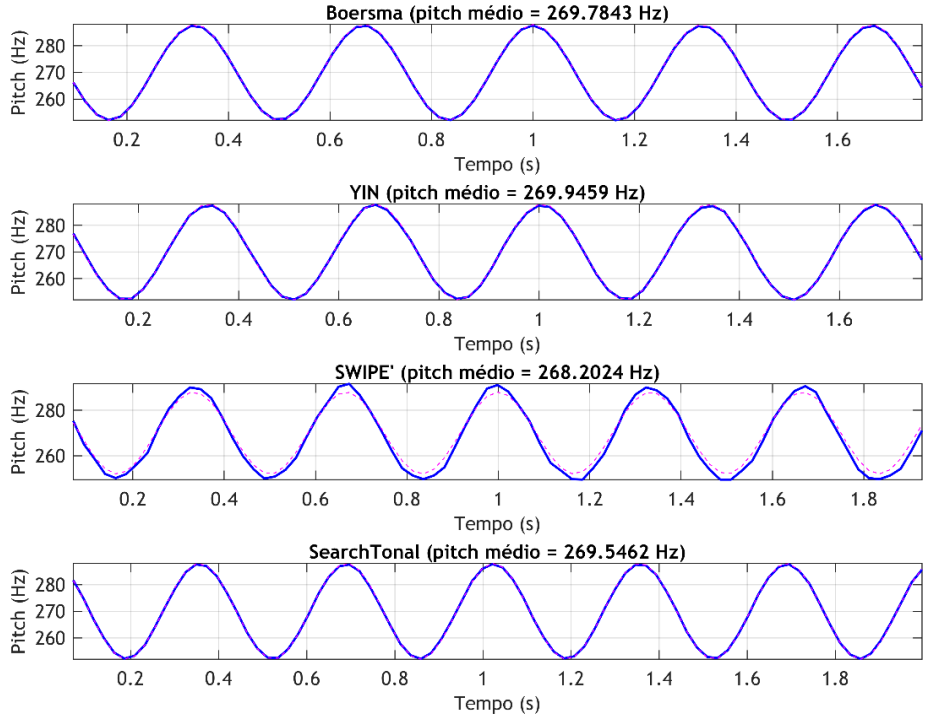

Figura 4.5 - Estimação do *pitch* dada por cada algoritmo (linha contínua azul) e sinal FM sintético com interferências harmónicas e sem ruído, com F0 de 270 Hz (linha tracejada magenta).

A partir dos resultados obtidos (ver tabela 4.2), conclui-se que o algoritmo SearchTonal tem o melhor desempenho no caso da ausência de interferências harmónicas. Quando estão presentes interferências harmónicas, o algoritmo Boersma apresenta um menor erro relativo, apesar de a diferença entre este e o erro relativo do SearchTonal não ser muito significativa. O algoritmo SWIPE' apresenta o pior desempenho, sendo que o seu erro relativo aumenta significativamente com a presença de interferências harmónicas.

De notar que, no caso dos algoritmos Boersma, YIN e SearchTonal, o erro relativo diminui na presença de interferências harmónicas, pois esta permite uma melhor identificação das sinusoides harmónicas do sinal por parte dos algoritmos.

Relativamente ao *pitch* médio, este não traduz o bom ou mau desempenho de um algoritmo. Por exemplo, apesar de o algoritmo SearchTonal ter um melhor erro relativo no caso de não haver interferências harmónicas, o *pitch* médio mais próximo do valor real de 270 Hz é dado pelo algoritmo Boersma. Isto porque quaisquer valores de frequência fundamental dados por um algoritmo que se afastem muito do valor real, influenciam a média final, ou seja, uma oscilação positiva pode ser anulada por uma negativa, a qual acaba por não afetar o valor médio do *pitch* estimado. O mesmo não se verifica com a utilização do erro relativo, pois este é diretamente proporcional à diferença absoluta entre a estimação do *pitch* e o *ground truth*, permitindo uma maior precisão na avaliação do desempenho do método de estimação.

Nos testes realizados com sinais FM sintéticos, sem e com interferências harmónicas, afetados por ruído com SNR de 50 dB, os resultados dos erros relativos obtidos para cada algoritmo de estimação são apresentados na tabela 4.3.

Tabela 4.3 - Erros relativos de estimação de F0 obtidos com os diferentes métodos para sinais FM sintéticos sem e com interferências harmónicas, afetados por ruído com SNR de 50 dB.

Algoritmo de estimação da frequência fundamental	Erro relativo (%) para sinal FM sintético sem interferências harmónicas	Erro relativo (%) para sinal FM sintético com interferências harmónicas
Boersma	0.12685 %	0.12686 %
YIN	0.14614 %	0.14609 %
SWIPE'	0.27162 %	0.27256 %
SearchTonal	0.32106 %	0.3219 %

Na figura 4.6 estão representadas as sinusoides do sinal FM sintético com frequência fundamental de 270 Hz (linha tracejada magenta) e da estimação do *pitch* dada por cada algoritmo (linha contínua azul), para o caso de não haver interferência harmónica. Na figura 4.7, os sinais representados consideram a presença de interferência harmónica. Em ambas as figuras é possível observar o *pitch* médio dado por cada algoritmo.

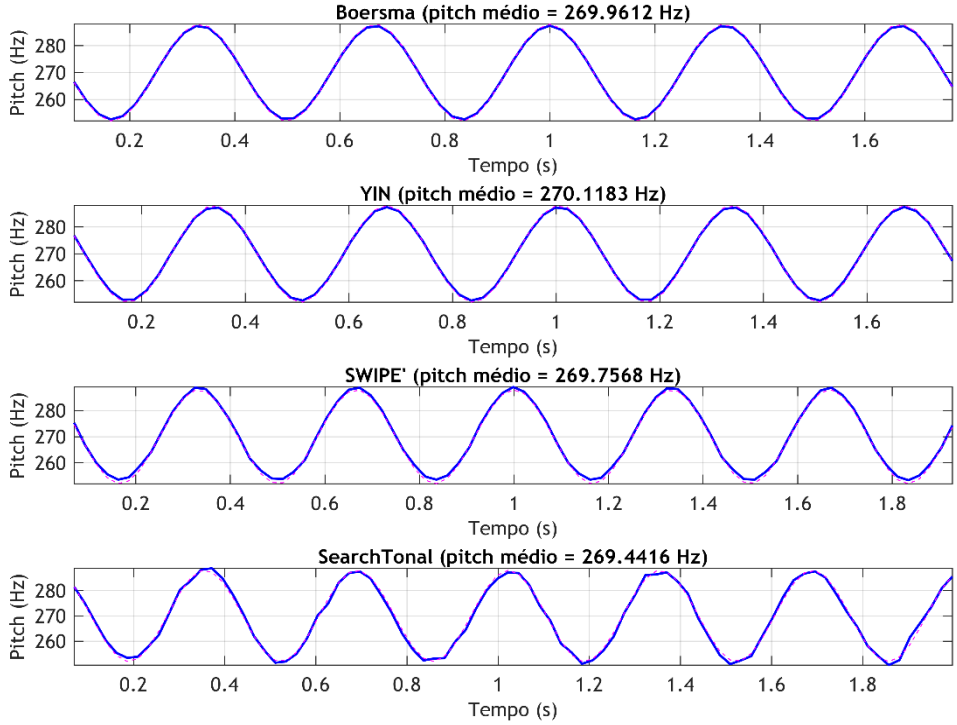

Figura 4.6 - Estimação do *pitch* dada por cada algoritmo (linha contínua azul) e sinal FM sintético sem interferências harmónicas e com ruído, com F0 de 270 Hz e SNR de 50 dB (linha tracejada magenta).

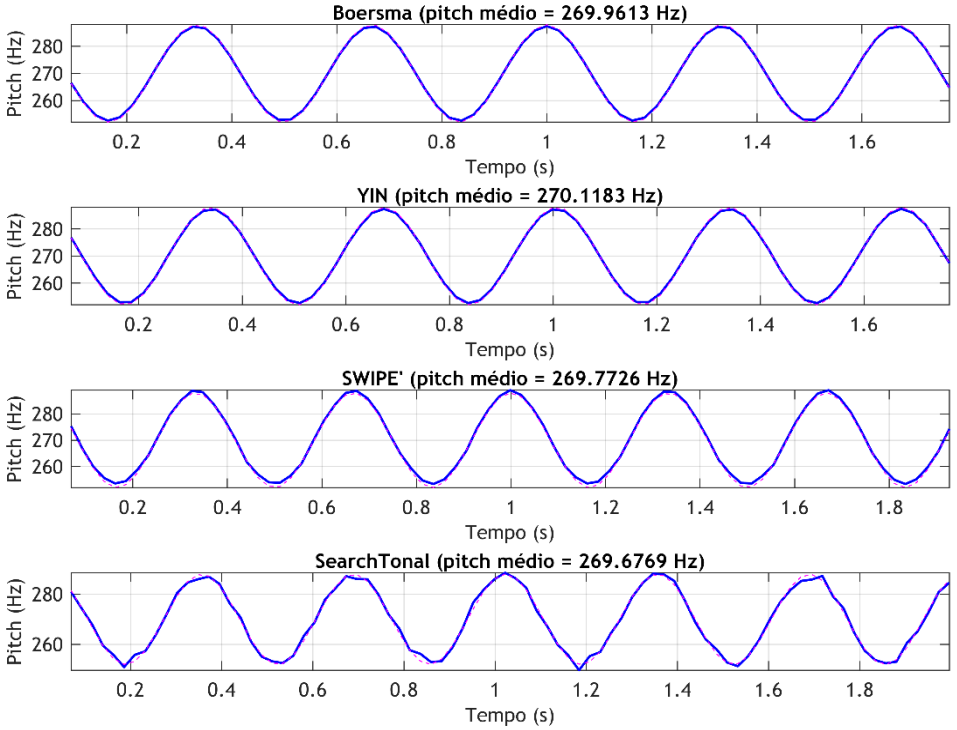

Figura 4.7 - Estimação do pitch dada por cada algoritmo (linha contínua azul) e sinal FM sintético com interferências harmónicas e com ruído, com F0 de 270 Hz e SNR de 50 dB (linha tracejada magenta).

A partir dos resultados obtidos (ver tabela 4.3), conclui-se que o algoritmo Boersma é o mais robusto, quer na presença ou não de interferências harmónicas. O algoritmo YIN apresenta resultados muito semelhantes ao método Boersma, sendo que o algoritmo SWIPE' apresenta um desempenho semelhante ao SearchTonal.

O desempenho do algoritmo SearchTonal é afetado pelo ruído, pois este faz com que alguns parciais do sinal sejam oprimidos e, por isso, não sejam detetáveis. Este é um problema para o algoritmo SearchTonal porque ele identifica todos os harmónicos que estão a pelo menos 5 dB acima do patamar de ruído e também requer que pelo menos três sinusoides harmónicas sejam detetadas para que uma estrutura harmónica seja reconhecida. Por outro lado, os resultados obtidos sugerem que, devido ao facto do SearchTonal esperar um número mínimo de parciais na estrutura harmónica, isto restringe o seu desempenho sob a influência de ruído. No entanto, note-se que os sinais de teste utilizados são sintéticos, pelo que os resultados podem ser diferentes, caso os sinais utilizados sejam de voz falada [38].

Conclui-se que a presença de ruído não afeta o desempenho dos algoritmos de estimação da frequência fundamental para sinais FM sintéticos sem e com interferências harmónicas, tal como se pode verificar pela análise da tabela 4.3, em que os erros relativos para ambos os cenários são semelhantes.

4.5. Síntese

Neste capítulo foram apresentados diferentes algoritmos de estimação da frequência fundamental utilizados em sinais sonoros de voz e música. Efetuou-se uma análise de desempenho e respetiva comparação de algoritmos utilizando sinais FM sintéticos. Concluiu-se que todos os métodos têm um desempenho apreciável do ponto de vista dos erros relativos obtidos face ao sinal FM sintético, sempre inferiores a 0.5 % (à exceção do algoritmo SWIPE' quando o sinal FM sintético tem interferências harmónicas e não é afetado por ruído), o que representa um limite de desempenho aceitável.

Deste modo, tratando-se o SearchTonal de um método preparado para a modelização e síntese harmónica, é concebível a sua utilização na investigação e desenvolvimento apresentados nos capítulos 5 e 6.

Capítulo 5

Análise, modelização e síntese harmónica

Neste capítulo apresenta-se a abordagem utilizada para a modelização de magnitude e de fase paramétrica de sinais sonoros de fala, bem como a abordagem de síntese harmónica no domínio das frequências. Descrevem-se os testes implementados para caracterizar objetiva e subjetivamente diferentes cenários de síntese utilizados e apresentam-se os resultados obtidos.

5.1. Introdução

Os sinais de fala são constituídos por duas componentes principais: i) uma componente periódica que resulta da vibração das pregas vocais na laringe; ii) uma componente de ruído resultante da passagem do ar turbulento por uma constrição da laringe ou pelo trato vocal [1]. Ambas as componentes são importantes no que diz respeito à variedade fonética e ao reconhecimento do orador. No entanto, a componente periódica vozeada contém maior informação acústica devido à sua estrutura de fase e à sua estrutura de magnitude.

O objetivo é a obtenção de modelos de magnitude e de fase espectrais totalmente paramétricos para poderem ser utilizados de forma flexível na representação e síntese de sinais de voz falada, preservando as suas características sonoras naturais. Isto requer que os modelos sejam holísticos e invariantes ao deslocamento. Holístico significa que um modelo deve ser capaz de representar todas as componentes relevantes do sinal no espectro até à frequência de Nyquist. Invariante ao deslocamento significa que, para um sinal quase-periódico, um modelo de representação deve ser válido, ou mudar muito lentamente, quando o sinal periódico é deslocado para a direita ou para a esquerda no tempo. Além disto, flexibilidade também requer que os modelos de magnitude espectrais sejam independentes dos modelos de fase.

Utilizando diferentes modelos de magnitude e de fase espectrais paramétricos é possível proceder à síntese e reconstrução harmónica de um sinal de fala no domínio das frequências. De modo a testar e caracterizar objetiva e subjetivamente resultados para diferentes cenários

de modelização harmónica, foram utilizados sinais de fala vozeados correspondentes a vogais sustentadas. Estes sinais fazem parte da base de dados destinada ao projeto FCT, dedicado à reconstrução da voz humana, no qual se insere esta dissertação.

Os testes foram realizados utilizando o *software* MATLAB e, tal como já referido, o método utilizado para a extração da estrutura harmónica dos sinais de voz falada é o SearchTonal.

5.2. Análise e modelização harmónica

A estrutura de processamento de sinal que realiza a análise e a modelização harmónica pode ser representada pelo diagrama de blocos da figura 5.1.

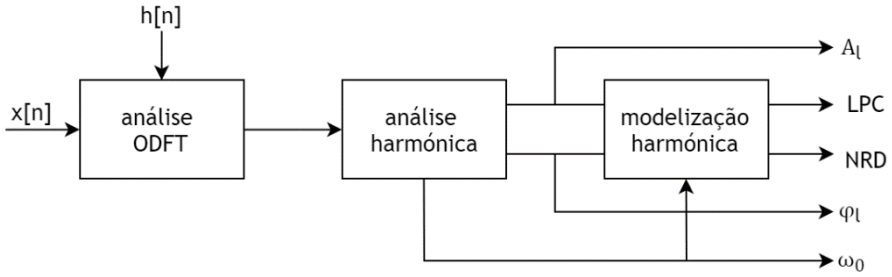

Figura 5.1 - Diagrama de blocos da análise e modelização harmónica de sinais de voz que contêm uma estrutura harmónica.

Antes da análise harmónica, é realizada uma análise inicial em que se aplica uma janela temporal, representada por $h[n]$ e correspondente à raiz quadrada de uma janela de Hanning deslocada, ao sinal de entrada, representado por $x[n]$. De seguida, é efetuado o cálculo da DFT de frequência ímpar (ODFT, *Odd-frequency Discrete Fourier Transform*) com uma sobreposição de sinal de 50 % [41]. Esta primeira análise é tipicamente utilizada em codificação de áudio, pois permite a reconstrução perfeita na ausência de modificação espectral [42] [43].

A análise harmónica diz respeito à estimação precisa da frequência, fase e magnitude de todas as componentes sinusoidais pertencentes à estrutura harmónica do sinal de voz, utilizando o estimador proposto no capítulo 3, na subsecção 3.5.3. Se l é o índice do harmónico, as magnitudes harmónicas são representadas por A_l e as fases harmónicas são representadas por φ_l. A frequência fundamental é representada por ω_0.

A modelização harmónica diz respeito à obtenção de dois modelos: i) modelo de magnitude holístico, representado pelo modelo preditivo linear (LPC, *Linear Predictive Model*); ii) modelo de fase invariante ao deslocamento, representado pelo atraso relativo normalizado (NRD, *Normalized Relative Delay*). É utilizado o modelo LPC devido à sua habilidade de capturar explicitamente as frequências dos formantes, os quais representam informação fonética importante [8] [44] [45]. Utiliza-se o modelo NRD, pois este é completamente independente, preservando a forma invariante do sinal [46] [47] [48]. Estas duas modelizações são abordadas e detalhadas nas próximas subsecções 5.2.1 e 5.2.2.

5.2.1. Modelização harmónica de magnitude holística

A estimação harmónica devolve as frequências e correspondentes magnitudes de todos os harmónicos, as quais estão assinaladas na figura 5.2 por triângulos vermelhos. O sinal representado na figura 5.2 representa então o espectro de magnitude de um segmento de voz falada sustentada referente à vogal /a/ da palavra "água" produzida por um orador do género feminino (linha contínua azul).

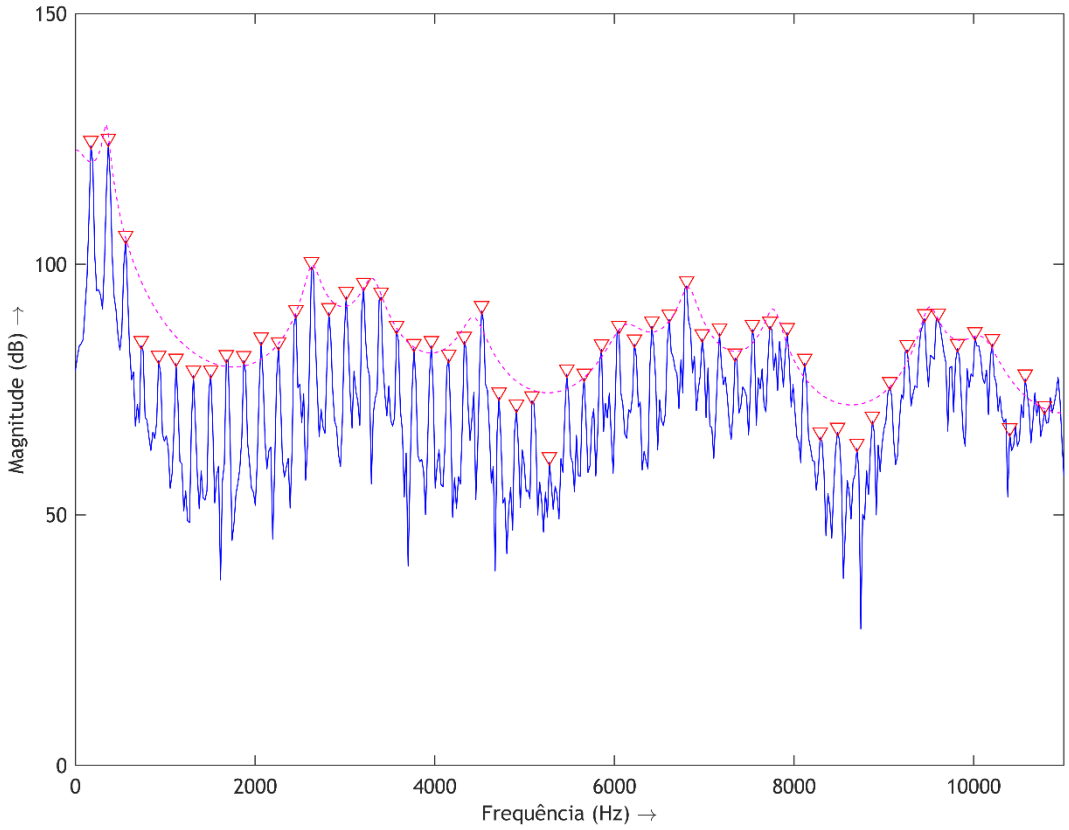

Figura 5.2 - Espectro de magnitude de um segmento de voz falada sustentada referente à vogal /a/ da palavra "água" produzida por um orador do género feminino (linha contínua azul). Os triângulos vermelhos assinalam todos os harmónicos pertencentes à estrutura harmónica e a linha tracejada magenta representa o modelo LPC do espectro de magnitude definido por todos os harmónicos.

Os passos realizados ao longo da modelização harmónica de magnitude são os seguintes:

1. estimação da densidade espectral de potência (PSD, *Power Spectral Density*) do espectro de magnitude por interpolação entre as magnitudes de todos os harmónicos detetados, numa escala em dB;
2. cálculo dos coeficientes de autocorrelação tirando vantagem do teorema de Wiener-Khinchin;
3. obtenção do modelo LPC utilizando a recursão de Levinson-Durbin [49].

Em suma, a modelização harmónica de magnitude é conseguida utilizando um modelo LPC de ordem 22, o qual é adequado para sinais com uma frequência de amostragem de 22050 Hz e para a diversidade de vogais da língua portuguesa.

5.2.2. Modelização harmónica de fase invariante ao deslocamento

Seja $x[n]$ um sinal harmónico periódico discreto nos tempos dado pela seguinte equação:

$$x[n] = \sum_{l=0}^{L-1} A_l \sin(n\omega_l + \varphi_l), \tag{5.1}$$

onde L são os harmónicos, até à frequência de Nyquist, que compõem o seu espectro de magnitude. Neste caso, os valores de fase dos L harmónicos, representados por φ_l, denotam a estrutura de fase do sinal.

No entanto, esta estrutura de fase não é invariante ao deslocamento, pois, deslocando a forma de onda em n para a direita ou para a esquerda, todos os valores de fase mudam. Se o sinal é harmónico, então a frequência de cada harmónico é dada por $\omega_l = (l+1)\omega_0$ e a equação (5.1) pode ser reescrita como [48]:

$$x[n] = \sum_{l=0}^{L-1} A_l \sin[(l+1)(n\omega_0 + \varphi_0) + 2\pi NRD_l]. \tag{5.2}$$

NRD é o atraso relativo normalizado definido pela seguinte equação [50]:

$$NRD_l = \frac{\varphi_l - (l+1)\varphi_0}{2\pi},\ l = 0, 1, \dots, L-1, \tag{5.3}$$

e reflete simplesmente um valor normalizado entre 0 e 1 que depende da diferença entre a fase do harmónico e a fase da frequência fundamental [48], pelo que o número de coeficientes NRD é igual ao número de harmónicos.

Outras propriedades importantes dos coeficientes NRD são as seguintes [48]:

- sendo uma característica relativa relacionada com a fase, o NRD da frequência fundamental é zero por definição;
- como representa a fase normalizada, o vetor de característica NRD pode ser *wrapped* (valores entre 0 e 2π radianos) e *unwrapped* (contrário de *wrapped*, isto é, valores representados de forma contínua);
- os coeficientes NRD são intrinsecamente invariantes ao deslocamento e também independentes da frequência fundamental.

Na figura 5.3 está representada uma sobreposição de todos os vetores de característica NRD *unwrapped* extraídos do segmento de voz falada sustentada referente à vogal /a/ da palavra "água" produzida por um orador do género feminino, representado na figura 5.2. Observa-se que os primeiros 20 coeficientes NRD são bastante consistentes, podendo ser representados por um modelo simples cujo suporte de frequência engloba os primeiros quatro formantes da voz [48]. A dispersão dos coeficientes NRD ocorre a partir do harmónico de índice 25 e deve-se ao facto de as magnitudes espectrais correspondentes atingirem o patamar de ruído [48]. Deste modo, a fase é fortemente afetada pela interferência do ruído e também pela dos restantes

harmónicos, pelo que a sua estimação é pouco fiável. Quando isto não acontece, pode dizer-se que os coeficientes NRD expressam relações de fase que, em complemento à magnitude dos harmónicos, definem completamente a invariância de forma de onda de um sinal harmónico, independentemente da sua frequência fundamental [46] [47] [48].

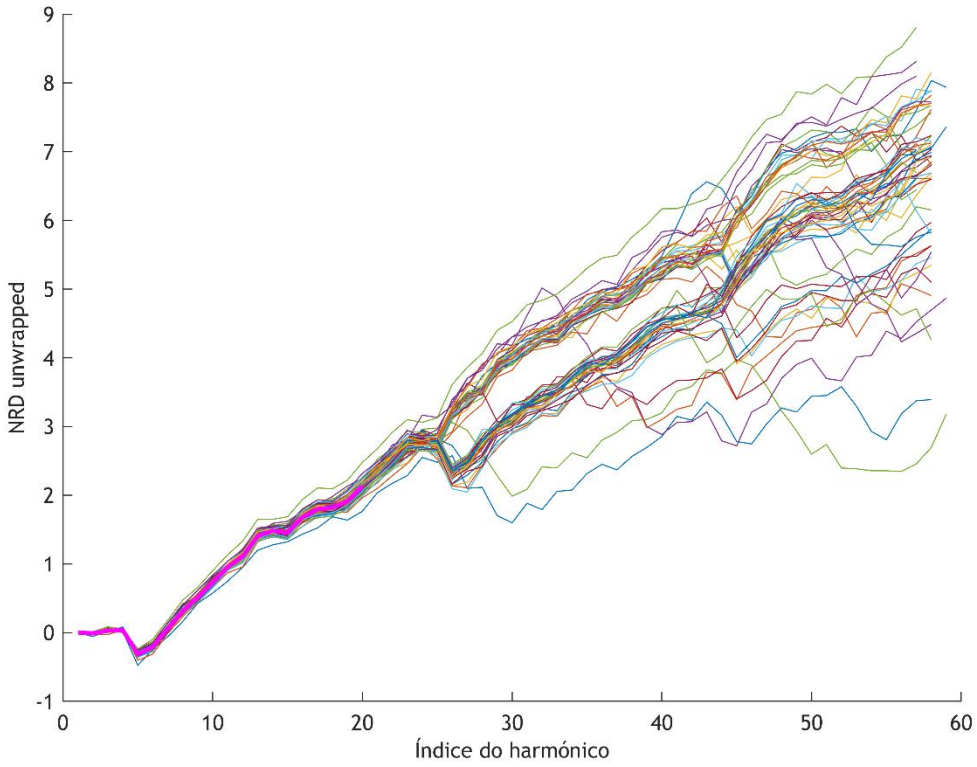

Figura 5.3 - Sobreposição dos vetores de característica NRD *unwrapped* extraídos do segmento de voz falada sustentada referente à vogal /a/ da palavra "água" produzida por um orador do género feminino. A linha contínua magenta representa a média do vetor NRD *unwrapped* até ao harmónico de índice 20.

Outros descritores de fase harmónica similares ao NRD foram também propostos por [51], [52] e [53].

5.3. Síntese harmónica no domínio das frequências

A síntese harmónica do sinal $x[n]$ no domínio das frequências pode ser representada pelo diagrama de blocos da figura 5.4.

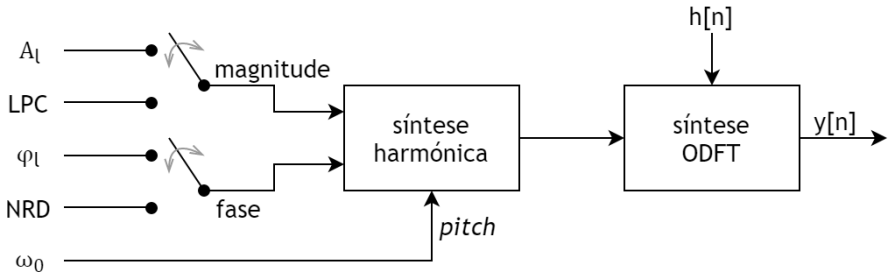

Figura 5.4 - Diagrama de blocos da síntese harmónica baseada em segmentos, utilizando os parâmetros exatos de magnitude ($\boldsymbol{A_l}$) e fase ($\boldsymbol{\varphi_l}$) ou, alternativamente, as aproximações correspondentes dadas pelo modelo harmónico LPC e pelo modelo NRD invariante ao deslocamento.

O processo de síntese é baseado em segmentos e espelha essencialmente os passos inversos da análise e modelização harmónica (ver figura 5.1), fazendo com que a combinação da análise e modelização harmónica com a síntese harmónica permita uma reconstrução perfeita do sinal na ausência de modificação espectral [42] [43]. A síntese pode ser realizada utilizando: i) ou os parâmetros exatos de magnitude, representado por A_l, e fase, representado por φ_l; ii) ou as aproximações correspondentes dadas pelo modelo harmónico LPC e pelo modelo NRD invariante ao deslocamento.

Um passo fulcral no processo de síntese do sinal é o mapeamento das magnitudes e das fases harmónicas em coeficientes espectrais ODFT ou bins, já que o domínio das frequências é discreto. Assume-se que o comprimento da transformada ODFT é representado por N e é igual ao comprimento da janela $h[n]$ de análise e síntese, de tal modo que $N \geq 3L$, em que L é o número de harmónicos. Assume-se também que o lóbulo principal da resposta em frequência da janela temporal tem uma largura que não excede $3x2\pi/N$. Em [14], [25] e [41] mostra-se que a janela que cumpre esta condição é a janela meio seno, ou seja, a janela que corresponde à raiz quadrada de uma janela de Hanning deslocada.

As equações de síntese utilizadas são as apresentadas em [41] com as modificações sugeridas em [54].

5.4. Testes e resultados

Nesta subsecção apresentam-se as condições de teste em que foram realizados os testes da síntese harmónica no domínio das frequências, bem como os resultados de avaliação objetivos e subjetivos.

5.4.1. Condições de teste

De modo a avaliar a qualidade da síntese harmónica no domínio das frequências, foram realizados testes para quatro cenários diferentes:

A) síntese utilizando as magnitudes e fases harmónicas exatas (tais como estimadas), ou seja, utilizando os parâmetros A_l e φ_l;
B) síntese utilizando o modelo harmónico LPC e as fases harmónicas exatas φ_l (tais como estimadas);
C) síntese utilizando o modelo harmónico LPC e o modelo NRD médio do orador;
D) síntese utilizando o modelo harmónico LPC, o modelo NRD médio do orador e fase sintética para a frequência fundamental (em substituição da fase original da frequência fundamental - de notar que esta modificação provoca falta de sincronismo em relação à forma de onda original o que invalida, por exemplo, a utilização de métricas como a SNR).

Em todos os testes foi forçado o alinhamento harmónico da estrutura harmónica.

Foram utilizados quatro sinais de teste correspondentes a duas vogais sustentadas, /a/ da palavra "água" e /i/ da palavra "ilha", produzidas por um orador do género feminino (SPF, *Speaker* Feminino) e por um orador do género masculino (SPM, *Speaker* Masculino). Todos os sinais têm uma duração de 1.45 segundos.

5.4.2. Avaliação objetiva e subjetiva

Para efeitos de avaliação objetiva calculou-se o nível de SNR entre o sinal original e o sinal após a síntese.

Na tabela 5.1 apresentam-se os valores da frequência fundamental média de cada orador para os sinais de fala produzidos e os valores SNR para cada sinal testado em cada um dos cenários de teste de síntese harmónica no domínio das frequências.

Tabela 5.1 - Valores da frequência fundamental média e valores de SNR para as vogais de fala sustentada testadas em cada um dos cenários de teste de síntese harmónica no domínio das frequências.

Orador	F0 média	Cenário A (A_l e φ_l)	Cenário B (LPC e φ_l)	Cenário C (LPC e NRD)	Cenário D (LPC, NRD e F0 sintética)
SPF /i/	187.98 Hz	24.14 dB	9.77 dB	9.71 dB	5.36 dB
SPF /a/	180.11 Hz	19.26 dB	10.28 dB	7.38 dB	2.03 dB
SPM /i/	113.65 Hz	25.54 dB	11.86 dB	11.51 dB	-4.76 dB
SPM /a/	109.75 Hz	23.79 dB	13.15 dB	11.24 dB	-2.57 dB

Analisando a tabela 5.1, verifica-se que o valor de SNR é maior no cenário de teste em que são utilizadas as magnitudes e fases harmónicas exatas (tais como estimadas) na síntese (cenário A). A utilização do modelo de magnitude harmónico LPC faz com que o valor de SNR diminua de forma significativa. No entanto, quer no cenário em que se utilizam as fases exatas (cenário B), quer no cenário em que se utiliza o modelo NRD médio do próprio orador (cenário C), os valores de SNR são semelhantes, concluindo-se que a utilização do modelo NRD não acrescenta artefactos disruptivos quando comparados com o cenário B. Por último, o SNR é particularmente pobre quando a síntese é realizada utilizando-se os modelos LPC e NRD juntamente com a fase sintética para a frequência fundamental, o que revela um desalinhamento entre os sinais originais e os sinais sintetizados.

Apesar dos resultados objetivos obtidos, é importante notar que estes podem não ser indicativos de um impacto percetual auditivo significativo. Por esta razão, os sinais sintéticos obtidos foram comparados auditivamente com os sinais originais e foi realizada uma avaliação subjetiva simples, apenas para perceber até que ponto as diferenças entre os sinais têm um impacto significativo perceptualmente.

À semelhança dos resultados obtidos na análise objetiva, verificou-se através de uma avaliação percetual que a qualidade sonora diminui do cenário A para o cenário D. É de notar que a síntese, utilizando os valores exatos (tais como estimados) das magnitudes e fases harmónicas, produz um sinal com características auditivas idênticas às do sinal original. No entanto, é de salientar que apesar dos restantes cenários de síntese produzirem sinais com características auditivas que os diferenciam do sinal original, tais artefactos não produzem alterações disruptivas. Desta forma, podemos afirmar que os modelos LPC e NRD são possíveis de serem utilizados para efeitos de síntese, não comprometendo a qualidade auditiva do sinal processado.

Na figura 5.5 estão representados os espectrogramas dos sinais: i) original produzido pelo orador do género feminino correspondente à vogal /i/; ii) sintético de acordo com o cenário de teste A; iii) sintético de acordo com o cenário de teste B; iv) sintético de acordo com o cenário de teste C; v) sintético de acordo com o cenário de teste D. Graficamente, são notórias as semelhanças entre os sinais obtidos após os diferentes cenários de síntese e o sinal original.

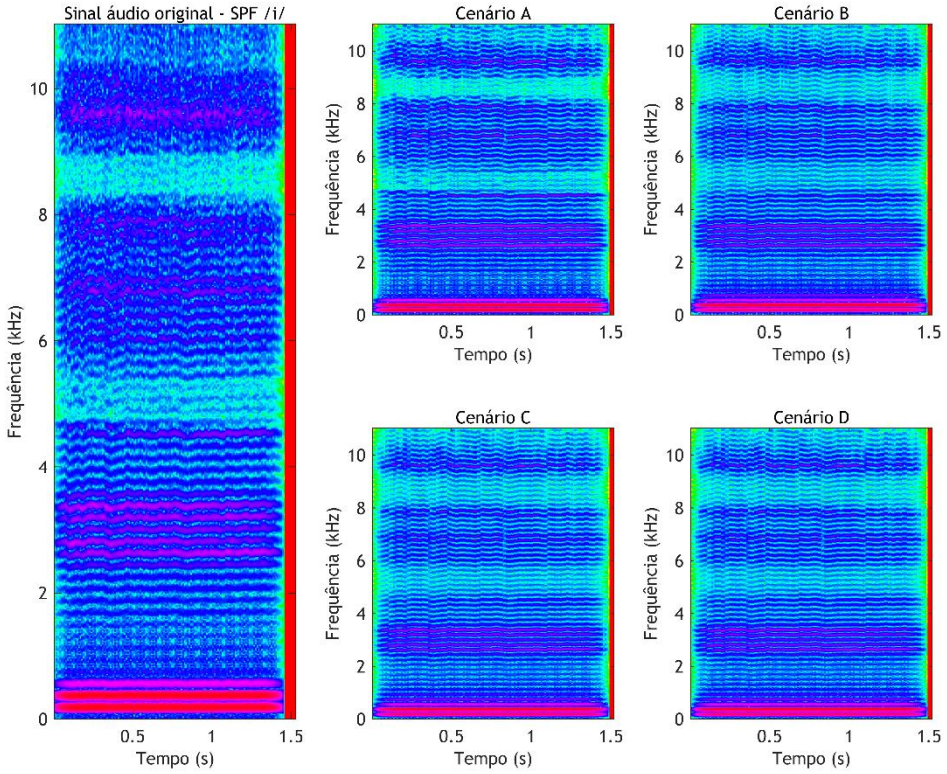

Figura 5.5 - Espectrogramas do sinal original produzido pelo orador do género feminino correspondente à vogal /i/ e dos sinais sintéticos após os diferentes cenários de teste.

No artigo [55] são apresentados e descritos testes de síntese harmónica e avaliação objetiva e subjetiva, de resultados idênticos aos realizados nesta secção, mas utilizando sinais sonoros de voz diferentes. O estudo de avaliação subjetiva foi realizado com a contribuição de diversos participantes, pelo que as conclusões retiradas anteriormente podem ser corroboradas e melhor sustentadas ao ter como referência este artigo.

5.5. Síntese

Neste capítulo foram descritos os procedimentos de modelização de magnitude e de fase paramétricos de sinais de fala vozeada e de síntese harmónica no domínio das frequências. Após a realização de testes para diferentes cenários de síntese harmónica e avaliação dos resultados, objetiva e subjetivamente, concluiu-se que a utilização dos modelos paramétricos de magnitude e fase, modelos LPC e NRD, respetivamente, não compromete a qualidade auditiva do sinal processado.

Sendo o principal objetivo desta investigação o implante das microvariações da frequência fundamental de um sinal de voz falada de um orador num outro, torna-se necessário a utilização de modelos paramétricos que preservem as características sonoras do sinal original. No capítulo seguinte recorre-se aos modelos LPC e NRD para efetuar transformação intencional das microvariações da frequência fundamental de sinais de voz falada.

Capítulo 6

Transformação intencional das microvariações da frequência fundamental de sinais de voz falada

O presente capítulo é dedicado à transformação intencional das microvariações da frequência fundamental de sinais de voz falada, recorrendo aos modelos paramétricos LPC e NRD. Efetua-se ainda uma avaliação dos resultados por meio de testes de perceção auditiva.

6.1. Aplanamento das microvariações da frequência fundamental

A primeira transformação intencional das microvariações da frequência fundamental realizada é o aplanamento das mesmas, definindo-se todo o trajeto de F0 como sendo o seu valor médio.

Foram obtidos resultados para oito sinais de voz falada, com uma duração de 1.45 segundos, correspondentes a duas vogais sustentadas, /a/ da palavra "água" e /i/ da palavra "ilha", produzidas por dois oradores do género feminino (SPF1 e SPF2) e por dois oradores do género masculino (SPM1 e SPM2). Para facilitar a análise da figura e das tabelas apresentadas com os resultados obtidos são utilizadas abreviações para referir os sinais avaliados, como por exemplo, SPF1 /i/ significa que a avaliação diz respeito à vogal /i/ produzida pelo orador 1 do género feminino.

De modo a avaliar o impacto percetual das diferenças entre os sinais originais e os respetivos sinais com as microvariações da frequência fundamental aplanadas, foram realizados testes de avaliação subjetiva. A avaliação foi realizada de acordo com a norma ITU-R BS.1116 [56] para a classificação de pequenas diferenças percetuais, utilizando também uma escala de qualidade contínua de 100 pontos tal como especificada na norma ITU-R BS.1284 [57]. Os participantes foram aconselhados a utilizar *headphones* de qualidade, pois algumas diferenças são subtis e

podem não ser notórias através das colunas do computador; também foram instruídos a adotar a seguinte escala de classificação percentual para avaliação das diferenças audíveis:

- impercetíveis (80 % a 100 %);
- percetíveis, mas não descaracterizadoras (60 % a 80 %);
- ligeiramente descaracterizadoras (40 % a 60 %);
- descaracterizadoras (20 % a 40 %);
- drásticas, muito descaracterizadoras (0 % a 20 %).

Nos testes percetuais participaram 17 ouvintes, em que 10 são do género feminino e 7 do género masculino, sendo a idade média de 31 anos (mínimo de 16 anos e máximo de 58 anos).

Na figura 6.1 estão representados graficamente os resultados obtidos, correspondentes às diferenças audíveis entre os sons originais e as respetivas versões aplanadas.

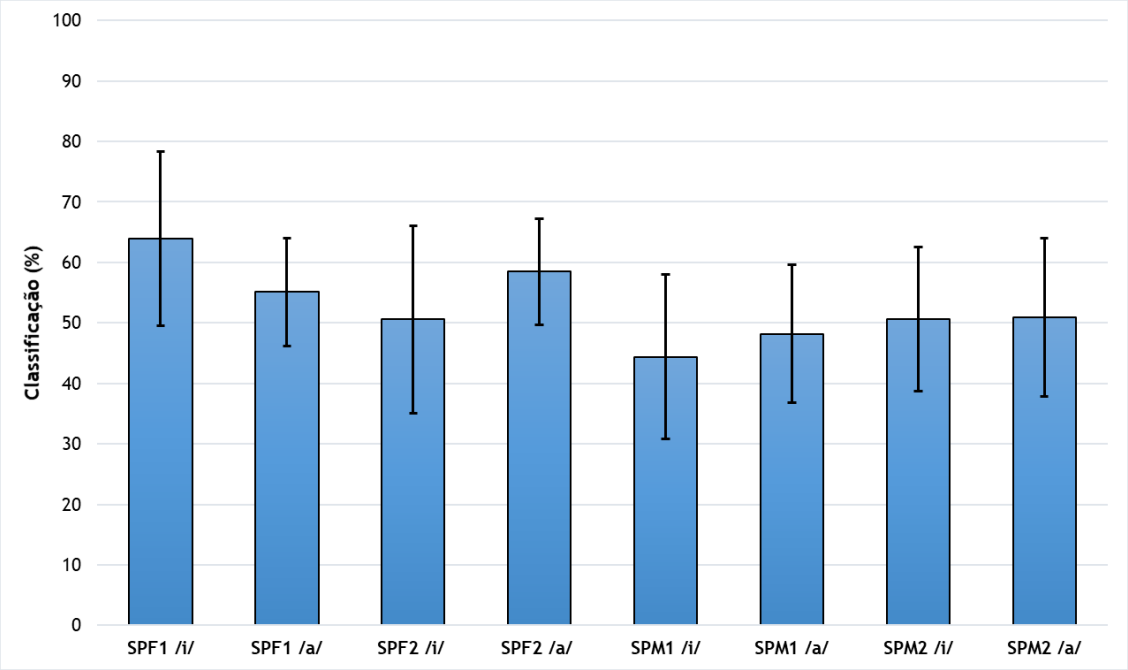

Figura 6.1 - Médias e intervalos de confiança (95 %) dos resultados de avaliação subjetiva.

Na tabela 6.1 estão apresentados os valores da classificação média e respetivos intervalos de confiança de 95 % para os sinais avaliados.

Na tabela 6.2 estão apresentados os valores da frequência fundamental média de cada orador e os valores da variação relativa do *pitch* de cada orador, isto é, o desvio padrão do *pitch* de cada orador a dividir pela sua frequência fundamental média.

Tabela 6.1 - Valores médios e intervalos de confiança (95 %) dos resultados de avaliação subjetiva.

Sinal avaliado	Classificação média	Intervalo de confiança (95 %) $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$
SPF1 /i/	63.94 %	63.94 % ± 14.40
SPF1 /a/	55.12 %	55.12 % ± 8.92
SPF2 /i/	50.59 %	50.59 % ± 15.47
SPF2 /a/	58.47 %	58.47 % ± 8.75
SPM1 /i/	44.41 %	44.41 % ± 13.57
SPM1 /a/	48.18 %	48.18 % ± 11.43
SPM2 /i/	50.59 %	50.59 % ± 11.92
SPM2 /a/	50.94 %	50.94 % ± 13.04

Tabela 6.2 - Valores da frequência fundamental média e da variação relativa do *pitch* de cada orador.

Sinal avaliado	F0 média	Variação relativa
SPF1 /i/	232.09 Hz	0.0075
SPF1 /a/	205.56 Hz	0.0076
SPF2 /i/	187.99 Hz	0.0061
SPF2 /a/	180.13 Hz	0.0078
SPM1 /i/	90.82 Hz	0.0105
SPM1 /a/	72.46 Hz	0.0049
SPM2 /i/	113.63 Hz	0.0062
SPM2 /a/	109.77 Hz	0.0080

Analisando a figura 6.1 e a tabela 6.1 verifica-se que a perceção auditiva de diferenças entre os sons originais e as respetivas versões aplanadas é induzida pelo perfil de voz de cada orador, independentemente da vogal produzida. Ou seja, não existe um padrão de classificação para o conjunto de vogais /i/ produzidas ou de vogais /a/ produzidas, pois os valores diferem consoante o orador. Verifica-se também que, apesar dos valores dos intervalos de confiança serem muito dilatados, em média, as classificações para o género feminino (média de 57.03 %) são ligeiramente superiores às do género masculino (média 48.53 %).

Os resultados de variação relativa do *pitch* de cada orador apresentados na tabela 6.2, apesar de não o comprovarem de forma inequívoca, sugerem que o ouvido humano é mais sensível às variações relativas; daí os resultados de avaliação subjetiva obtidos para o género feminino serem ligeiramente superiores aos resultados obtidos para o género masculino. Este pressuposto é evidenciado para o caso do orador SPM1 /i/, que tem a variação relativa mais elevada e o valor de classificação mais baixo de todos os oradores. O orador SPM2 /a/ tem também um valor de variação relativa mais elevado e um valor de classificação inferior aos dos oradores do género feminino.

Também se realizou o *t-test* estatístico para determinar se a diferença entre as médias de dois conjuntos de valores são estatisticamente significativas. O resultado devolvido designa-se de *p-value* e a sua comparação com o nível de significância definido permite aceitar ou rejeitar a hipótese nula. O nível de significância utilizado foi de 5 %, correspondente a um intervalo de confiança de 95 %, pelo que, se o *p-value* for menor do que 5 %, então a diferença entre as médias de dois conjuntos de valores é estatisticamente significativa e a hipótese nula é rejeitada. Caso contrário, a hipótese nula não é rejeitada.

No caso da avaliação percetual que foi realizada, a hipótese nula foi definida como: "os sinais de voz comparados não suscitam impacto percetual notório". Efetuou-se o *t-test* para cada orador e as diferentes vogais, ou seja, os dois conjuntos de valores utilizados em cada *t-test* foram os resultados obtidos para os sinais aplanados correspondentes às vogais /i/ e /a/ do mesmo orador. Isto porque, tal como já foi referido anteriormente, as diferenças percetuais entre os sons originais e as respetivas versões aplanadas são induzidas pelo perfil de voz de cada orador, independentemente da vogal produzida.

Na tabela 6.3 estão apresentados resultados obtidos do *p-value* após os *t-tests* realizados para cada orador, bem como quais os conjuntos de valores utilizados em cada *t-test*.

Tabela 6.3 - *p-values* para cada orador.

Conjunto de valores 1	Conjunto de valores 2	*p-value*
SPF1 /i/	SPF1 /a/	0.15
SPF2 /i/	SPF2 /a/	0.13
SPM1 /i/	SPM1 /a/	0.60
SPM2 /i/	SPM2 /a/	0.97

Analisando os resultados dos *t-tests*, verifica-se que todos os *p-values* obtidos são superiores a 5 %, ou seja, as diferenças entre as médias dos conjuntos de valores não são estatisticamente significativas, pelo que não se pode rejeitar a hipótese nula. Assim, o aplanamento dos sinais de voz para um mesmo orador não suscita diferenças percetivas notórias entre si. Porém, o próprio aplanamento da frequência fundamental exerce um impacto percetual claramente notório. Isto pode ser corroborado pelo facto de, em média, a classificação obtida para todos os sinais utilizados nos testes ser de 52.78 %, ou seja, as diferenças audíveis entre os sinais originais e os respetivos sinais aplanados são ligeiramente descaracterizadoras.

Conclui-se, assim, que as microvariações da frequência fundamental conferem naturalidade à voz e o seu aplanamento tem um impacto percetual notório.

6.2. Implantação dos padrões extraídos de um orador num outro

A segunda transformação intencional das microvariações da frequência fundamental realizada é a implantação das microvariações da frequência fundamental de um orador num outro. O procedimento realizado foi adicionar à frequência fundamental média de um orador as microvariações da frequência fundamental de um outro orador, forçando também a estrutura harmónica de fase através da utilização do modelo paramétrico NRD do outro orador.

Os sinais de voz falada utilizados foram os mesmos que na secção 6.1. A implantação dos padrões de um orador num outro foi realizada tendo em conta que os sinais têm de corresponder à mesma vogal e os oradores têm de ser do mesmo género. Por exemplo, para a vogal sustentada /i/ foram obtidos os resultados da implantação das microvariações e do modelo NRD do orador SPF2 no orador SPF1 e vice-versa. Para facilitar a análise das figuras e das tabelas apresentadas com os resultados obtidos, utilizaram-se as seguintes letras para representar as versões dos sinais modificados, tanto para oradores do género feminino como masculino:

- a letra A representa a implantação das microvariações da frequência fundamental do orador 2 no orador 1;
- a letra B representa a implantação das microvariações da frequência fundamental do orador 1 no orador 2.

A avaliação do impacto percetual após a implantação das microvariações da frequência fundamental de um orador num outro foi realizada através da classificação percentual de quanto as características dos sons originais dos oradores são percetíveis nas versões modificadas (A e B). À semelhança dos testes dedicados ao aplanamento realizados na secção 6.1, os participantes foram aconselhados a utilizar *headphones* de qualidade, pois algumas diferenças são subtis e podem não ser notórias através das colunas do computador; também foram instruídos a adotar a seguinte escala de classificação contínua de 100 pontos para avaliação da audibilidade das características do som de referência nas versões modificadas:

- totalmente percetíveis, isto é, os sons são iguais (80 % a 100 %);
- muito percetíveis (60 % a 80 %);
- percetíveis (40 % a 60 %);
- pouco percetíveis (20 % a 40 %);
- nada percetíveis, isto é, os sons são totalmente diferentes (0 % a 20 %).

Os participantes foram os mesmos que realizaram os testes na secção 6.1.

Nas figuras 6.2 e 6.3 estão representados graficamente os resultados obtidos da avaliação subjetiva do impacto percetual após a implantação das microvariações da frequência fundamental de um orador num outro para os oradores do género feminino e masculino, respetivamente. As abreviações abaixo das respetivas versões A e B, como por exemplo, SPF1 /i/, dizem respeito ao som original, em relação ao qual são feitas as avaliações comparativas.

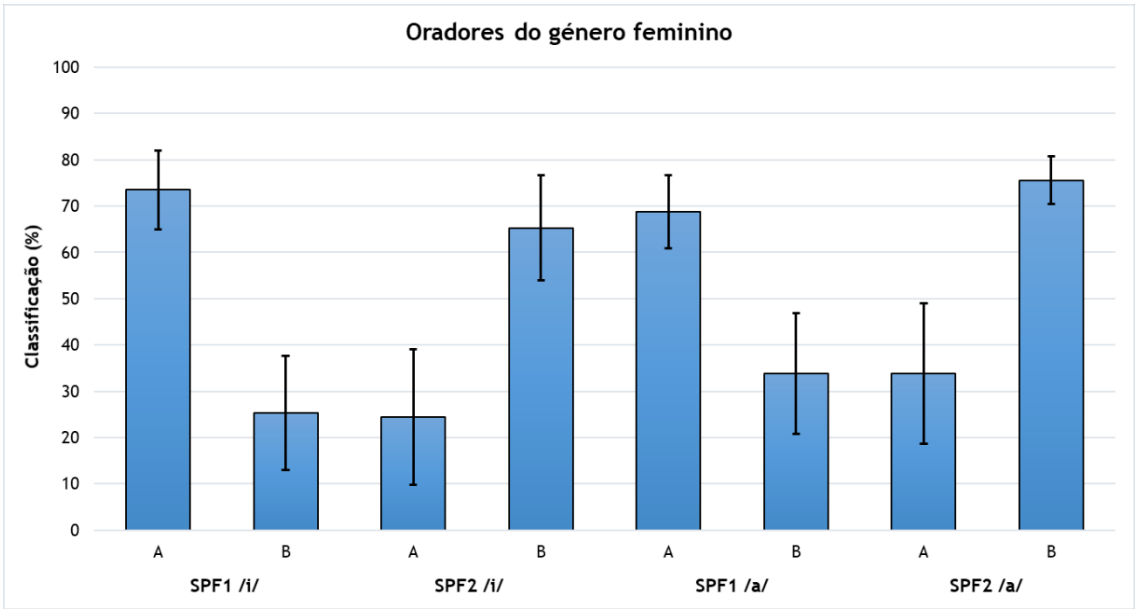

Figura 6.2 - Médias e intervalos de confiança (95 %) dos resultados de avaliação subjetiva para os oradores do género feminino.

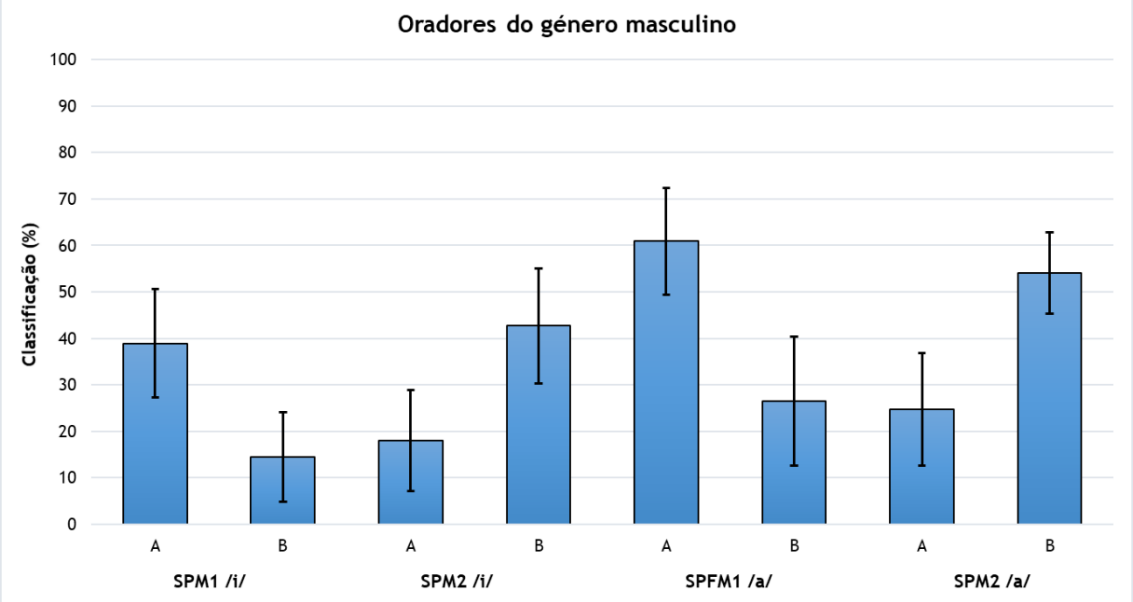

Figura 6.3 - Médias e intervalos de confiança (95 %) dos resultados de avaliação subjetiva para os oradores do género masculino.

Nas tabelas 6.4 e 6.5 estão apresentados os valores da classificação média e respetivos intervalos de confiança de 95 % para os sinais avaliados referentes aos oradores do género feminino e masculino, respetivamente.

Tabela 6.4 - Valores médios e intervalos de confiança (95 %) dos resultados de avaliação subjetiva para os oradores do género feminino.

Sinal de referência	Versão do sinal modificado	Classificação média	Intervalo de confiança (95 %) $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$
SPF1 /i/	A	73.53 %	73.53 % ± 8.51
	B	25.29 %	25.29 % ± 12.38
SPF2 /i/	A	24.41 %	24.41 % ± 14.60
	B	65.29 %	65.29 % ± 11.37
SPF1 /a/	A	68.76 %	68.76 % ± 7.93
	B	33.82 %	33.82 % ± 13.00
SPF2 /a/	A	33.82 %	33.82 % ± 15.09
	B	75.59 %	75.59 % ± 5.22

Tabela 6.5 - Valores médios e intervalos de confiança (95 %) dos resultados de avaliação subjetiva para os oradores do género masculino.

Sinal de referência	Versão do sinal modificado	Classificação média	Intervalo de confiança (95 %) $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$
SPM1 /i/	A	38.94 %	38.94 % ± 11.70
	B	14.41 %	14.41 % ± 9.66
SPM2 /i/	A	18.00 %	18.00 % ± 10.91
	B	42.70 %	42.70 % ± 12.36
SPM1 /a/	A	60.94 %	60.94 % ± 11.51
	B	26.47 %	26.47 % ± 13.90
SPM2 /a/	A	24.70 %	24.70 % ± 12.04
	B	54.12 %	54.12 % ± 8.75

De um modo geral, a partir das figuras 6.2 e 6.3 verifica-se que existe um padrão visual no que diz respeito às classificações médias obtidas. Quando a comparação é realizada entre o sinal de referência e a versão em que as microvariações são as do orador de referência implantados sobre a frequência fundamental média do outro orador, as classificações são inferiores. Quando o sinal modificado é a própria referência com implantação das microvariações da frequência fundamental de outro orador, a comparação deste com o sinal de referência resulta em valores de classificação superiores.

Analisando as tabelas 6.4 e 6.5, verifica-se também que o ouvido humano é mais sensível à implantação de padrões cruzados nos oradores do género masculino. Isto está relacionado com o facto da frequência fundamental média destes ser mais baixa do que a dos oradores do género

feminino (ver tabela 6.2 da secção 6.1), pelo que o período do ciclo glótico é mais prolongado, fazendo com que o ouvido tenha tempo para dedicar a sua atenção à forma de onda.

Também nesta secção, tal como na secção 6.1, foi realizado o *t-test* estatístico para testar a hipótese nula, desta vez definida como: "as microvariações da frequência fundamental de um dado orador são percetíveis quando são implantadas num outro orador". O nível de significância utilizado foi de 5 %, correspondente a um intervalo de confiança de 95 %. Efetuou-se o *t-test* para uma dada vogal correspondente a um dado orador, ou seja, os dois conjuntos de valores utilizados em cada *t-test* foram os resultados obtidos para as versões de sinais modificados A e B correspondentes a uma dada vogal de um dado orador.

Na tabela 6.6 estão apresentados resultados obtidos do *p-value* após os *t-tests* realizados entre as versões modificadas A e B de cada orador, correspondentes a uma dada vogal.

Tabela 6.6 – *p-values* para as versões modificadas A e B de cada orador, correspondentes a uma dada vogal.

Orador de referência	*p-value*
SPF1 /i/	1.42E-07
SPF2 /i/	3.00E-05
SPF1 /a/	9.47E-06
SPF2 /a/	8.51E-06
SPM1 /i/	2.13E-05
SPM2 /i/	1.73E-04
SPM1 /a/	1.03E-05
SPM2 /a/	4.00E-05

Analisando os resultados dos *t-tests*, verifica-se que todos os *p-values* obtidos são inferiores a 5 %, ou seja, as diferenças entre os conjuntos de valores (A e B) são estatisticamente significativas, pelo que se pode rejeitar a hipótese nula. Assim, as microvariações da frequência fundamental de um dado orador não são percetíveis quando são implantadas num outro orador. Isto corrobora as conclusões retiradas a partir das figuras 6.2 e 6.3 e das tabelas 6.4 e 6.5.

Assim, conclui-se que a frequência fundamental média é uma característica predominante na captação da atenção do ouvido humano, fazendo com que este não valorize de forma significativa outras características do sinal sonoro, tal como a forma de onda, a qual depende da estrutura harmónica de fase.

6.3. Síntese

Neste capítulo foram realizadas duas transformações intencionais das microvariações da frequência fundamental em sinais de voz falada, recorrendo aos modelos paramétricos LPC e NRD: o aplanamento das microvariações da frequência fundamental e a implantação das microvariações da frequência fundamental de um orador num outro.

Após a avaliação dos resultados obtidos por meio de testes de perceção auditiva, concluiu-se que as microvariações da frequência fundamental conferem naturalidade à voz. No entanto, estas microvariações não são capazes de modificar a assinatura sonora de um dado orador quando implantadas num outro orador. Assim, reforça-se a ideia de que a frequência fundamental média é um atributo do sinal em destaque na caracterização do orador.

Capítulo 7

Conclusões e trabalho futuro

Neste capítulo apresentam-se as conclusões retiradas ao longo do trabalho desenvolvido e após os diferentes resultados obtidos. Fazem-se ainda sugestões para trabalho futuro.

7.1. Conclusões do trabalho

No final deste trabalho conclui-se que os objetivos propostos foram cumpridos. Destacam-se as seguintes conclusões retiradas dos resultados de investigação e desenvolvimento realizados:

- o estimador proposto de estimação precisa da frequência de componentes sinusoidais individuais tem o melhor desempenho entre todos os estimadores que utilizam a janela sinusoidal, a qual é uma escolha conveniente para aplicações áudio baseadas em MDCT;
- o algoritmo SearchTonal tem um bom desempenho na estimação da frequência fundamental de estruturas harmónicas, estando já preparado para a realização da modelização e síntese harmónica;
- os modelos paramétricos de magnitude e fase, modelos LPC e NRD, respetivamente, podem ser utilizados na síntese harmónica de sinais no domínio das frequências, pois não comprometem a qualidade auditiva dos sinais processados;
- a transformação intencional das microvariações da frequência fundamental de sinais de voz falada por aplanamento destas retira naturalidade à voz;
- a frequência fundamental média da voz é uma característica percetual predominante na captação da atenção do ouvido humano, pelo que as microvariações da frequência fundamental e a estrutura harmónica de fase não são capazes de modificar a assinatura sonora de um dado orador quando implantadas sobre o valor médio da frequência fundamental de um outro orador.

7.2. Trabalho futuro

A partir dos resultados obtidos e das conclusões retiradas anteriormente, apresentam-se sugestões para trabalhos futuros, tendo em conta o projeto FCT no qual se enquadra esta dissertação:

- **efetuar melhorias no algoritmo SearchTonal**. Após os testes realizados no capítulo 4 conclui-se que ainda se pode melhorar o desempenho deste algoritmo quando existe presença de ruído. Sabe-se que este problema surge do facto do algoritmo SearchTonal esperar um número mínimo de parciais detetados na estrutura harmónica, pelo que a presença de ruído oprime os harmónicos, induzindo o algoritmo em erro;
- **realizar um estudo que abranja todas as vogais produzidas por todos os oradores da base de dados da qual o projeto FCT dispõe.** De modo a se poder realizar uma avaliação mais alargada do impacto que tem a transformação intencional das microvariações da frequência fundamental de sinais de voz falada, será vantajoso realizar diferentes testes para todos os sinais sonoros pertencentes à base de dados, permitindo desta forma uma análise estatística mais formulada e com resultados mais realistas;
- **realizar os mesmos testes para palavras em contexto real.** Os testes realizados nesta dissertação representam o início do desenvolvimento de um projeto FCT em que o contexto de análise foram vogais sustentadas isoladas, o que representa um cenário ideal na análise das transformações de sinal descritas e analisadas anteriormente. O passo seguinte será a utilização de vogais em contexto de fala natural, isto é, vogais retiradas de palavras, e desta forma analisar o desempenho dos métodos de estimação da frequência fundamental de estruturas harmónicas e o impacto da realização de transformações intencionais das microvariações da frequência fundamental aqui apresentados.

Referências

[1] Aníbal Ferreira. Análise acústica, percetiva e visual da voz. Em Ana P. Mendes, editor, *Vocologia do Fado*, capítulo 4. 2016.

[2] Aníbal Ferreira et al. Mecanismos humanos de produção e perceção audiovisual. Em Fernando Pereira, editor, *Comunicações Audiovisuais: Tecnologias, Normas e Aplicações*, capítulo 3. 2009.

[3] T. Drugman, P. Alku, A. Alwan e B. Yegnanarayana. Glottal source processing: From analysis to applications. *Computer Speech and Language*, 28(5): 1117-1138, 2014.

[4] Aníbal Ferreira e Carlos Salema. Som, luz e cor. Em Fernando Pereira, editor, *Comunicações Audiovisuais: Tecnologias, Normas e Aplicações*, capítulo 2. 2009.

[5] M. Wang e M. Lin. An Analysis of Pitch in Chinese Spontaneous Speech. *International Symposium in Tonal Aspects of Languages*, 2004.

[6] I. R. Murray e J. L. Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108, 1993.

[7] B. C. J. Moore. An Introduction to the Psychology of Hearing. Academic Press, 4 edição, 1997.

[8] A. M. Kondoz. Digital Speech: Coding for Low Bit Rate Communication Systems. John Wiley & Sons, Inc., 1994.

[9] T. Painter e A. Spanias. Perceptual coding of digital audio. *Proceedings of the IEEE*, 88(4): 451–512. 2000.

[10] D. Hoppe, M. Sadakata e P. Desain. Development of real-time visual feedback assistance in singing training: a review. *Journal of Computer Assisted Learning*, 22(4): 308–316, 2006.

[11] A. Klapuri, M. Davy e T. Virtanen. Part III: multiple fundamental frequency analysis. Em A. Klapuri, M. Davy, editores, *Signal Processing Methods for Music Transcription*. Springer, 2006.

[12] E. Jacobsen e P. Kootsookos. Fast, Accurate Frequency Estimators. *IEEE Signal Processing Magazine*, 24(3): 123-125, 2007.

[13] J. Sundberg. The Science of the Singing Voice. *Northern Illinois University Press*. 1987.

[14] A. Ferreira e R. Sousa. DFT-based frequency estimation under harmonic interference. *Final Program and Abstract Book - 4th International Symposium on Communications, Control, and Signal Processing, ISCCSP 2010*. 2010.

[15] A. V. Oppenheim, R. W. Schafer e J. R. Buck. Discrete-time Signal Processing. Prentice-Hall Inc., 2 edição, 1998.

[16] R. Sousa e A. Ferreira. Non-iterative frequency estimation in the DFT magnitude domain. *Final Program and Abstract Book - 4th International Symposium on Communications, Control, and Signal Processing, ISCCSP 2010*, 2010.

[17] F. J. Harris. On the use of windows for harmonic analysis with the Discrete Fourier Transform. *Proceedings of the IEEE*, 66(1): 51–83, 1978.

[18] M. D. Macleod. Fast nearly ml estimation of the parameters of real or complex single tones or resolved multiple tones. *IEEE Transactions on Signal Processing*, 46(1): 141–148, 1998.

[19] J. Schoukens, R. Pintelon e H. Van Hamme. The interpolated fast Fourier transform: a comparative study. *IEEE Transactions on Instrumentation and Measurement*, 41(2): 226–232, 1992.

[20] B. G. Quinn. Estimation of frequency, amplitude, and phase from the DFT of a time series. *IEEE Transactions on Signal Processing*, 45(3): 814–817, 1997.

[21] J.D. Klein. Fast algorithms for single frequency estimation. *IEEE Transactions on Signal Processing*, 54(5): 1762–1770, 2006.

[22] T. Grandke. Interpolation algorithms for discrete Fourier transforms of weighted signals. *IEEE Transactions on Instrumentation and Measurement*, 32(2): 350–355, 1983.

[23] B. G. Quinn. Frequency estimation using tapered data. Em *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, páginas III73-III76, 2006.

[24] Y. Dun e G. Liu. A Fine-Resolution Frequency Estimator in the Odd-DFT Domain. *IEEE Signal Processing Letters*, 22(12): 2489-2493, 2015.

[25] A. Ferreira e D. Sinha. Accurate and robust frequency estimation in the odft domain. Em *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, páginas 203-206, 2005.

[26] M. Lagrange e S. Marchand. Estimating the instantaneous frequency of sinusoidal components using phase-based methods. *Journal of the Audio Engineering Society (JAES)*, 55:385-, 2007.

[27] A. Camacho. *Swipe: a swatooth waveform inspired pitch estimator for speech and music*. Tese de doutoramento, Universidade da Florida, 2007.

[28] D. Jouvet e Y. Laprie. Performance analysis of several pitch detection algorithms on simulated and real noisy speech data. Em *2017 25th European Signal Processing Conference (EUSIPCO)*, páginas 1614-1618, 2017.

[29] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of sampled sound. Em *IFA Proceedings 17*, páginas 97-110, 1993.

[30] A. de Cheveigné e H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111: 1917-30, 2002.

[31] D. Talkin. A Robust Algorithm for Pitch Tracking (RAPT). Em W. B. Kleijn, K. K. Paliwal, editores, *Speech Coding and Synthesis*. Elsevier, páginas 495-518, 1995.

[32] H. Kawahara e H. Katayose, A. de Cheveigné e R. D. Patterson. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity. Em *Porc EUROSPEECH*, volume 6, 1999.

[33] Praat. Disponível em http://www.fon.hum.uva.nl/praat/. Acesso em 24 de janeiro de 2019.

[34] T. Drugman e A. Alwan. Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics. Em *INTERSPEECH*, 2011.

[35] A. M. Noll. Cepstrum Pitch Determination. *The Journal of the Acoustical Society of America*, 41: 293-309, 1967.

[36] G. Ravindran, S. Shenbagadevi e V. Salai Selvam. Cepstral and linear prediction techniques for improving intelligibility and audibility of impaired speech. *Journal Biomedical Science and Engineering*, 3: 85-94, 2010.

[37] J. Ventura. *Biofeedback da Voz Cantada*. Tese de mestrado, Faculdade de Engenharia da Universidade do Porto, 2011.

[38] J. Ventura, R. Sousa e A. Ferreira. Accurate analysis and visual feedback of vibrato in singing. *5th International Symposium on Communications Control and Signal Processing*, *ISCCSP 2012*, 2012.

[39] R. Costa. *Reorganização espectral de sinais e fala na banda de [0, 2] kHz*. Tese de doutoramento, Faculdade de Engenharia da Universidade do Porto, 2009.

[40] V. Almeida. *Relação entre características objetivas da voz cantada e seus atributos artísticos e estéticos*. Tese de mestrado, Faculdade de Engenharia da Universidade do Porto, 2012.

[41] A. Ferreira. Accurate estimation in the ODFT domain of the frequency, phase and magnitude of stationary sinusoids. *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

[42] P. P. Vaidyanathan. Multirate Systems and Filter Banks. Prentice-Hall, 1993.

[43] H. Malvar. Signal Processing with Lapped Transforms. Artech House, Inc., 1992.

[44] S. A. Zahorian e A. J. Jagharghi Spectral-shape features versus formants as acoustic correlates for vowels. *Journal of the Acoustical Society of America*, 94(4): 1966–1982, 1993.

[45] L. Rabiner e B.-H. Juang. Fundamentals of Speech Recognition. *Prentice-Hall, Inc.*, 1993.

[46] T. F. Quatieri e R. J. McAulay. Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, 40(3): 497–510, 1992.

[47] M. R. Portnoff Time-scale modification of speech based on short time Fourier analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(3): 374–390, 1981.

[48] A. J. Ferreira e J. M. Tribolet. A holistic glottal phase related feature. Em *21st International Conference on Digital Audio Effects (DAFx-18), Aveiro, Portugal*, 2018.

[49] M. H. Hayes. Statistical Digital Signal Processing and Modeling. John Wiley & Sons Inc., 1996.

[50] R. Sousa e A. Ferreira. Importance of the relative delay of glotal source harmonics. Em *AES 39th International Conference, Denmark*, 2010.

[51] I. Stylianou. *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification*. Tese de doutoramento, École Nationale Supérieure des Télécommunications, France, 1996.

[52] R. D. Federico. Waveform preserving time stretching and pitch shifting for sinusoidal models of sound. Em *COST-G6 Digital Audio Effects Workshop*, páginas 44-48, 1998.

[53] I. Saratxaga, I. Hernaez, D. Erro, E. Navas e J. Sanchez. Simple representation of signal phase for harmonic speech models. *Electronic Letters*, 45 (381), 2009.

[54] A. Ferreira e D. Sinha. Advances to a Frequency-Domain Parametric Coder of Wideband Speech. Em *AES 140th Convention, Paris, France*, 2016.

[55] A. Ferreira, J. Silva, F. Brito e D. Sinha. Subjective impact of holistic phase and magnitude descriptors in fully parametric harmonic speech representation and synthesis. Em *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2019.

[56] ITU-R Recommendation BS.1116-3. *Methods for the subjective assessment of small impairments in audio systems*. 2015.

[57] ITU-R Recommendation BS.1284-2. *General methods for the subjective assessment of sound quality*. 2019.