FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Vozeamento sintético de voz disfónica através da síntese digital de estruturas harmónicas em tempo real

Nélio David de Freitas Gonçalves

Mestrado em Engenharia Eletrotécnica e de Computadores

Orientador: Prof. Dr. Aníbal João de Sousa Ferreira

28 de julho de 2023

Resumo

A fala sussurrada é definida, ao nível fisiológico, pela falta de atuação das pregas vocais durante a enunciação de fonemas que de outro modo fariam uso delas. Apesar da utilidade pontual que a discrição do sussurro oferece, o impacto negativo que este pode ter na qualidade de vida de pacientes que sofrem de algum tipo de afonia, e que se vêm por isso limitados exclusivamente a este modo de comunicação, é significativo. As vias mais comuns de reabilitação incluem, por exemplo, a electrolaringe, a voz esofagal e sistemas informáticos como o *text-to-speech*, que, apesar de poderem restituir alguma medida de voz ao indivíduo afetado, ainda opõem grandes obstáculos à recuperação de uma fala natural.

O DyNaVoiceR é um projeto, financiado pela Fundação para a Ciência e Tecnologia (FCT), que propõe o desenvolvimento de um assistente de fala não-intrusivo capaz de identificar e converter fala sussurrada em fala normal em tempo-real. Os trabalhos que promove têm contribuído para a construção de um ambiente de processamento próprio que reúne, e aplica, métodos de segmentação e classificação de sinais de fala, de extração de características espectrais, de vozeamento sintético, entre outros. Esta dissertação tem como objetivo testar e validar o uso do kit TM32F746G Discovery como plataforma de hardware para a implementação simplificada de uma componente do assistente como prova de conceito.

Desenvolveram-se algoritmos de vozeamento baseados na síntese de uma forma de onda periódica arbitrária no domínio das frequências, utilizada como substituto da componente harmónica em falta na voz disfónica, incorporada no sinal de voz pela modulação com a envolvente espectral do sinal de sussurro. Com vista a validar os métodos propostos, conduziram-se testes subjetivos onde a inteligibilidade e naturalidade de excertos de áudio vozeados artificialmente segundo o algoritmo foi avaliada. Esta investigação levou a uma importante conclusão sobre como o carácter ruidoso do sinal de sussurro impacta a percepção do áudio, tendo mostrado que a falta de congruência interfere com a reconstrução de sinal devido à falta de congruência na sobreposição de segmentos consecutivos de sinal. Os resultados dos testes apoiam a solução proposta como via para um vozeamento de teor mais natural.

Abstract

Whispered speech is defined, at a physiological level, by the lack of vibration on the part of the vocal folds on otherwise voiced phonemes. In spite of its usefulness when it comes to providing one with the means with which to communicate discreetly, the negative impact it has on the social and professional lives of patients suffering from aphonia, who may be restricted to this mode of speech, is very significant. Common rehabilitation procedures include, but are not restricted to the electrolarynx, esophageal speech and assistive technology such as text-to-speech, which, even though they succeed at giving back a sense of speech to the afflicted individual, they are still far from constituting practical and, most importantly, natural-feeling and sounding solutions.

DyNaVoiceR is an FCT funded project aimed at developing a non-intrusive speech assistant capable of identifying and converting whispered speech into normal speech in real-time. The research projects it has prompted have resulted in several contributions to the development of a dedicated signal processing environment that integrates speech segmentation algorithms, classifiers, methods for signal feature extraction, among others. This dissertation is focused on testing the use of the TM32F746G Discovery kit as host for the aforementioned environment as proof-of-concept for a striped-down and computationally efficient voice assistant product.

An algorithm was then developed wherein an arbitrary periodic waveform, synthesized in the frequency domain and modulated by the spectral envelope of the whispered speech signal, was used as substitute for the missing periodic component of the dysphonic voice. In order to validate this approach, a set of subjective tests was put forward to evaluate the gains in inteligibility and naturalness of the artificially voiced speech excerpts. This investigation was helpful in showing how the noise character of whispered speech signal interferes with signal reconstruction, revealing how the lack of a stationary source model can render overlapping speech segments incongruous, thereby degrading the synthesized audio. The test results support the claim that the proposed method for overcoming said problems is effective in producing a more natural-sounding artifical voice.

"He said, «I dream of colour music
And the intricacies of the machines that make it possible»
I said, «You are nothing if not inconsistent»"

Jhonn Balance

Conteúdo

Lista de Figuras

Abreviaturas e Símbolos

PTE	Punção Traqueoesofágica
EL	Electrolaringe
FCT	Fundação para a Ciência e Tecnologia
DyNaVoiceR	Dysphonic to Natural Voice Reconstruction
DMA	Direct-Memory Access
V/UV	Voiced/Un-Voiced
LPC	Linear Predictive Coding
SNR	Signal-to-Noise Ratio
MELP	Mixed Excitation Linear Prediction
CELP	Code-Excited Linear Prediction
DSP	Digital Signal Processing
ODFT	Odd Discrete Fourier Transform
OLA	Overlap-Add
PSD	Power Spectral Density

Capítulo 1

Introdução

Neste capítulo introdutório são apresentados o problema e motivação do presente estudo, o enquadramento da dissertação no contexto da área científica em que está inscrita, e a estrutura do presente documento sob forma de tópicos.

1.1 Enquadramento e Motivação

A fala constitui uma das mais importantes, e poderosas, ferramentas de comunicação ao dispor do ser humano. Não só se vê como veículo primeiro da linguagem, estando na raiz da dimensão discursiva da transmissão do pensamento, mediado segundo a língua, na qualidade de troca simbólica, como também é capaz de a transcender pelo controlo da fonética. A entoação, volume, dicção, a assinatura vocal e a prosódia complementam a oralidade pela capacidade de moldar o sentido de uma mensagem falada. O modo como um aumento do *pitch* durante o término de uma frase torna distinguível uma afirmação de uma interrogação, ou de como o tom de voz pode ser indicativo do estado emocional do orador, realçam a importância destas qualidades não apenas na composição de uma fala natural e perceptível, mas também na forma como se implicam sobre a capacidade individual para a sociabilidade como um todo.

O sussurro é caracterizado pela falta da atuação das pregas vocais na produção de sons de outra forma vozeados. Como tal, fonemas proferidos neste registo são desprovidos de uma componente periódica forte, têm fraca projeção, não comportam a assinatura vocal do locutor e são mais suscetíveis a interferência por ruído. Sussurrar intencionalmente pode ser uma forma eficaz de enunciar algo de forma discreta num cenário que assim o exige, como, por exemplo, numa sala de estudo comunitária. No entanto, o que caracteriza este tipo de fala aplica-se por inteiro a pacientes que sofrem de algum tipo de afonia, o que fundamenta parte do problema a abordar.

A perda de voz, ou afonia, define uma condição patológica na qual um indivíduo é impedido de produzir sons vozeados de forma natural, devido, por exemplo, a perturbações a nível da musculatura do trato vocal, como no caso da *Muscle Tension Dysphonia*, para a qual foi explorada uma solução baseada em terapia vocal funcional em [6], ou como consequência de danos provocados ao nível da laringe.

A reabilitação vocal em situações de remoção parcial ou total da laringe no seguimento de um processo de laringectomia, surge maioritariamente sobre a forma da punção traqueoesofágica (PTE), voz esofágica ou pela utilização da electrolaringe (EL) [7, 8]. Para além de, no caso do PTE, ser exigida uma intervenção cirúrgica complexa e invasiva, estas soluções apresentam sérias limitações quanto à produção de uma fala de cariz natural, facilmente inteligível e comfortável para o paciente. Com efeito, a PTE e a EL são métodos *hands-on*, *i.e* requerem uma contração manual da zona da punção para que se produza o som da voz. Verificou-se, quanto a este último ponto, e é notado em [8], que existe uma quantidade não negligenciável de pacientes de PTE que optam por remover a prótese, enquanto o uso dos métodos eletrónicos, como a EL, tendem a perdurar. Alternativas computacionais como o *text-to-speech* são de modo geral mais cómodas, mas impossibilitam a comunicação em tempo real, pois o utilizador é obrigado a digitar a mensagem que pretende converter.

O objetivo último do projeto da Fundação para a Ciência e Tecnologia (FCT) no qual se insere este trabalho, o DyNaVoiceR, é o de desenvolver um assistente de fala não intrusivo capaz de converter fala sussurrada em fala normal em tempo-real. O projeto divide-se em 5 tarefas principias. São estas, segundo [9]:

- A – Idiosyncratic voice signal analysis and modeling
- B – Perceptually natural synthesis of periodic voicing components
- C – DyNaVoiceR system integration and real-time implementation
- D – DyNaVoiceR usability tests and fine-tuning
- E – Management

Considerando o problema da restituição de voz nos casos apresentados, e os largos obstáculos que os métodos correntes ainda interpõem entre o indivíduo e o domínio confortável da sua fala, tornam-se evidentes os benefícios que uma solução deste tipo oferece.

1.2 Objetivos

Esta dissertação está associada ao segundo segmento da fase B do projeto Dysphonic to Natural Voice Reconstruction (DyNaVoiceR), intitulada "*Perceptually natural voicing implantation and prosodic control*", e visa a realização de uma prova de conceito simplificada do assistente através da sua implementação num *kit* STM32F746G Discovery, cujas capacidades para realização do processamento pretendido serão testadas. Isto constitui um importante passo na trajetória desta pesquisa, pois apesar de perfazerem 20 anos desde que foi colocada uma proposta deste tipo [10], ainda não existe uma concretização física do assistente pretendido.

Este trabalho tem por principais objetivos os elencados a seguir:

- A elaboração de algoritmos de vozeamento baseados em processos de análise e síntese assentes em transformada ODFT.

- Programar e testar o STM32F746G Discovery. Explorar os limites e oportunidades oferecidas pelo uso do *Direct-Memory Access* (DMA).
- Ensaiar, realizar a validação funcional do assistente e avaliar a qualidade objetiva e subjetiva do seu desempenho.

1.3 Estrutura da Dissertação

Além do capítulo introdutório, constam deste documento as seguintes secções e temas:

- Capítulo 2 - ***Background* teórico**, onde serão expostos alguns conceitos fundamentais relacionados com a produção de fala e as perspetivas computacionais adotadas no sentido de a modelar.
- Capítulo 3 - **Revisão do Estado da Arte**, que compõe uma imagem do cenário atual da pesquisa sobre o tópico de conversão de fala.
- Capítulo 4 - **Plataforma de Processamento**, onde se introduz o funcionamento do *kit* STM32F746G Discovery em regime de controlo de transferências de memória segundo DMA, e onde é medido o atraso de processamento mínimo introduzido pelo processamento em bloco.
- Capítulo 5 - **Estrutura de Análise e Síntese**, neste capítulo expõe-se o funcionamento da estrutura de análise síntese que é adotada, ligando a escolha de janela e transformada aos requisitos do sistema, e abordando também a implementação e validação da operação do algoritmo que o implementa no *kit*.
- Capítulo 6 - **Síntese e adição nas frequências de uma onda de amplitude constante**, onde um método para a adição de uma forma onda periódica arbitrária, de F0 constante, sintetizada nas frequências segundo transformada ODFT a uma onda analógica captada em tempo real é demonstrado e concretizado num algoritmo adaptado ao funcionamento no *kit*, seguido de testes de validação do mesmo.
- Capítulo 7 - **Modulação por Envolvente Espectral**, apresentam-se neste capítulo dois algoritmos onde uma forma de onda arbitrária é modulada pela envolvente espectral do sinal de sussurro captado em tempo-real. O final do capítulo é dedicado aos testes subjetivos que foram conduzidos de modo a avaliar a qualidade dos algoritmos de vozeamento desenvolvidos.
- Capítulo 8 - **Conclusões e Trabalho Futuro**, oferece um resumo das principais conclusões que são extraídas desta investigação e sugere tarefas que deem continuidade ao trabalho realizado.

Capítulo 2

Background

2.1 Introdução

De modo a introduzir o problema computacional colocado pela dificuldade em aferir uma componente periódica de um sinal gerado por uma voz afónica, segue-se neste capítulo uma revisão sobre os conceitos fundamentais associados à fala e sua produção.

2.2 Produção de Fala

A produção da fala pode ser vista, segundo um modelo discutido em [11], como o produto final de um sistema, a cadeia da fala, constituído por três níveis:

1. Nível linguístico: responsável por gerar o conteúdo simbólico da fala, e onde o principal elemento determinante é a gramática específica à língua do falante.

2. Nível fisiológico: nível que abarca a estrutura do aparelho fonador.

3. Nível acústico: associado às propriedades físicas do som produzido.

Apesar da importância que o primeiro destes níveis acarreta no cenário atual de reconhecimento e síntese de fala [12, 13], nesta secção, serão abordados apenas os últimos dois.

2.2.1 Aparelho Fonador

O aparelho fonador é formado por um conjunto diverso de órgãos não especializados, e pode ser decomposto, em três regiões identificadas segundo a sua posição relativa à glote, e que se distinguem pela função que interpretam na produção de fala. São estas: a **região sub-glótica**, composta por elementos do sistema respiratório, nomeadamente, os pulmões, brônquios, e traqueia, que controlam o fluxo de ar durante a fala e influenciam a dinâmica da mesma; a **laringe**, que alberga as pregas vocais e a glote, isto é, a abertura entre as pregas, e, por último, o tracto vocal, ou **região supraglótica**. A Figura 2.1 ilustra e identifica os órgãos estruturantes destas últimas duas regiões do aparelho vocal.

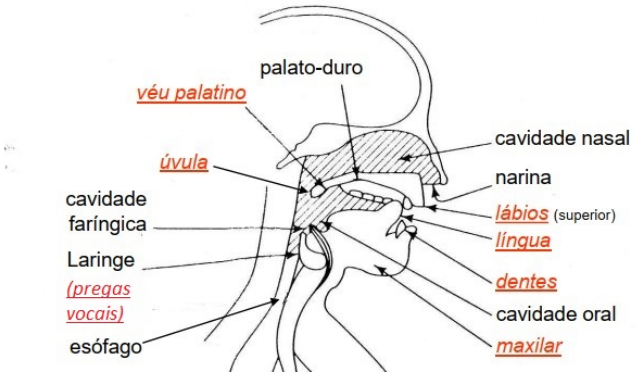


Figura 2.1: Secção sagital média do aparelho vocal. Figura adaptada de [1]

2.2.1.1 Vozeamento

A fala vozeada tem início no momento em que o ar impulsionado pelos pulmões, subindo pela traqueia, vence a resistência oferecida pelas **pregas vocais**. A partir deste ponto, a flexibilidade, a tensão da musculatura das pregas, a diminuição repentina da pressão sub-glótica e a força de Bernoulli [2, 1] levam a um fecho rápido destas, restringindo a passagem de ar até que a pressão atinja novamente o mesmo nível. Deste modo, o fluxo de ar pela laringe acontece como uma sequência de impulsos correspondentes às massas de ar expelidas aquando a separação das pregas.

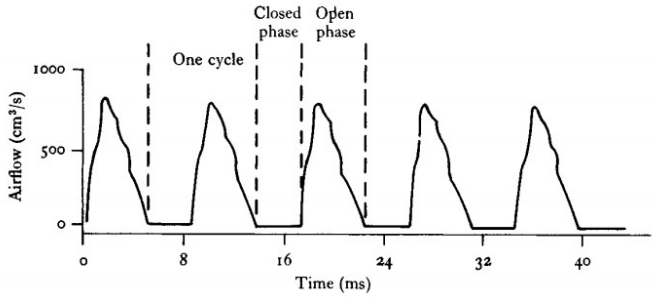


Figura 2.2: Forma de onda do fluxo de ar na laringe. Adaptado de [2]

A taxa a que este ciclo se repete oferece ao som produzido, neste caso, a voz, a sua frequência fundamental e, consequentemente, a sua estrutura harmónica. É de notar que a tensão das pregas pode ser controlada para que estas vibrem a diferentes frequências, e é esse controlo que permite gerir a entoação, que é uma das componentes da prosódia.

Ao nível das frequências, o espectro do pulso glotal, no intervalo que contempla as componentes de maior energia, tem o aspeto que pode ser verificado na Figura 2.3. Apesar de alguns harmónicos excecionalmente pouco pronunciados, o decaimento da energia das componentes em frequência apresenta alguma regularidade.

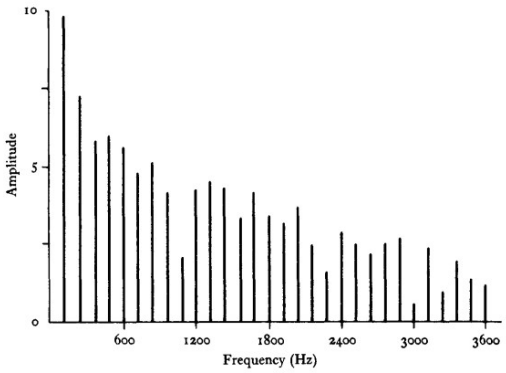


Figura 2.3: Espectro do impulso glotal. Adaptado de [2]

2.2.1.2 Articulação e Formantes

O sinal produzido vê-se modificado pela ação das diferentes ressonâncias que têm origem nas variações da geometria tubular do **tracto vocal**, e pelas obstruções que a posição dos principais articuladores, identificados na legenda da Figura 2.1 pela cor vermelha, introduzem no fluxo de ar. Diferentes geometrias do tubo implicam diferentes ressonâncias e, como tal, a produção de diferentes fones. As regiões da envolvente espectral acentuadas pelo comportamento do tracto são designadas na literatura por frequências formantes.

As três primeiras formantes são tidas normalmente como bons indicadores acústicos de vogais [14], sendo que as duas primeiras são as que mais pesam na caracterização de fonemas, concentrando também a maior parte da energia do sinal. Consoantes não-vozeadas não apresentam componente periódica forte, por falta de atuação das pregas, e são produzidas em grande parte por movimentos fricativos do ar junto aos articuladores mais externos, tais como os dentes, os lábios e a língua. A última etapa no trajeto do sinal periódico é a irradiação pelas cavidades nasal e oral. Na figura seguinte podem ser notadas as proeminências correspondentes às formantes.

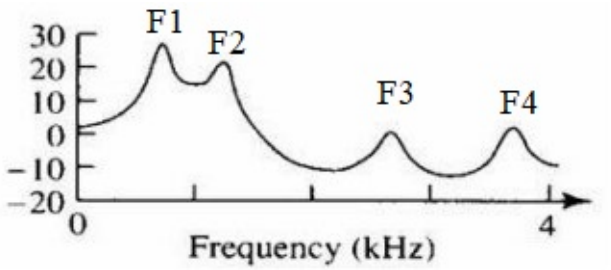


Figura 2.4: Envolvente do espectro do sinal acústico da vogal /a/, tal como pronunciada na palavra inglesa *father*. Adaptado de [1]

Ao moldar a forma de onda e o espectro do sinal vozeado, o tracto vocal, e a região supraglótica como um todo, atuam como um filtro acústico pelo qual são processados os sinais gerados nas camadas inferiores, sejam eles vozeados ou não-vozeados.

2.2.2 Modelo Fonte-Filtro

A conclusão anterior funda o princípio no qual se sustenta o conceito do modelo fonte-filtro, introduzido por Gunnar Fant [15]. Segundo este, o aparelho produtor de fala pode ser caracterizado por uma fonte de sinal de natureza periódica, no caso de sons vozeados, ou ruidosa, para não-vozeados, seguido de um banco de filtros que emula a resposta em frequência do tracto vocal, e, por conseguinte, introduz e localiza as formantes ao definir a envolvente do espectro.

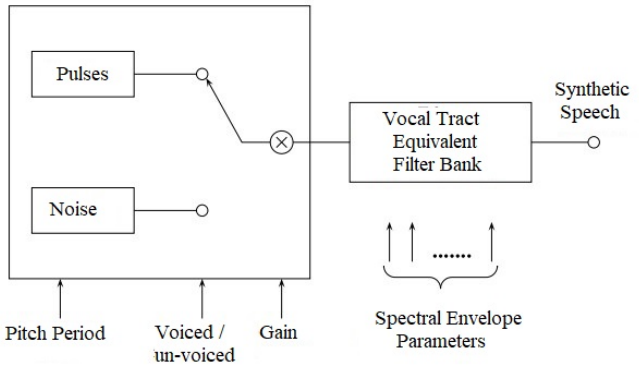


Figura 2.5: Modelo simplificado da produção de fala. Adaptado de [3]

Este modelo funda uma importante técnica de processamento e codificação de áudio, o *Linear Predictive Coding* (LPC). Neste, procura-se parametrizar o processo de fala segundo os parâmetros anotados na Figura 2.5. São estes: o modo de articulação, vozeado ou não vozeado (V/UV), implicando, respetivamente, uma excitação por um trem de impulsos periódicos ou por ruído; o período do *pitch*, no caso vozeado; o ganho e parâmetros do filtro estimador. A resposta do tracto é aproximada por um filtro *all-pole*, pelo que somente as ressonâncias ficam ao alcance direto da representação, em detrimento de anti-ressonâncias que ocorrem com maior preponderância nos sons nasalados não-vozeados.

Uma vez estimada a resposta do filtro, esta pode ser desconvoluida do sinal, processo designado na literatura como *inverse-filtering*, do qual a saída, o sinal branqueado, é conhecido como resíduo, ou *residue*. Os parâmetros do modelo de filtro são calculados com base na minimização do erro quadrático médio. O ganho computacional está implicado na possibilidade de reconstrução do sinal original a partir do resíduo e dos parâmetros do filtro, utilizando-os para construir o filtro LPC que devolve ao espectro a envolvente equivalente do tracto. Na perspetiva do descodificador, o *residue* identifica-se com a excitação original da fala.

É importante notar que na generalidade dos casos práticos de processamento computacional de voz, se considera que o sinal processado corresponde a um curto segmento de fala, obtido pela aplicação de uma análise deslizante com janela. Somente para intervalos breves é justificável supor que o sinal é aproximadamente estacionário, ou seja, que as suas características espectrais e probabilísticas não se alteram significativamente durante o período observado. A caracterização do sistema é tanto melhor quanto mais o fone possa ser sustido ao longo do tempo, o que ocorre

na produção de vogais, mas o que dificilmente se verifica no caso de sons oclusivos, que têm um carácter percussivo e, portanto, altamente não-estacionário.

2.3 Conclusões

Da exposição levada a cabo neste capítulo podemos concluir como é que os princípios da produção de voz influenciam a perspetiva computacional empregue para os abordar. Além do mais, percebem-se quais as origens dos desafios colocados no tratamento de fala afónica, onde a componente periódica está ausente.

No capítulo seguinte apresenta-se o estado da arte, incluindo o trabalho desenvolvido na FEUP no âmbito da pesquisa do DyNaVoiceR.

Capítulo 3

Revisão Bibliográfica - Conversão de sussurro

Neste capítulo reveem-se alguns dos mais significativos trabalhos realizados no âmbito da conversão de fala sussurrada.

3.1 MELP

Importa destacar inicialmente o trabalho de 2002 de Morris e Clements [10]. Os autores optam por utilizar um modelo modificado do codificador de *mixed excitation linear prediction* (MELP). Este modelo partilha muitas das características fundamentais do LPC, como uma descrição completamente paramétrica do sistema (conforme indicado na Figura 2.5), mas introduz uma maior *nuance* a nível da excitação ao adicionar um *jitter* ao *pitch* fundamental, o que o aproxima de um modelo mais natural da excitação vocal tal como ocorre no aparelho de fala humano [16]. A predição linear neste trabalho está assente num LPC de 10ª ordem.

Ao modelo MELP mais tradicional é acrescentado um estágio de pré-processamento de onde se pretendem extrair a estimação de *pitch* e os valores de deslocação de formantes. Estes, replicam *shifts* de frequência e modificações de banda conhecidos entre as contrapartes vozeadas e sussurradas de fonemas idênticos, um detalhe também contemplado em [17]. À cabeça desse estágio é colocado um filtro de fase mínima cuja resposta em magnitude é aferida através do logaritmo da diferença média verificada entre o espectro da enunciação de vogais normais e sussurradas. A saída do filtro é usada tanto como fonte para a estimação da F0 como para o *shift* de formantes.

Os resultados deste último ponto variam em qualidade conforme o filtro de *smoothing* utilizado numa etapa posterior, contudo, o primeiro é algo contencioso, já que a estimação do *pitch* fundamental F0 provém de diferenças associadas à intensidade dos fonemas produzidos em cada registo, sem sugestão de um controlo que emule um padrão de fala normal. Mcloughlin nota em [18] como, por estar dependente da comparação de segmentos não-vozeados com amostras de fala normal, o processo de melhoramento espectral desta abordagem não é adequado para uma aplicação em tempo real.

3.2 CELP

No trabalho supramencionado [18] e em [12], Sharifzadeh e Mcloughlin propõem um conjunto de modificações ao anterior modelo MELP, oferecendo especial relevo à melhoria espectral de fonemas sussurrados no respeitante à correção da localização e largura de banda das formantes, sublinhando a importância que o primeiro destes pontos tem no impacto perceptual de vogais reconstruídas.

A solução apresentada por Mcloughlin recorre a um codificador do tipo *code-excited linear prediction* (CELP) também ele adaptado às condições do problema. No que diz respeito às questões colocada no parágrafo anterior, são utilizados modelos probabilísticos, uma *Probability Mass-Density Function*, com base em conhecimento apriorístico da relação entre pares de fonemas com e sem vozeamento para determinar e corrigir a estrutura das formantes.

A maior deriva do modelo CELP tradicional dá-se no modo como a excitação é colhida. Num caso típico, a informação de *pitch* é extraída de uma predição de longo termo que caracteriza o comportamento periódico do sinal a sintetizar. Essa informação é então aplicada sobre um sinal gaussiano de média nula extraído do *codebook*, que indexa diferentes sinais desse mesmo tipo para serem usados como excitação, ao invés do binário estrito, V/UV do LPC (ver secção 2.2.2). No artigo referido, dada a falta de periodicidade do sussurro, a informação de longo-termo não é aplicável, e utiliza-se uma técnica de inserção de *pitch templates* que introduz uma série de fatores de *pitch* que tem por base o número de formantes do fonema classificado como alvo de vozeamento.

Mcloughlin levanta algumas considerações acerca dos principais limites desta solução em [12]. Destaca a falta de resposta concreta ao problema do tratamento de fonemas que devem ser não-vozeados, e aponta também para o facto de que a classificação de fonemas independente de utilizador é explicitamente deixada de parte em [18]. A decisão quanto ao nível do *pitch* a aplicar também é problemática, pois o *contour* empregue na técnica anterior é de carácter curvilíneo, de modo a emular a evolução de típica do *pitch* ao longo de uma frase. No entanto o baixo SNR dos sons desprovidos de *pitch* dificulta a classificação, e reduz consideravelmente a qualidade sistema.

3.3 Conversão Paramétrica

Ainda em [12], o autor apresenta uma possível solução para os dois problemas levantados em [18] e que foram apontados no parágrafo anterior.

A conversão paramétrica tem como objetivo delinear um *countour* plausível de F0 guiado pelas relações observadas entre as trajetórias das formantes e do *pitch* durante a fala.

No método aqui discutido, após uma análise deslizante, determinam-se valores candidatos de formantes por um critério dado em função da energia média calculada em cada frame, e constrói-se uma mistura/soma de cossenos cujas amplitudes refletem a energia calculada. O F0 é posteriormente obtido pela expressão:

$$F0 = \xi|F3 - F2| + \alpha|F2 - F1|$$

onde $F_{1,2,3}$ representam o valor das frequências formantes, e, ξ e α, constantes usadas para regular o valor médio do *pitch*.

Este trabalho representa um passo limitado no sentido de aproximar uma prosódia de fala natural, atingindo, como decorre da expressão anterior, apenas o seguimento da fundamental com as formantes ao longo da fala, o que não é necessariamente realista.

3.4 Estimação de pitch através do espectro

Konno *et al*, no artigo [17], expõem uma metodologia para a recuperação de *pitch*, que não depende de *input* externo vozeado adaptado ao falante. Esta incide sobre a questão da formatação do *pitch countour* durante a acentuação, e o impacto sobre a inteligibilidade que essa mesma confere às palavras proferidas e sintetizadas no contexto de uma linguagem *pitch-accent*, como é o caso do japonês, onde diferenças semânticas podem ser introduzidas em palavras homógrafas pela alteração da entoação, isto é, onde existe, uma ligação linguística direta à forma como o *pitch* evolui ao longo da locução.

A informação que serve de fonte à escolha de F0 é extraída de um número de ensaios nos quais foram gravadas as pronúncias de 5 vogais sussurradas por 5 oradores, 3 masculinos e 2 femininos. Estes ensaios consistem no sussurro de cada uma das vogais individuais por parte de um orador que ouve em simultâneo, através de auscultadores, um tom sinusoidal puro como guia. Os oradores foram instruídos a tentar produzir o tom guia através do sussurro. Deste modo, a cada gravação é associado um *intended pitch* que corresponde à frequência da sinusoide pura auscultada.

A gravação é submetida a uma análise espectral de onde se extraem os outputs de um filtro de bancos Mel. Verificou-se a necessidade de construir um preditor de *pitch* específico ao locutor, pois o tom percebido no sussurro é impreciso, levando a que o guia puro escolhido por cada um dos locutores divirja significativamente. O *predictor* é consequência de uma múltipla regressão linear de parâmetros calculados segundo a análise da filtragem perceptual, e pelo guia puro que dá a frequência almejada pelo falante. Finalmente, a fala é sintetizada através de um *vocoder*.

Os resultados mostraram uma boa recuperação da acentuação, mas o desempenho da síntese é notado como pobre.

3.5 Pesquisa na FEUP

De particular relevo para o trabalho do projeto DyNaVoiceR identifica-se um trabalho de investigação preliminar de 2016 [19] e a tese de mestrado, de 2015, [20] com este relacionada. O método proposto segue a estrutura de um codificador perceptual de alta qualidade totalmente paramétrico que emprega um método de análise e síntese baseado em transformadas. Processam-se

trechos de áudio amostrados a 22,05 kHz, garantindo a acomodação do sinal numa banda de 10 kHz. O autor de [19] extrai um conjunto de conclusões acerca da estimação de *pitch* e envolvente espetral, a segmentação, e a segmentação dos fonemas que devem ser alvo de vozeamento. Uma das ilações indica que resultados de aspeto natural exigem, durante a análise, uma segmentação fonética de janela temporal melhor que 10 ms [19]. Esta reflete-se na opção de analisar o sinal segundo duas resoluções temporais: 5,8 e 23,2 ms; além do mais conclui-se acerca da inviabilidade de um método estatístico para a classificação de fonemas, optando-se pelo cálculo e comparação de descritores dos segmentos, como a energia do sinal e a sua distribuição ao longo das frequências.

Importa frisar a importância dada em [19] à preservação da prosódia e informação idiossincrática, e o modo como o *contour* de F0 é modelado. Neste caso, realizou-se a análise da evolução de *pitch* da fala de um jornalista televisivo português, notando-se a repetição de um padrão durante o início de sílabas vozeadas e ditongos. Estimado por uma média, esse mesmo padrão é então forçado aos segmentos artificialmente vozeados, reiniciando-se a cada novo segmento.

A naturalidade da fala produzida, contudo, aparece limitada pelo processo de estimação de envolvente. Foram testados dois métodos, tendo ambos resultado na síntese de vogais com um perfil muito plano, e próximo do aspeto não vozeado, do que o desejado. Este resultado leva o autor a concluir acerca da necessidade de um alargamento da base de dados de envolventes espectrais protótipo, com vista a tornar mais eficiente a análise e mais natural a síntese.

3.6 DyNaVoiceR

O projeto DyNaVoiceR tem acolhido diversos trabalhos de pesquisa sobre o tópico da conversão de fala sussurrada em fala normal enquadrados nas tarefas listadas na Secção 1.1. Quanto aos resultados obtidos, resumidos em [21], destacam-se abaixo os do primeiro bloco nas suas respetivas sub-tarefas.

- A.1 *Dysphonic voice database and feature characterization* - A construção de uma base de dados que inclui gravações de voz de 15 oradores do género masculino e 15 do género feminino. As gravações incluem a enunciação de vogais, dissílabos e de trechos de texto nos dois modos de articulação, normal e sussurrado. Este segmento do projeto foi liderado pela Universidade de Aveiro.

- A.2 *Accurate harmonic analysis and modelling of natural voiced sounds* - A criação de uma estrutura de análise e controlo de *features* de sinal e realização de testes subjetivos onde se mediu o impacto perceptual de micro-variações de F0 e da fase dos harmónicos vocais. O primeiro caso não suscitou diferenças significativas na assinatura vocal; o segundo, revelou que existe informação idiossincrática a retirar da fase da estrutura harmónica de vozes com *pitch* baixo. Estas resultados fundam a investigação reportada em [22], que dá seguimento ao estudo da identificação de orador com base em atributos vocais de fase, F0 e magnitude de espectro. Esta sub-tarefa foi desenvolvida na FEUP.

- A.3 *Accurate vocal tract filter modelling* - Foi composta uma base de dados de *templates* da magnitude de espectral baseadas em LPC a partir de amostras vocais obtidas em A.1, com o propósito de serem usadas como filtro para a síntese de vogais. Daqui surgiu a motivação para um estudo da correlação cruzada entre os modelos espectrais, obtidos da base de dados, de vogais de um mesmo orador. Este trabalho foi liderado pela U.Aveiro, e realizado na FEUP.
- A.4 *Accurate glottal source estimation and modelling* - Este módulo, desenvolvido pela Faculdade de Medicina da Universidade do Porto, reforçou os resultados obtido em A.2 quanto à informação idiossincrática que pode ser extraída do pulso glotal.
- A.5 *Adaptive phonetic segmentation techniques in dysphonic voice* - Desenvolvimento de estratégias para a segmentação de sons oclusivos.

A tese de mestrado de 2020 de Marco Oliveira [15], investigador que já contribuíra para o projeto durante a fase A.3, retoma a modelização da envolvente espectral com base no ambiente do DyNaVoiceR cujo desenvolvimento foi detalhado. Nesta, exploram-se sobretudo técnicas de modelização de envolvente e identificação de vogais sussurradas.

Para a análise e modelização de envolvente, o autor recorre a uma seleção de gravações presentes na base de dados de recolhida no âmbito da atividade A.1. Para o caso das vogais sussurradas, utiliza um modelo LPC de 22ª ordem que recebe os coeficientes de autocorrelação resultantes da aplicação do teorema de Wiener-Khintchine sobre os valores da função Densidade Espectral de Potência da transformada de Fourier discreta da porção de sinal janelado. No que diz respeito às vogais vozeadas, utiliza funções já implementadas no ambiente DyNaVoicer para extração de F0. Num primeiro teste, as envolventes pré-processadas foram usadas como base para a ressíntese de segmentos à escolha. Verificou-se que a definição fraca das formantes resultou numa síntese de má qualidade a nível perceptual.

A abordagem tomada para a classificação de segmentos segue a ideia, introduzida em A.3, de usar como filtro num segmento alvo o modelo de envolvente espectral, extraído de uma biblioteca de características colhidas de amostras de diferentes fonemas. O modelo a aplicar será aquele que verifica uma maior correlação estatística entre a envolvente extraída do segmento e o modelo presente na biblioteca. É importante notar que, neste contexto, a biblioteca de modelos é específica do orador. Dos testes sobre a metodologia, concluiu-se que esta era apta para operar em tempo real; que a taxa de acerto da classificação variava muito conforme a vogal testada, pelo que exige a exploração de um método mais robusto, ou uma análise mais abrangente de diferentes fonemas, e que a utilização de modelos no domínio cepstral representam ganhos computacionais, pela redução do número de parâmetros utilizados, sem grandes impactos a nível da taxa de acerto do algoritmo.

3.7 Resumo do capítulo

Da revisão bibliográfica aqui realizada podemos notar quais é que são as principais mecânicas e abordagens que têm sido aplicadas ao problema da conversão de fala, delineando também o

modo como os problemas da aferição de *pitch*, e da caracterização de um espectro que esteja fundamentalmente ligado à fala sussurrada ainda se mantêm.

Capítulo 4

Plataforma de Processamento

4.1 Sobre o STM32F746G Discovery

A plataforma proposta neste trabalho para acomodar o ambiente de processamento de áudio desenvolvido no decurso das sucessivas etapas do projeto DyNaVoiceR, como referido na Introdução, é o STM32F746G Discovery. Trata-se de um *kit* de desenvolvimento com um processador 212 MHz Arm Cortex-M7 que integra um número de *features* apropriadas à receção, como unidade de teste, de aplicações em fase de design. Este dispositivo foi adotado pela FEUP, no ano letivo de 2021/2022, como principal plataforma para a componente laboratorial da unidade curricular (UC) *Fundamentals of Signal Processing* (FunSP) (código L.EEC025).

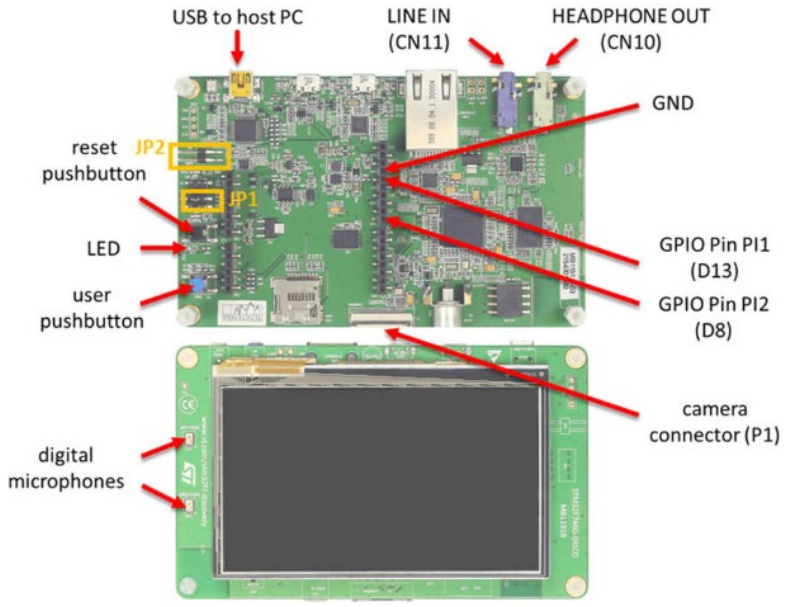


Figura 4.1: A placa STM32F746G Discovery. Adaptada dos apontamentos da UC de FunSP.

A *kit* é capaz de processar som captado em tempo real por intermédio de um codec Wolfson WM8994, pelo qual se gerem as saídas e entradas de áudio (conforme indicado na Figura 4.1),

e de realizar o traçado simultâneo de gráficos no *display*. O código desenvolvido e implementado é gerido pelo ambiente Keil MDK-ARM que permite compilar, importar, e depurar, com um *debugger* ST-Link/V2 incorporado, programas em linguagem "C".

4.2 Configuração de DMA

O *framework* da estrutura de processamento a ser implementada, cujos estágios aparecem detalhadas em secções posteriores deste documento, é herdado da plataforma desenvolvida em Matlab no contexto do DyNaVoicer. Neste, o processamento está assente num método *frame-based*, baseado em *buffers* de 512 amostras, de 16 bits, obtidas a uma taxa de amostragem de 22,05 kHz. O kit dispõe já de mecanismos que permitem configurar a transferência de dados em tramas, nomeadamente, sobre a forma de *block-based* I/O *interrupts* geridos por DMA. O DMA consiste num controlador que intermedeia e controla as transferências de dados entre um periférico como, neste caso, o *bus* I/O do *kit*, e a memória do sistema de forma autónoma, reduzindo a carga exercida sobre o CPU.

No STM32F746G, o método instalado para o processamento *frame-based* assistido por DMA apresenta um esquema de *buffer* múltiplo em modo *ping-pong*, cujo funcionamento e relação com o estágio de *Digital Signal Processing* (DSP) se encontram ilustrados na Figura 4.2.

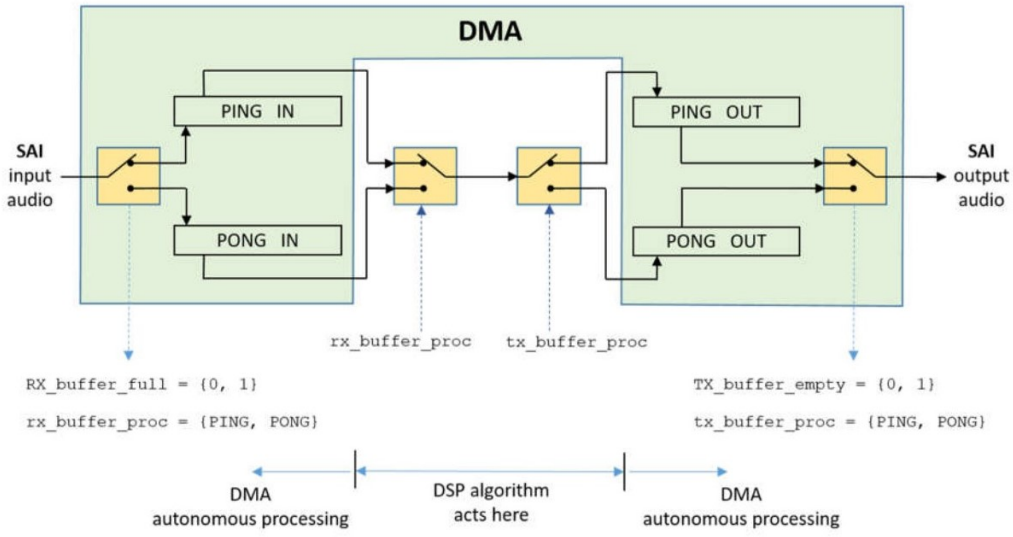


Figura 4.2: Representação esquemática da operação do DMA.

O tamanho do bloco de amostras à entrada, representando o conjunto do total das amostras transferidas segundo os *buffers* PING e PONG, ou seja, o total de *sampling moments*, é dado pela definição da constante PING_PONG_BUFFER_SIZE à qual é atribuido o valor de 512. A frequência de amostragem, por sua vez, é dada como argumento da função de inicialização do sistema do *kit*.

4.3 Medição do atraso de processamento mínimo

Um dos principais requisitos funcionais do sistema a desenvolver é o da operação do assistente em tempo-real. Ao problema colocado pela necessidade de produzir uma fala sintetizada que seja complementar à fala produzida por um orador, formando um único percepto, é acrescida uma exigência temporal cuja natureza se prende ao atraso introduzido pela operação da operação de conversão A/D e D/A do kit. A latência *end-to-end* deste deve ser suficientemente baixa para que os sons não sejam percebidos como ecos um do outro. Este fenómeno da percepção de dominância de uma frente de onda face às suas repetições, causados por reflexão sonora ou, como neste caso, pela emissão de sinais similares por fontes distintas, é conhecido na literatura como efeito de precedência. Para além deste fenómeno central de precedência e fusão sonora, o problema da latência sobre a percepção da fala pode ainda implicar outros artefactos, dos quais podemos destacar as perturbações audiovisuais de desfazamento entre a produção do som e o seu correspondente visual pela importância que podem ter sobre a qualidade do assistente.

Em [23], Litovsky et al. fazem um levantamento de alguns valores já estudados para o limiar de fusão, ou seja, os valores de latência para os quais não são perceptíveis ecos distintos. Dentro desses, podemos destacar os estudos de Haas (1951), e de Lochner (1958), onde se utilizaram estímulos de fala transmitidos por altifalantes num ambiente *free field*, e onde se aponta para um limiar superior de atraso de 30-40 ms, no primeiro, e de 50 ms no segundo. Outro valor relevante para estimar o impacto perceptual do atraso de processamento pode ser depreendido dos resultados de [24], onde se estuda o efeito audio-visual da introdução de atrasos entre sinais acústicos e visuais, i.e o movimento da face, durante a fala. No trabalho citado é observada uma tolerância de 80 ms para este tipo de assincronia.

Este levantamento permite definir, senão uma meta, sendo aparente que o caso mais desejável será aquele onde o atraso não supera os 50 ms, pelo menos um termo de referência para o efeito expectável produzido pelo atraso. Esta investigação dá então o motivo desta primeira fase de testes, onde se pretende medir o *delay* introduzido ao nível mais primitivo do sistema, contando apenas com o seguimento da entrada para a saída, sem qualquer modificação, passando apenas pelo de *buffering* associado ao método de DMA que rege todo o processo.

4.4 Metodologia e discussão de resultados

O procedimento para a medida da latência consiste na comparação do desvio no tempo entre uma onda dente-de-serra, gerada por uma fonte de sinal, e aquela que é produzida à saída do kit que a tem como entrada para diferentes comprimentos do *buffer* DMA. Além do material já referido, foram utilizados adaptadores para divisão dos dois canais *stereo* à entrada e saída do kit, e um osciloscópio para a representação das ondas e medição dos intervalos. O programa utilizado consiste numa versão modificada do programa exemplo stm32f7_loop_dma.c, onde são colocados nos *buffers* à saída os valores à entrada, invertendo-se somente o sinal destes últimos por motivos de visualização da forma de onda.

Aponte-se que, dada a estrutura apresentada na figura 4.2, onde se verifica que o processamento compreende uma transferência de PING_PONG_BUFFER_SIZE amostras tanto à entrada como à saída do algoritmo DSP, que o atraso mínimo esperado deverá ser de $\Delta_{proc} = 2 * PING\_PONG\_BUFFER\_SIZE/fs$, representando fs a frequência de amostragem. Este atraso de processamento limita a frequência da onda escolhida para a medição, pois o atraso só é mensurável para valores inferiores a $1/Delta_{proc}$.

Tendo por objetivo o estudo da evolução da latência conforme o tamanho do *buffer* DMA, apresenta-se na tabela seguinte o histórico da sequência de medições realizadas. Dada a relevância do caso em que o *buffer* tem um comprimento de 512 amostras, repetiu-se a sua medição entre cada caso de comprimento diferente de modo a avaliar o desvio face ao valor esperado.

Buffer len	delay (ms)	f0 (Hz)
256	24,4	20
512	48,4	20
128	10,8	20
512	47,2	20
1024	91,6	10
512	48	20

Tabela 4.1: Histórico de medições

A partir destes valores, podemos concluir que a latência mínima introduzida pelo processamento de uma única *frame* de 512 amostras, isto é, o tempo implicado na recepção e transmissão da mesma, é de, em média, 47,9 ms. Contudo, como será detalhado no próxima secção, o processamento dos dados será realizado sobre pares de *frames* e, como tal, ao atraso de processamento irá acrescer um atraso algorítmico. Este atraso consiste na necessidade de retenção de uma *frame* durante um ciclo, o que implica um atraso equivalente a uma recepção. Tendo por base o valor medido para o atraso com *buffers* de 512 amostras, podemos então calcular o atraso associado à recepção, ou transmissão, isolada de uma *frame*, o que corresponde a metade do valor médio medido, ou seja, a um valor de aproximadamente 23,95 ms. O atraso total esperado será então de em torno de 71 ms.

Este valor foi verificado experimentalmente quando, após a implementação do algoritmo que inclui a estrutura de análise e síntese, se mediu novamente o atraso para as mesmas condições (i.e, o mesmo tipo de sinal à entrada). Mediu-se um atraso de 71,2 ms, o que está de acordo com o valor teórico calculado.

4.5 Resumo do capítulo

Neste capítulo apresentou-se a plataforma de *hardware* que será alvo de teste na próxima fase da dissertação, analisaram-se os modos como o atraso de processamento pode interferir e impactar a percepção e produção da voz modificada, e mediu-se a média do atraso mínimo real,

comparando-o com o valor estimado pelos parâmetros da leitura de dados (i.e, tamanho dos *buffers* e frequência de amostragem).

Capítulo 5

Estrutura de Análise e Síntese

Nos próximos capítulos deste trabalho, uma primeira aproximação ao mecanismo de vozeamento será tentada através da emulação do funcionamento de uma EL, isto é, pela adição de uma onda periódica com frequência fundamental F0 e forma de onda arbitrária ao sinal de voz. Para isso, utilizou-se um método de síntese no domínio das frequências, por transformada *Odd Discrete Fourier Transform* (ODFT), baseado no algoritmo desenvolvido em [4], replicando-se os métodos de análise e síntese nos tempos pelo uso de um método *Overlap-Add* (OLA), com 50% de sobreposição e janela sinusoidal. Neste capítulo será introduzido o funcionamento da estrutura de análise e síntese.

5.1 Algoritmo DSP - Overlap-Add e ODFT

Lembrando a posição do bloco DSP no contexto do esquema da acesso à memória, representado na Figura 4.2, o algoritmo DSP irá processar e transmitir blocos de 512 amostras geridos pelo DMA conforme um ciclo que atua de cada vez que um bloco de dados DMA é preenchido. A Figura 5.1 ilustra de forma simplificada a cadeia de sinal do algoritmo.

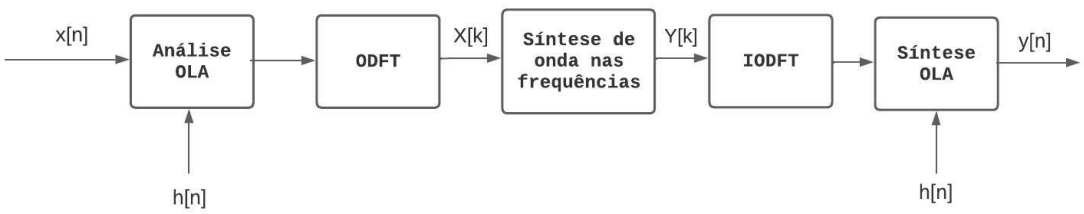


Figura 5.1: Diagrama de Blocos - Algoritmo DSP

Nesta Figura, os blocos OLA Análise/Síntese representam, respetivamente, os processos de análise e de síntese por OLA com janela sinusoidal e 50% de sobreposição, os blocos ODFT/I-ODFT as formas direta e inversa da ODFT, e h[n] a função da janela sinusoidal.

A janela utilizada corresponde à raíz quadrada da janela de Hanning deslocada de $\frac{1}{2}$ [25], e é dada nos tempos pela equação:

$$h_s(n) = sin\frac{\pi}{N}(n+0.5), \quad n = 0,1,...,N-1. \tag{5.1}$$

O uso deste tipo de janela é comum no campo da codificação de áudio [4] pois, como investigado em [26], quando utilizada numa estrutura OLA seguida por transformada, aplicando o janelamento tanto à sequência de análise como à sequência recuperada pela transformação inversa, torna-se possível obter reconstrução perfeita do sinal.

O mecanismo OLA consiste, neste caso, na segmentação de um sinal de áudio em janelas com 1024 amostras de comprimento deslocadas de um *step* de 512 amostras, pelo que se impõe uma sobreposição de 50%. O sinal reconstruído será dado pela concatenação das consecutivas porções sobrepostas, como ilustra a Figura 5.2:

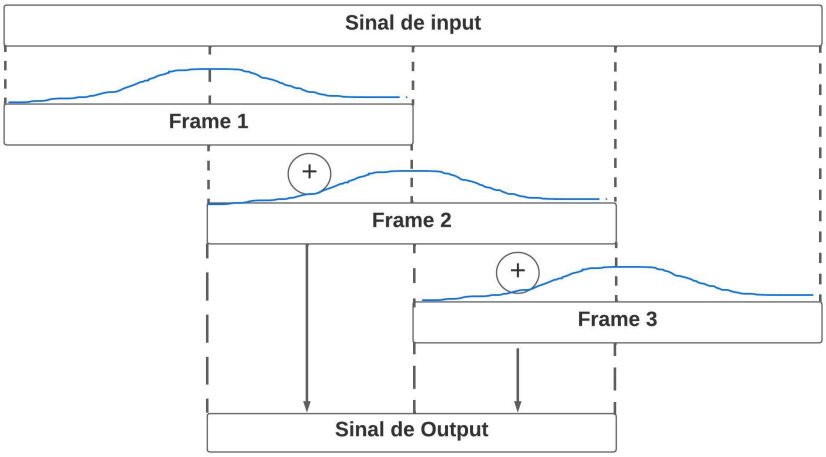


Figura 5.2: Esquema de funcionamento do algoritmo *Overlap Add*

Dado que, como referido à cabeça desta secção, o algoritmo DSP opera segundo um ciclo onde se processam blocos de 512 amostras, as *frames* de comprimento 1024 são resultado da concatenação de blocos de ciclos consecutivos: considerando uma iteração i, as primeiras 512 amostras de uma *frame* i correspondem ao bloco de dados captado durante a iteração $i-1$ e, as últimas, ao bloco à entrada do sistema durante a iteração vigente. O janelamento está representado graficamente pela linha azul que ilustra a envolvente da função de janela. É importante, contudo, apontar que a criação de cada *frame* implica a multiplicação do segmento de dados que lhe corresponde pela janela em dois momentos: uma durante a criação da própria *frame*, na fase de análise, e, novamente, antes de se somarem as partes sobrepostas entre *frames*, na fase de síntese.

Tal como a escolha de janela, a opção de se utilizar a ODFT é motivada pela vantagem que esta variante traz à etapa de análise, nomeadamente, a maior definição que é obtida nas frequências mais altas e mais baixas da banda [27], um detalhe particularmente relevante para a extração da *Power ODFT* do sinal que será abordada no próximo capítulo. A representação nas frequências que se obtém desta transformada apresenta, face à DFT, um deslocamento de $\frac{\pi}{N}$ para a direita em

todas as amostras em frequência, resultado da multiplicação do sinal nos tempos pela exponencial $exp_n = e^{-j\frac{n\pi}{N}}$ antes da aplicação da transformada DFT. Este último passo implica a necessidade de se multiplicar o sinal recuperado durante a DFT inversa pelo conjugado de exp_n, para que o efeito nos tempos por seja revertido. Outro aspecto a destacar é de que, a ODFT de um sinal real no intervalo normalizado de $[0, \pi]$ é dado por $\frac{N}{2}$ coeficientes únicos [28], ao passo que a DFT é composta por $\frac{N}{2} + 1$.

5.2 Implementação no *kit*

Nesta secção, o código C que será apresentado implementa os quatro blocos que foram destacados da Figura 5.1 na secção anterior. O excerto e a figura que se seguem apresentam e ilustram o funcionamento da etapa de análise nos tempos. Sublinhe-se que a variável currBuff, onde são concatenados os blocos de dados das iterações sucessivas, diz respeito a um *array* de comprimento 1024 (N) de uma *struct* COMPLEX composta por dois valores *float* correspondentes à parte real e imaginária do sinal (real e imag). A manipulação da fase do sinal só se tornará necessária no domínio das frequências, pelo que a parte imaginária nos tempos, dado que este é um sinal real, é nula.

```
for (i = 0; i< N2; i++) // N2 corresponde a N/2
{
left_sample =  *rx_buf++; //canal esquerdo
right_sample = *rx_buf++; //canal direito

currBuff[i].real = (old_inBuff[i]*sinW[i]); // Coloca no buffer as 512 amostras
    captadas no ciclo anterior
currBuff[i].imag = (float32_t) 0.0;
currBuff[i+N2].real = (((float32_t)left_sample) * sinW[i+N2]);
currBuff[i+N2].imag = (float32_t) 0.0;
old_inBuff[i] = ((float32_t)left_sample);    //substitui o buffer antigo pelos
    valores novos do bloco DMA
}
```

Listing 5.1: Análise

```
ifftFlag = 0; //flag = 0 => FFT Direta
arm_cmplx_mult_cmplx_f32((float32_t *)(currBuff),(float32_t *)(direxp),(float32_t
    *)(currBuff), N); //produto complexo pela expn
arm_cfft_f32(&arm_cfft_sR_f32_len1024, (float32_t *)(currBuff), ifftFlag,
    doBitReverse);
// Inverse FFT----------------------------
ifftFlag = 1; // flag = 1 =>  FFT inversa
arm_cfft_f32(&arm_cfft_sR_f32_len1024, (float32_t *)(currBuff), ifftFlag,
    doBitReverse);
arm_cmplx_mult_cmplx_f32((float32_t *)(currBuff),(float32_t *)(invexp),(float32_t
    *)(currBuff), N); //produto complexo pelo conjugado de expn
```

Listing 5.2: ODFT e IODFT

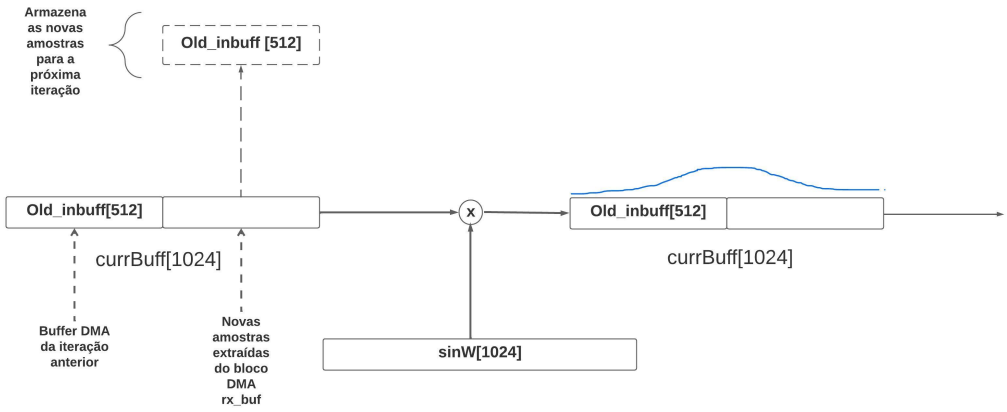


Figura 5.3: Esquema de funcionamento do algoritmo *Overlap Add* (Análise)

Na listing 5.1, o apontador rx_buf corresponde ao vetor de dados à entrada do sistema (equivalente ao rx_buf_proc ilustrado em 4.2), e compreende os valores de 512 *sampling instants*, sendo que cada um desses é composto por um par de amostras stereo agrupadas de forma alternada no vetor.

As funções arm_cmplx_mult_cmplx_f32() e arm_cfft_f32() fazem parte da biblioteca de processamento digital de sinal da CMSIS [29], e tratam, respetivamente, da multiplicação entre *arrays* de complexos e da aplicação *in-place* da DFT.

```
1 for(i = 0; i < N2; i++) {
2     outSample = (old_outBuff[i] + (currBuff[i].real*sinW[i]));
3
4     *tx_buf++ = (int16_t) (outSample); //left out
5     *tx_buf++ = (int16_t) (outSample); //right out
6
7     old_outBuff[i] = ((currBuff[i+N2].real)*sinW[i+N2]);}
```

Listing 5.3: Reconstrução

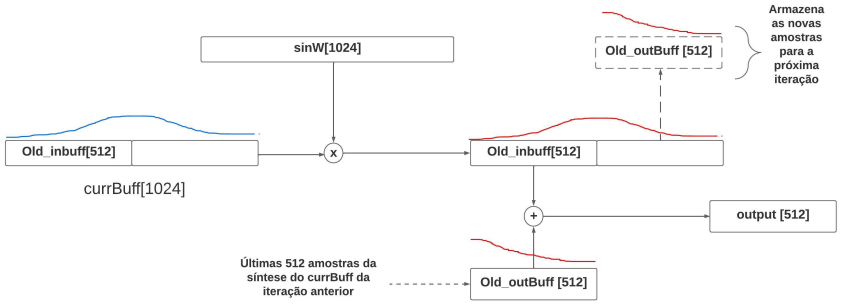


Figura 5.4: Esquema de funcionamento do algoritmo *Overlap Add* (Síntese)

Sublinhe-se que o bloco de 512 amostras recebidas numa dada iteração é guardado em memória em duas ocasiões diferentes, mas que só é implicado no cálculo das amostras transmitidas

durante a iteração seguinte, pelo que o *step* de 512 *samples* que a sobreposição requer é responsável por introduzir o *delay* de uma *frame*, como mencionado no capítulo anterior.

5.3 Testes e Validação

A sequência de testes realizados em torno do programa anterior pretende garantir: a reconstrução perfeita do sinal, pois que todas operações implicadas no processo devem ser sem perdas; a estabilidade do sinal ao longo de toda a banda de frequências (i.e até à frequência de Nyquist), e avaliar a capacidade de processamento do kit procurando definir uma margem dentro da qual o kit possa continuar a operar em tempo-real sem artefactos.

5.3.1 Reconstrução de Sinal e *Headroom* de Processamento

A fim de medir a qualidade do sinal produzido na placa e o funcionamento correto de cada um dos blocos de processamento implementados, utilizou-se um gerador de função para produzir um sinal de *input* bem definido como referência, comparando-o então com sinal de *output*, fazendo variar os estágios de processamento que estão em operação, começando por testar o funcionamento do OLA, seguido do OLA+FFT e terminando com o OLA+ODFT.

(a) *OLA* (b) $OLA + FFT$ (c) $OLA + ODFT$

Figura 5.5: Visualizações de Onda Dente-De-Serra em Três Casos

Na Figura 5.5 podemos verificar o comportamento dos diferentes algoritmos quando colocados em série. Nestas fotografias, o canal 1 (gráfico amarelo) está ligado diretamente à fonte de sinal, enquanto o segundo canal (gráfico azul) está ligado ao canal esquerdo do output do *kit*. A onda de referência é uma dente-de-serra de frequência 1,3 kHz e amplitude de 5Vpp. No anexo do documento encontram-se fotos aos testes de outros tipos de onda. A onda gerada corresponde à esperada contanto se considere a banda em que o kit opera, motivo pelo qual sobrevivem somente os harmónicos de baixa frequência.

Foi importante conhecer, antes de terem sido implementados novos módulos de processamento, o limite da capacidade de cálculo do *kit*. Como tal, fez-se um teste simples que consistiu em forçar a repetição do cálculo das transformadas direta e inversa (Listing 5.2) até que fossem manifestados problemas na representação digital obtida pelo osciloscópio.

Verificou-se que até as 31 iterações do ciclo em que se encerrou o cálculo, o *output* continuou estável, mantendo a onda a estrutura harmónica que lhe é própria. A partir desse limiar deixou de

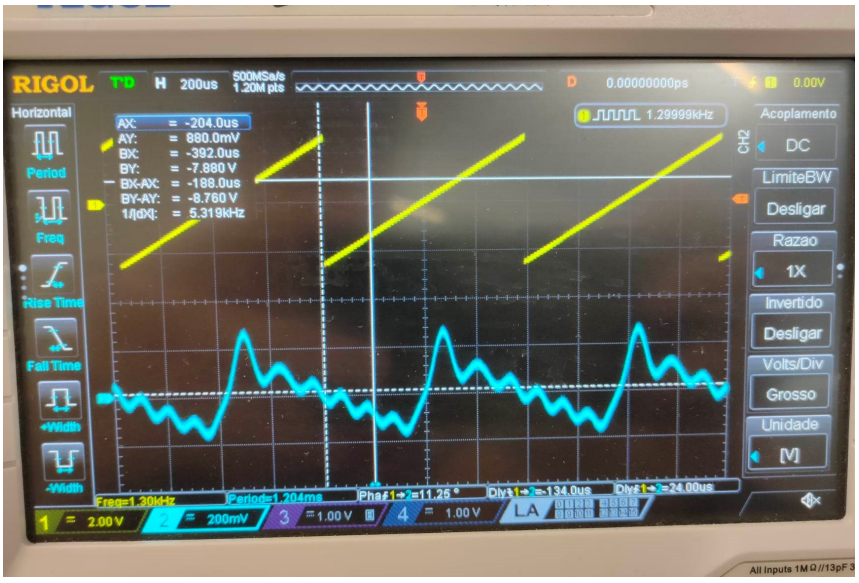


Figura 5.6: Onda dente-de-serra com 31 repetições da FFT

ser possível obter uma representação gráfica da onda de saída. Isto significa que a carga computacional admitida pelo *kit* é 30 vezes superior à que a implementação da estrutura de análise e síntese constitui, o que deixa uma boa margem para a incorporação de novos blocos de processamento.

5.3.2 Discussão dos resultados

Os testes levados a cabo nesta etapa do trabalho têm a importância de validar o funcionamento da estrutura sobre o qual assentarão os módulos posteriores do algoritmo.

Vimos que o *output* gerado em cada um dos estágios de análise e síntese era congruente com o esperado dentro da banda de frequências em que se está a operar, a menos de uma inversão do sinal, e que, com uma capacidade para a repetição do conjunto FFT+IFFT de trinta e uma vezes, que existe ainda um *headroom* de cálculo considerável, o que permitirá o desenvolvimento continuado do algoritmo.

5.4 Resumo do capítulo

Neste capítulo, expôs-se o funcionamento da estrutura de análise e síntese sobre a qual será elaborada a etapa de síntese em frequência. Apontaram-se as vantagens que os tipos de janela e transformada escolhidos trazem ao processamento ideado, e verificou-se o cumprimento dos requisitos de reconstrução da implementação do algoritmo no *kit*, além de ter sido verificada a capacidade para um continuado desenvolvimento do algoritmo pela medição da margem de processamento aferida através do número de vezes que o cálculo da transforma pode ser repetido sem que se sobrecarregue o *kit*.

Capítulo 6

Síntese e adição nas frequências de uma onda periódica de amplitude constante

Neste capítulo, será apresentada a primeira abordagem a um método de vozeamento artificial operado em tempo-real no *kit*. O método aqui proposto baseia-se no princípio de funcionamento da EL, pretendendo-se, portanto, introduzir no sinal de fala sussurrada um sinal harmónico de amplitude e F0 constantes, tomando partido de determinadas características da estrutura de análise apresentada no capítulo anterior para gerar, de um modo computacionalmente económico, uma forma de onda mais estruturada do que uma simples sinusoide através de síntese de sinal no domínio das frequências.

6.1 Algoritmo de Síntese de Onda nas Frequências

Em [4], propõe-se um método para estimação precisa das características de sinal de sinusoides estacionárias. Nesse estudo, demonstra-se como, em função da separação dos canais da ODFT em $\frac{2\pi}{N}$, se torna possível detetar sinusoides com uma frequência normalizada de $\omega = \frac{2\pi}{N}\ell$ simultaneamente por dois canais da ODFT, nomeadamente, nos *bins* ℓ e $\ell - 1$. De facto, considerando só a gama de frequências positivas:

$$X_O[K] = \frac{NA}{4}[\delta[k-\ell+1]e^{j(\phi-\frac{\pi}{2N})} + \delta[k-\ell]e^{j(\phi+\frac{\pi}{2N}+\pi)}] \tag{6.1}$$

e considerando $N >> 1$, e $\phi = 0$, a expressão é aproximada para:

$$X_O[K] \approx \frac{NA}{4}[\delta[k-\ell+1] - \delta[k-\ell]] \tag{6.2}$$

ou, se $\phi = \pi$

$$X_O[K] \approx \frac{NA}{2}[\delta[k-\ell] - \delta[k-\ell+1]] \tag{6.3}$$

Decorre, portanto, que uma sinusoide pode ser gerada ao se forçar na *frame* de dados que resulta da ODFT um par de impulsos nos *bins* ℓ e $\ell - 1$, sendo que a escolha de *bin* ℓ determina

a F0 da onda sintetizada. Contudo, a escolha de *bin* ℓ é condicionada pela estrutura da ODFT como um banco de filtros, onde cada canal está centrado em múltiplos ímpares da frequência normalizada $\frac{\pi}{N}$, a ter um valor par de modo a evitar a inversão de sinal de uma *frame* para a seguinte, atendendo a que a sobreposição entre estas é de 50%.

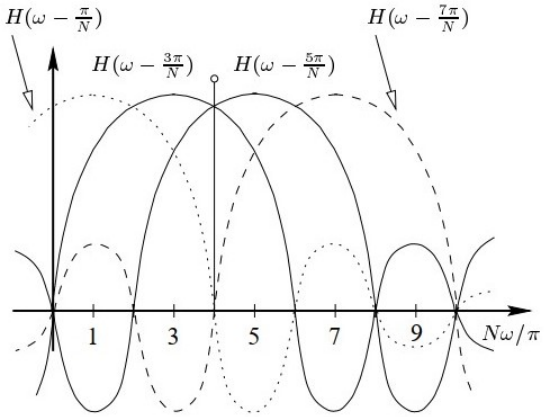


Figura 6.1: Resposta em frequência dos quatro primeiros canais do banco de filtros ODFT, adaptado de [4]

6.1.1 Síntese de Sinusoide

A verificar a hipótese anterior, elaborou-se o seguinte programa, onde, num *array* COMPLEX de comprimento 1024, se introduz um pico no *bin* de frequência 26 e, consequentemente, um pico simétrico no *bin* 25. A primeira metade deste *array* é composto, de resto, por valores nulos, enquanto a segunda parte corresponde ao simétrico conjugado da primeira, pois este é um sinal real apesar de estar representado como um complexo (com parte imaginária nula). O *bin* ℓ escolhido deverá corresponder a uma F0 de

$$F0 = \frac{\ell}{N} FS \tag{6.4}$$

onde ℓ dá o valor do *bin*. A sinusoide gerada, neste caso, terá uma frequência de 560Hz.

```
for(i = 0; i < N; i++) //inicializa o array com valores nulos
{
    fData[i].real = 0.0;
    fData[i].imag = 0.0;
}

for(i = 1; i < 2; i++) // l e harmonicos
{
    fData[i*26].real = 5.0E06;
    fData[i*26].imag = 0.0;

    fData[(i*26)-1].real = -5.0E06;
    fData[(i*26)-1].imag = 0.0;
```

```
14
15      fData[N-1-(i*26)].real =  fData[(i*26)].real;
16      fData[N-1-(i*26)].imag = -fData[(i*26)].imag;
17
18      fData[N-(i*26)].real = fData[(i*26)-1].real;
19      fData[N-(i*26)].imag = -fData[(i*26)-1].imag;
20  }
```

Listing 6.1: Síntese de sinusoide nas frequências

O método de síntese aqui utilizado foi estudado em [30] no seguimento do trabalho supracitado, tomando proveito da equação de síntese acima indicada (Eq. 6.3). O código aqui listado constitui uma aproximação prática dessa mesma equação de síntese.

O excerto do Listing 6.1 foi usado para produzir a sinusoide apresentada na Figura 6.1, contudo, a forma como o ciclo iniciado na linha 7 está estruturado prevê a inclusão de harmónicos da onda fundamental, o número dos quais será dado pelo número de iterações do ciclo.

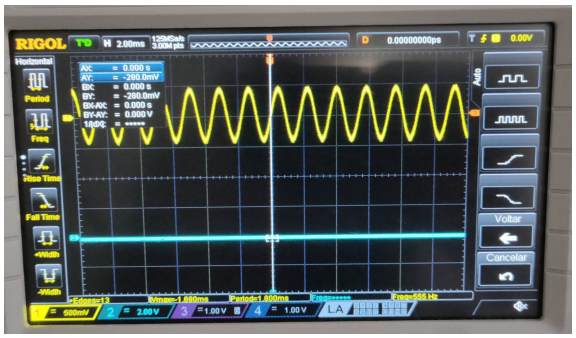


Figura 6.2: Sinusoide gerada segundo o Listing 6.1

6.1.2 Síntese de onda estruturada

Dado o intuito de incorporar a onda gerada ao sinal de voz, é relevante investigar o perfil espectral da onda a produzir. Não só o número da harmónicos que deverá constituir a onda, como também o perfil espectral poderão ter um importante impacto na qualidade do som produzido. Visto que a onda periódica é sintetizada visando cumprir a função que o impulso glotal cumpre na produção natural de fala (ver Secção 2.2.1.1), tirou-se proveito da caracterização da onda periódica tal como trabalhada no modelo Liljencrants-Fant [31], onde o espectro do pulso glotal apresenta um decrescimento monótono ao longo da frequência.

Levando também em consideração a forma como a energia do sinal de voz se concentra nos primeiros formantes, optou-se por reproduzir uma onda dente-de-serra, onde a magnitude dos harmónicos evolui de forma inversamente proporcional ao índice dos mesmos. Com esta alteração, e generalizando a partir da síntese da forma de onda sinusoidal, visto no Listing 6.2, para dar lugar à adição da onda ao *buffer* de dados *currBuff*, obtém-se o seguinte programa:

```
1 int ell = 26;
2 for(i = 1; i < 9; i++)
```

```
{
    currBuff[i*ell].real += (float32_t)(5.0E04/i);
    currBuff[(i*ell)-1].real += (float32_t)(-5.0E04/i);
}

for(i = 0; i<N2; i++)
{
    //preencher a segunda metade do buffer com o conjugado da primeira
    currBuff[(N-1-i)].real = currBuff[i].real;
    currBuff[(N-1-i)].imag = -currBuff[i].imag;
}
```

Listing 6.2: Síntese de onda estruturada nas frequências

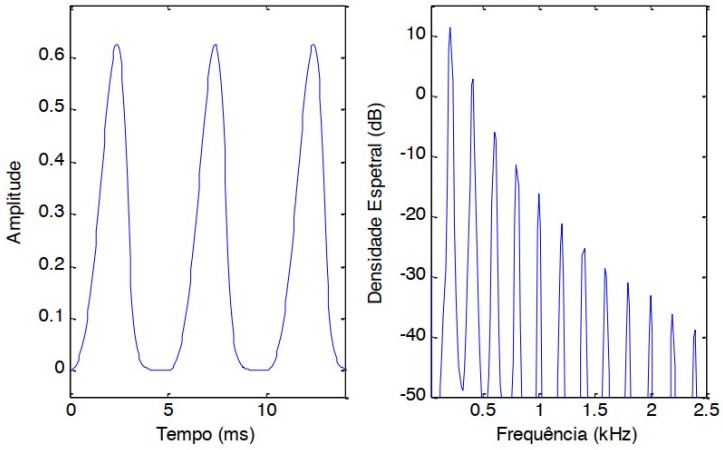


Figura 6.3: Forma temporal e respectivo espectro da onda glotal segundo o modelo idealizado Liljencrants-Fant, adaptado de [5].

Além do método aqui apresentado poderíamos optar por sintetizar a onda dente-de-serra no domínio dos tempos, o que exigiria uma maior complexidade em termos de código. Um método nos tempos foi, contudo, desenvolvido em Matlab, mas a qualidade do método, inferior ao da versão desde já conseguida, não justificou a sua inclusão nos testes percetivos.

6.2 Validação

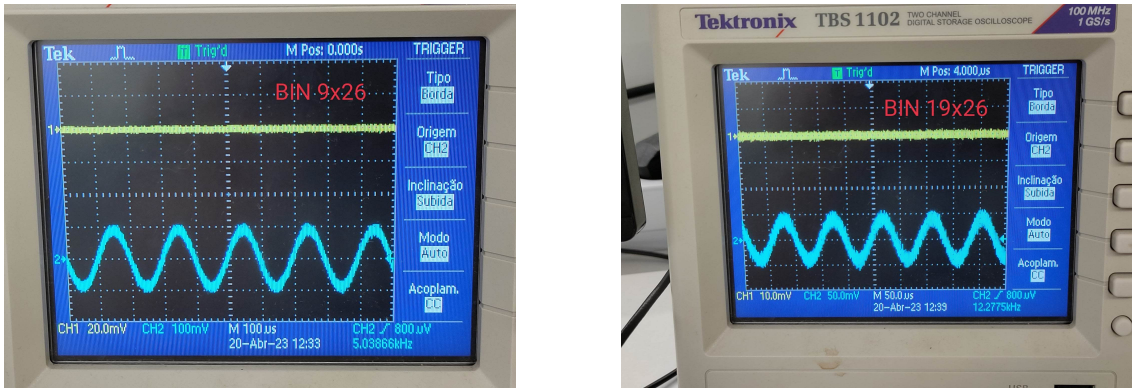
A validar o funcionamento correto do programa, fez-se um grupo de verificações que incluem: a verificação da boa formação de cada um dos harmónicos ao longo da gama disponível, limitada sobretudo pela frequência de amostragem; a síntese da onda dente-de-serra até o referido limite e, por fim, a garantia de que a adição da onda ao sinal de entrada é efetuado.

6.2.1 Estabilidade dos Harmónicos da banda disponível

Dois dos factores que limitam o número máximo de múltiplos da frequência fundamental que podem ser sobrepostos no sinal são: a frequência de amostragem de 22,05 kHz, o que implica o

corte à frequência de Nyquist em 11 kHz, e o comprimento do *buffer*, como, por se tratar de um sinal real, o índice dos harmónicos dado por $i * \ell$ não poderá ultrapassar o valor de 511, a fim de não alterar a segunda metade do vetor que terá que ser dada pelo simétrico conjugado da primeira.

Os testes foram conduzidos sobre uma F0 de 560 Hz, com $\ell = 26$, o que implica um limite de $\frac{FS}{2\ell} \approx 19$ harmónicos, colocando o último múltiplo no índice $i = 494$. As figuras seguintes ilustram a representação que foi obtida ao produzir os nono e décimo-nono harmónicos de forma isolada.



(a) $Bin = 9 * 26$ (b) $Bin = 19 * 26$

Figura 6.4: Visualização dos harmónicos 9 e 19

Apesar de se poder identificar alguma degradação da qualidade do sinal, os resultados confirmam que as sinusoides são formadas corretamente até ao limite calculado .

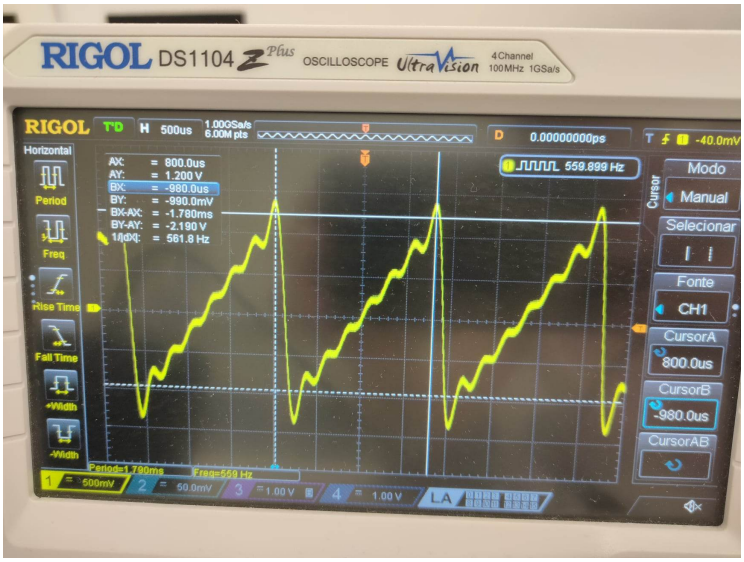
6.2.2 Produção da onda dente-de-serra e adição a sinal acústico

Tendo validado a síntese das sinusoides isoladas, verificou-se o comportamento da onda dente-de-serra resultante da inclusão simultânea das mesmas no espectro, pretendendo-se avaliar a qualidade da onda formada ao longo de toda a gama. Dos casos testados, destacam-se dois de particular interesse e que são apresentados na Figura 6.5. O primeiro diz respeito à onda composta por 8 harmónicos, limiar a partir do qual o som produzido passa a ter um carácter marcadamente artificial e desagradável à audição, e outro, o caso onde a banda é utilizada até ao limite calculado na subsecção anterior.

Nos dois casos é apreciável a disposição adequada dos harmónicos, não sendo também detetável qualquer artefacto sonoro ao auscultar o áudio produzido.

O último passo da validação consistiu em observar o efeito da adição da função dente-de-serra a um sinal exterior à placa. Esta etapa é de importância crítica dado que permite concluir acerca da possibilidade de fazer interagir sinais de naturezas distintas, neste caso, dar uma componente harmónica digital, gerada em frequência, a um sinal inicialmente analógico. As fotos apresentadas na Figura 6.5 já utilizam o método de adição ao *buffer* de entrada, mas foram conduzidas com um *input* vazio.

A validação passa então pela introdução de uma sinusoide de baixa frequência, neste caso, com uma frequência equivalente a um décimo da F0 da onda sintetizada, para obter uma boa



(a) 8 Harmónicos

(b) 19 Harmónicos

Figura 6.5: Visualização de ondas dente-de-serra compostas por 8 e 19 harmónicos

Figura 6.6: Adição de dente-de-serra de amplitude constante a sinusoide de baixa frequência

representação de 10 períodos da dente-de-serra para cada período da sinusoide. O resultado está ilustrado na Figura 6.6.

6.3 Resumo do capítulo

Neste capítulo foi apresentado um método de síntese de formas de onda no domínio das frequências com base na transformada ODFT e uma adaptação do algoritmo que o implementa em C. Verificou-se também a adequação do *kit* à plataforma de processamento digital e validou-se o funcionamento do mesmo. A síntese correta da forma de onda arbitrária e a sua incorporação a um sinal analógico captado em tempo-real são também indicativos da validade do método como ferramenta de vozeamento.

Capítulo 7

Modulação por Envolvente Espectral

O algoritmo desenvolvido até este ponto permite-nos atestar quanto à viabilidade da atribuição de uma estrutura harmónica sintetizada no domínio das frequências a um sinal de voz sussurrada, com base numa forma de onda periódica arbitrária, como método preliminar de vozeamento. Acresce que os resultados desde já obtidos apontam para a adequação da implementação aos requisitos funcionais do sistema, nomeadamente à operação do sistema em tempo real.

Neste capítulo, será apresentado um método de extração, filtragem e modulação da forma de onda periódica sintetizada em frequência pela *Power Spectral Density* (PSD) da transformada do sinal de sussurro. Deste modo, com uma magnitude da PSD filtrada, a estrutura harmónica usada para colorir o espectro da fala há-de apresentar um correlato direto com uma característica do sinal obtido em tempo-real. Esta camada de processamento contribui então para uma aproximação a um modelo de produção de fala, ocupando a forma de onda sintetizada o lugar da excitação glotal em falta, por definição, ao sussurro.

Para além do programa preparado para utilização no kit, são também expostas, e usadas no contexto de testes subjetivos, variantes do algoritmo análogo em Matlab, além de ser também avançada uma versão melhorada baseada na aplicação simplificada, e local, de técnicas desenvolvidas no contexto DyNaVoiceR.

7.1 Coloração Espectral do Sinal de Voz Sussurrada

Um dos maiores desafios que o vozeamento artificial de fala sussurrada enfrenta prende-se com a dificuldade que existe em extrair características espectrais do sussurro. O algoritmo, tal como foi apresentado até agora, opera o vozeamento preliminar com uma forma de onda arbitrária que é independente da atividade do sinal em análise. A etapa que então se introduz, consiste em utilizar a energia do sinal de fala sussurrada, traduzido na PSD, ou *Power ODFT*, do mesmo, como modulador da envolvente espectral da forma de onda sintetizada nas frequências. O sinal resultante será então colorido pela forma de onda periódica que é adicionada, dado que é a estrutura harmónica da forma de onda arbitrária que caracterizará o espectro desse.

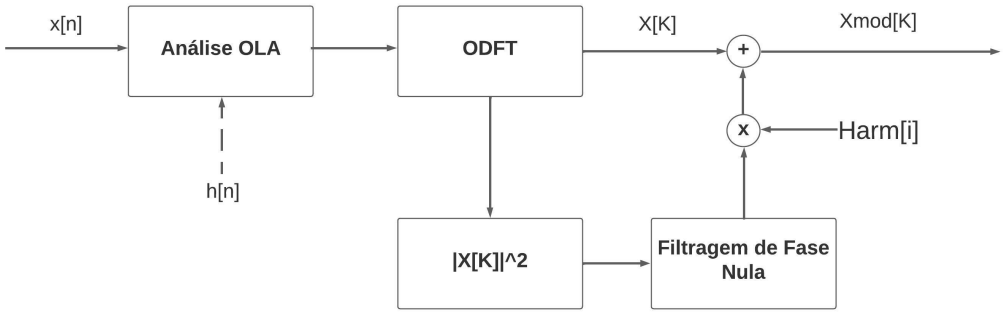


Figura 7.1: Diagrama de blocos do novo algoritmo

A Figura 7.1 ilustra o funcionamento do novo algoritmo, onde Xmod[] representa o sinal modulado em frequência, e Harm[] a sequência de harmónicos que constituem a onda sintetizada. A amplitude do sinal é dada pelo raiz da soma dos quadrados entre as partes real e imaginária do espectro do sinal, pois que se trata de um sinal complexo.

Este processo segue o método de extração de *features* que em [15] são empregues no processo de identificação de vogais, sendo que, nesse caso, a *Power* ODFT vê-se comparada, por meio de correlação, com modelos de vogais já definidos. Nesse trabalho, o intuito é o de identificar o modelo LPC que melhor aproxima a envolvente espectral da *frame* em análise. O algoritmo que apresentamos nesta secção não apresenta ainda a capacidade para aplicar um modelo de produção de fala, pelo que a envolvente é suavizada de modo a evitar um *fitting* excessivo ao espectro do sussurro. A filtragem é feita com recurso a um filtro de média, de comprimento 3, aplicado duas vezes em sequência, uma no sentido causal e outra no sentido anti-causal, compensando assim *delay* introduzido, tornando-o nulo (i.e trata-se de um filtro de fase nula). Considerando os requisitos da aplicação em tempo-real, ao invés da filtragem sequencial, computacionalmente mais exigente por necessitar uma inversão do sentido do sinal entre cada filtro, utilizou-se o filtro equivalente de resposta ao impulso triangular da Figura 7.2, que é dado pela convolução entre dois filtros de média de três coeficientes.

Como referido na introdução do capítulo, foram desenvolvidas também algumas variantes deste algoritmo. A primeira, e mais simples, consiste na substituição da forma da estrutura harmónica que foi até então utilizada, por uma onde todos os harmónicos têm igual amplitude. O cancelamento do termo de divisão, dado, lembremos, pelo índice do *bin* do harmónico, faz com que esta onda branqueada corresponda à primeira derivada da dente-de-serra original. Tanto esta como a versão colorida, como será descrito na próxima secção deste capítulo, foram utilizadas para vozear frases sussurradas, ao passo que a próxima técnica se prende exclusivamente ao vozeamento de vogais, extraindo, contudo, um resultado experimental importante que advém da investigação feita sobre o problema da melhoria da qualidade do áudio produzido segundo estes dois algoritmos.

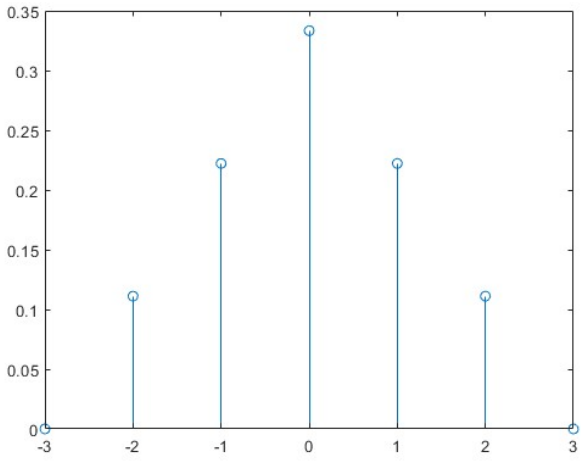


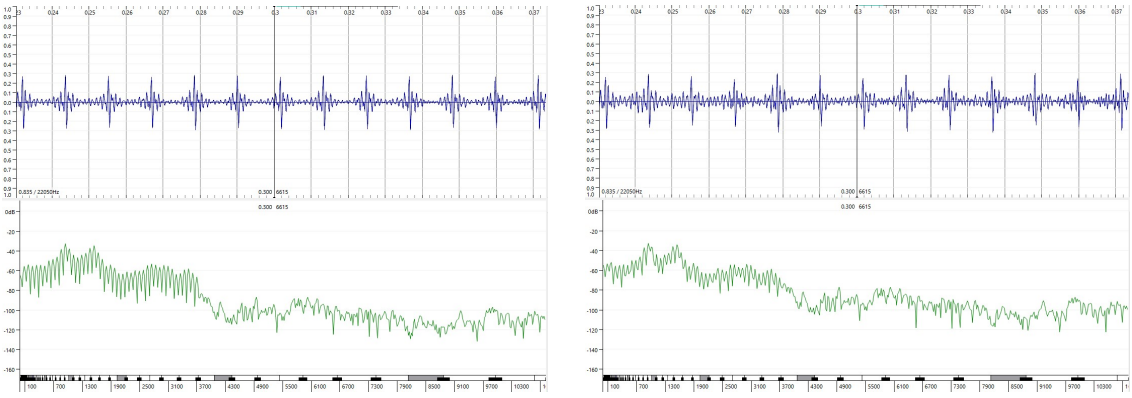
Figura 7.2: Coeficientes do filtro de fase zero da aplicação em tempo-real do algoritmo

7.2 Algoritmo de Vozeamento por Modelo Espectral Médio de Vogal

7.2.1 Regulação do peso do sinal de voz sussurrada

Durante o desenvolvimento dos algoritmos anteriores, conduziram-se alguns testes de vozeamento em ficheiros de voz gravados no contexto do projeto DyNaVoiceR (ver Secção 3.6). Constatou-se que, apesar da correta modulação do sinal, a voz produzida tinha uma qualidade desagradável e pouco regular. Este problema motivou a primeira alteração a nível do algoritmo de síntese em frequência, que foi o de adicionar um coeficiente regulador da amplitude do sinal de voz sussurrada.

Para os testes, isolou-se o áudio de um orador que pronuncia a vogal /a/, extraído de uma gravação maior que inclui as 5 vogais /a/, /e/, /i/, /o/, /u/ pronunciadas em sequência, sem intervalos de silêncio entre si.



(a) Peso do sinal de voz: 30% (b) Peso do sinal de voz: 90%

Figura 7.3: Representação temporal e espectral do sinal de vogal /a/ sintetizada

Produziram-se dez ficheiros de áudio sintetizado em que a onda dente-de-serra é espetralmente

colorida pela envolvente espectral de uma dada vogal e onde se fez variar a percentagem da amplitude do sinal de voz que era adicionada ao sinal final. Verificou-se que, apesar da proximidade que o espectro dos sinais produzidos exibem, a qualidade do áudio produzido para uma amplitude de sinal de sussurro abaixo dos 30% do valor original era significativamente mais agradável ao ouvido. Este resultado permite concluir acerca do impacto que a falta de regularidade e estacionaridade da produção de fala sussurrada tem na percepção do som.

7.2.2 Modelização de espectro médio de vogal

A solução proposta para o problema levantado pelas observações anteriores requer uma nova configuração do algoritmo até agora estudado, e que está ilustrada no diagrama de blocos da Figura 7.3.

O desenvolvimento deste algoritmo assenta em testes de vozeamento realizados sobre uma gravação das 5 vogais sussurradas continuamente, como foi descrito na subsecção anterior. Verificou-se que a falta de regularidade das características espectrais do sinal revela-se nas variações significativas que ocorrem a nível do espectro de diferentes janelas consecutivas correspondentes a uma mesma vogal, já que, trabalhando com segmentos de 1024 amostras, e considerando a sobreposição de 50% de segmentos adjacentes, o som de cada vogal surge repetido no comprimento de duas janelas. Uma variação excessivamente grande entre *frames* de análise implica uma falta de congruência durante a sobreposição de segmentos na fase de síntese, o que degrada a qualidade do som produzido.

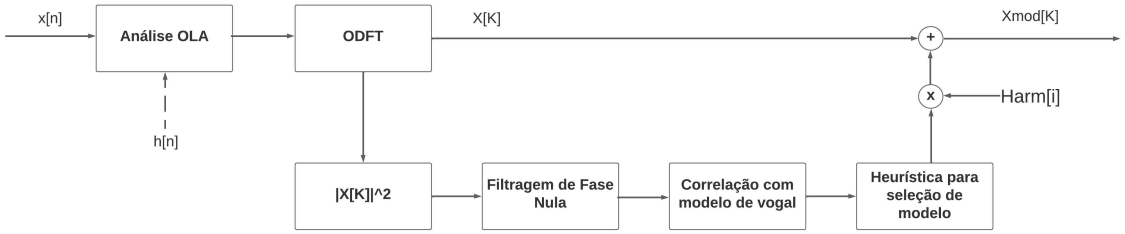
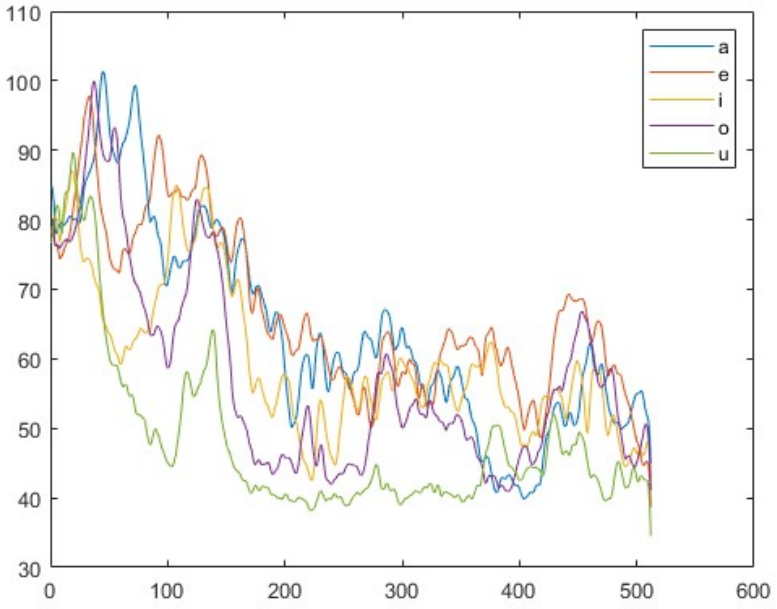


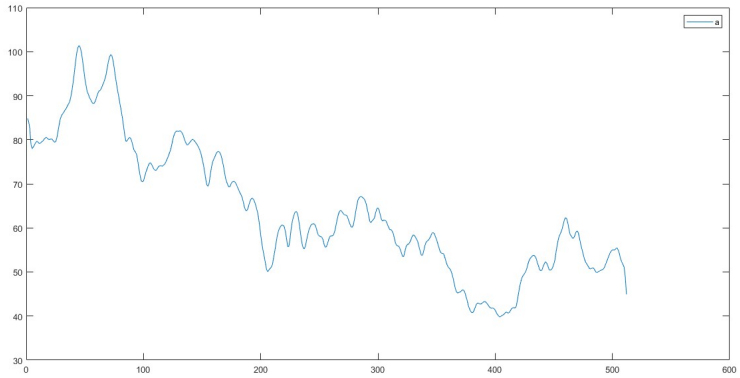
Figura 7.4: Diagrama de blocos do processamento em frequência do algoritmo modificado

Optou-se então por criar uma pequena biblioteca de modelos espectrais das vogais patentes na gravação, calculados a partir da média dos espectros das janelas que dizem respeito a cada uma das vogais. O novo algoritmo passa então por, ao invés de utilizar a *Power* ODFT da *frame* em análise como modulador espetral da onda sintetizada, por correlacionar essa mesma *Power* ODFT com os modelos pré-definidos, e eleger, segundo uma heurística que utiliza os coeficientes de correlação como termos de comparação, o modelo que deve ser utilizado na modulação.

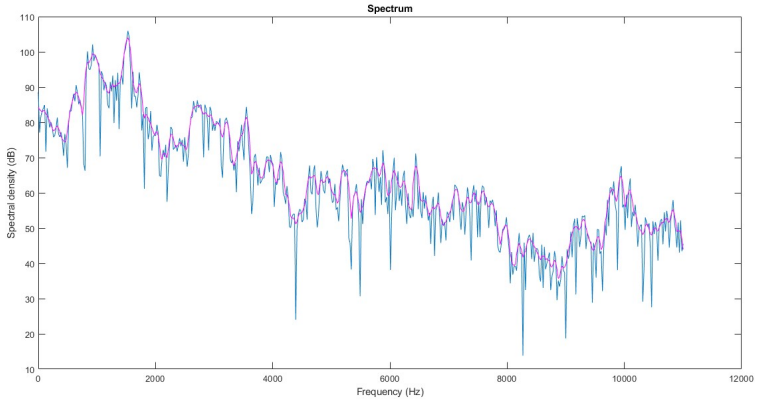
Segundo esta operação, o modelo escolhido é mantido ao longo da duração da vogal até que se verifique uma transição de fonema, o que aproxima melhor a regularidade que o aparelho fonador apresenta durante a produção da vogal sustentada, sendo a resposta do tracto e dos articuladores aproximadamente constante ao longo de todo o período.



(a) Modelos espectrais das cinco vogais



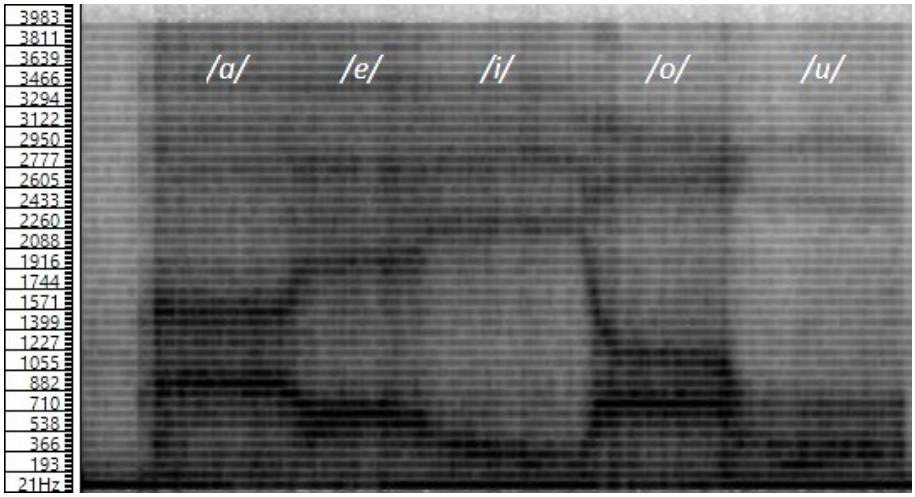
(b) Modelo espectral da vogal /a/



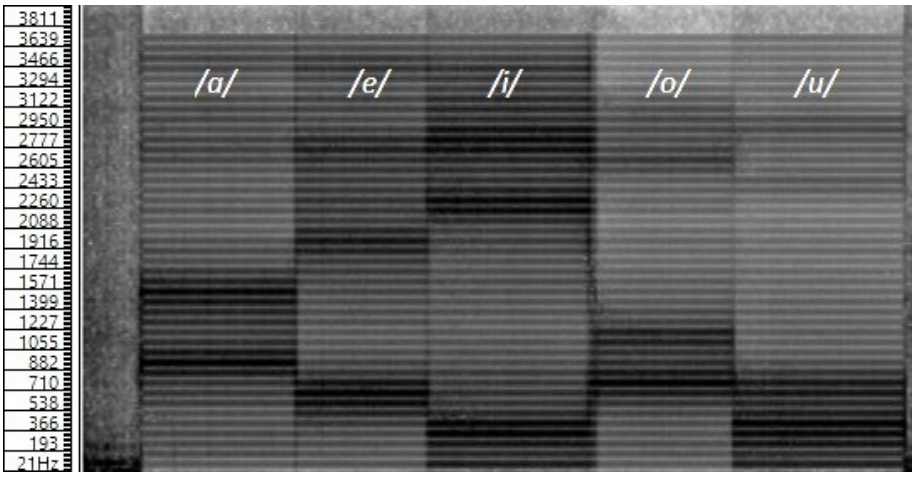
(c) Sobreposição do modelo de /a/ (curva suavizada) com a Power ODFT da frame

Figura 7.5: *Espectros dos modelos de vogal*

Comparando os espetros dos ficheiros de voz sintetizados segundo os algoritmos apresentados nas secções 7.1 e 7.2 sobressai a melhor definição das bandas de frequência. Nos testes aqui ilustrados, utilizou-se para coloração do espectro uma onda dente-de-serra com uma F0 de 86 Hz, dado que o orador é masculino, com um número de harmónicos cuja frequência limite é de 4 kHz, pois que é abaixo deste valor que se concentra a maior parte de energia do sinal e onde se localizam os harmónicos e formantes mais relevantes.



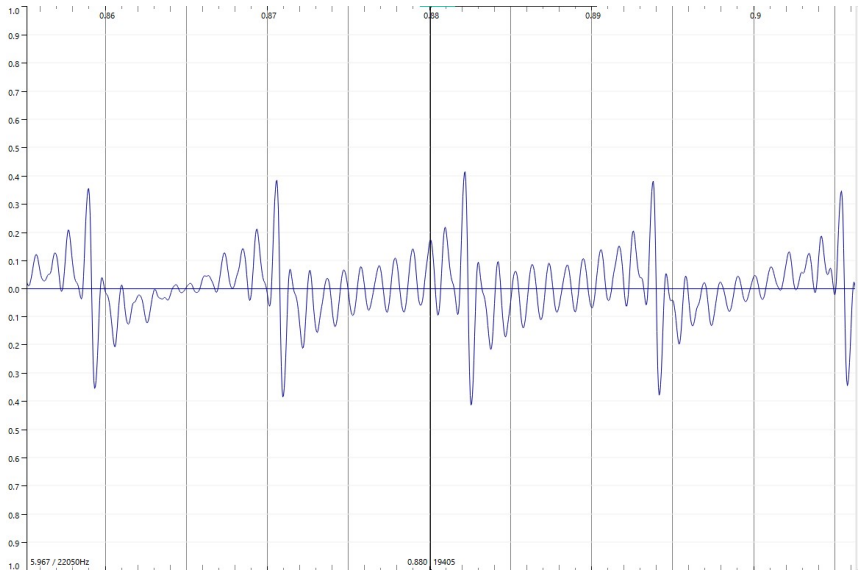
(a) Espectrograma resultante do primeiro método



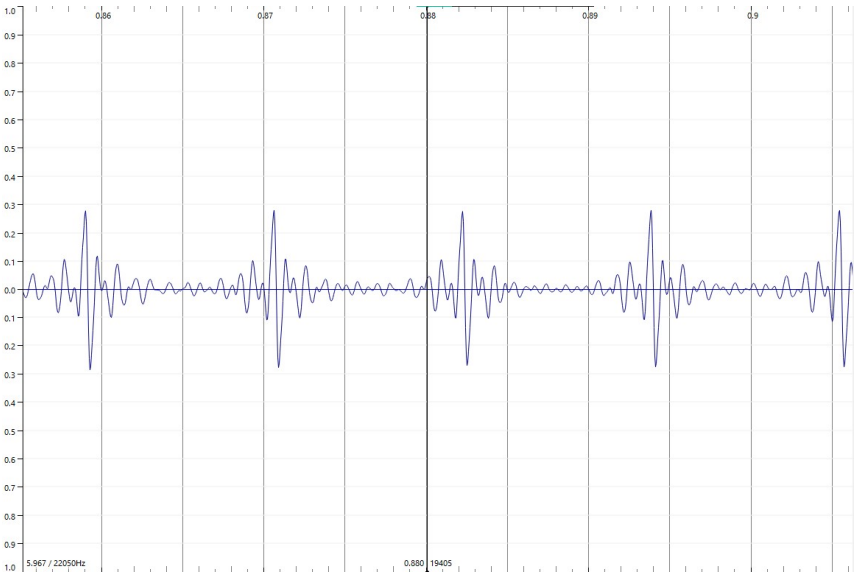
(b) Espectrograma resultante do novo método

Figura 7.6: Espectrogramas dos ficheiros de áudio sintetizados de vogais sussurradas

Uma observação destes mesmos sinais no domínio dos tempos revela ainda um aspecto importante que se prende ao perfil espectral da onda periódica usada para a coloração do espectro. No método melhorado é empregue uma onda com perfil espectral plano, querendo dizer que os harmónicos da onda partilham a mesma amplitude, para a coloração referida, e que apresentam um perfil análogo ao do ruído branco [5]. O método mais simples, como referido anteriormente, utiliza uma onda dente-de-serra, cujo o perfil é caracterizado pelo decrescimento monótono da amplitude do seu espectro.



(a) Representação temporal do sinal produzido segundo o primeiro método



(b) Representação temporal do sinal produzido segundo o novo método

Figura 7.7: Representação temporal dos ficheiro de áudio coloridos espectralmente com onda dente-de-serra e a sua derivada.

Nas capturas da Figura 7.7, é possível verificar como a forma temporal da onda utilizada na coloração da envolvente espectral do áudio transparece na envolvente temporal do sinal, destacando-se, da primeira, a forma da dente-de-serra, enquanto em 7.7b o perfil é plano.

Devido à falta de meios que permitissem obter uma modelização mais refinada do sistema de produção de voz em tempo-real, a implementação no *kit* ficou limitada ao primeiro método com uma modificação ligeira, passando-se então a utilizar a média entre a *Power* ODFT de duas janelas consecutivas, ao invés da que é extraída unicamente da *frame* da iteração vigente.

7.3 Testes Subjetivos

Nesta secção, serão apresentados os procedimentos, e resultados, dos testes subjetivos que foram realizados com vista a avaliar o desempenho dos algoritmos explorados neste capítulo no tocante à qualidade perceptual do áudio.

Foram elaborados dois tipos de desafio que foram aplicados a diferentes conjuntos de áudios:

- Testes de inteligibilidade- Pediu-se aos participantes que avaliassem, numa escala de valores de -2 a 2, com passo discreto de 0.5, a melhoria percebida da inteligibilidade de um par de ficheiros de áudio vozeados artificialmente quando comparados com um ficheiro de referência composto pela respetiva gravação de áudio sussurrado. Neste caso, a escala de classificação utilizada foi a seguinte:
  - 2: Melhoria significativa da inteligibilidade do áudio.
  - 1: Melhoria perceptível, mas pouco significativa.
  - 0: Não é percebida qualquer mudança.
  - -1: Perceptível degradação da inteligibilidade do áudio.
  - -2: Degradação severa da inteligibilidade.
- Testes de naturalidade: Nestes teste, opuseram-se duas versões de uma mesma gravação vozeada artificialmente segundo métodos distintos. Aos participantes indicou-se que fizessem uma avaliação da naturalidade relativa das duas versões segundo uma escala de +2 a 2 distribuída do seguinte modo:
  - 2: A versão A é muito mais natural que B.
  - 1: A versão A é mais natural que B.
  - 0: As diferenças entre as versões são irrelevantes.
  - -1: A versão B é mais natural que A.
  - -2: A versão B é muito mais natural que A.

No contexto destes testes, entende-se por inteligibilidade a distinção clara dos diferentes fonemas pronunciados por um orador, e, por naturalidade, a semelhança a uma voz produzida naturalmente por um ser humano.

As gravações de voz sussurrada utilizadas correspondem ao ficheiro de vogais mencionado na Secção 7.2.2 e às gravações de outros dois oradores, um feminino e outro masculino, a enunciar uma mesma frase. O primeiro ficheiro serviu de base a dois cenários de avaliação diferentes. No **primeiro cenário**, do ficheiro sussurrado produziram-se duas versões vozeadas artificialmente: uma com recurso a uma EL, modelo *Provox Trutone Emote*, e outra com o algoritmo de síntese em frequência com coloração espectral, como implementado no *kit*. No **segundo cenário**, mantêm-se o sinal de referência e a gravação da EL, rebatidos desta vez contra o ficheiro vozeado segundo o algoritmo que utiliza o modelo médio de vogais introduzido na última secção. Os outros dois ficheiros de voz **dão origem, cada um, a um cenário**, onde se contrapõem ao sinal sussurrado uma versão A, vozeada segundo o algoritmo de síntese em frequência com espectro colorido por onda dente-de-serra, e uma versão B, onde o mesmo algoritmo é utilizado, mas no qual o espectro é colorido pela derivada do sinal sintetizado anterior, ou seja, é colorido por um espectro branco onde os harmónicos da onda arbitrária têm todos a mesma amplitude. Para cada um destes 4 cenário fez-se um teste de inteligibilidade e um teste de naturalidade.

Contam-se 13 participantes nos testes percetivos, onde 8 se identificam com o género masculino e 5 com o feminino, sendo a idade média de 29 anos (máximo de 48 e mínimo 21).

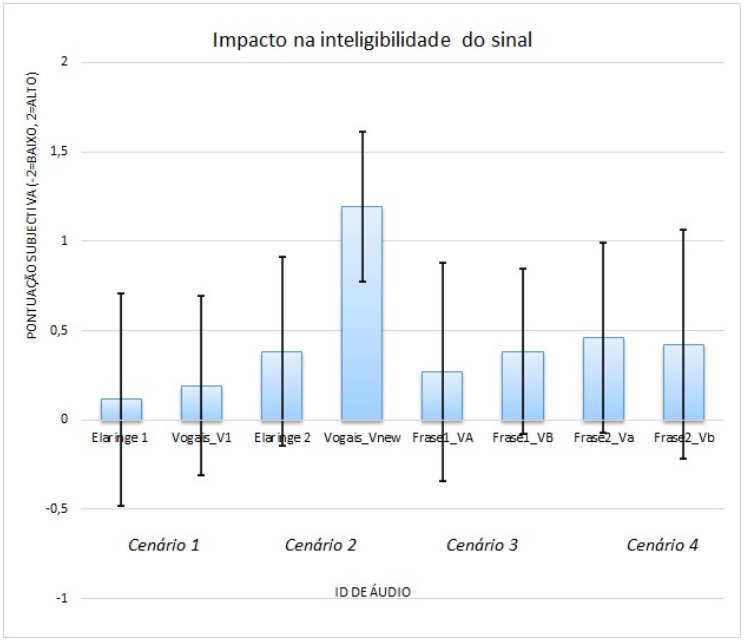


Figura 7.8: Média e intervalos de confiança dos testes subjetivos de inteligibilidade

Nas figuras 7.8 e 7.9 faz-se a representação gráfica das médias das pontuações atribuídas pelos participantes aos diferentes ficheiros e cenários nos dois contextos que foram preparados. Essa informação, traduzida numericamente nas médias e intervalos de confiança de 95%, encontram-se nas Tabelas 7.1 e 7.2.

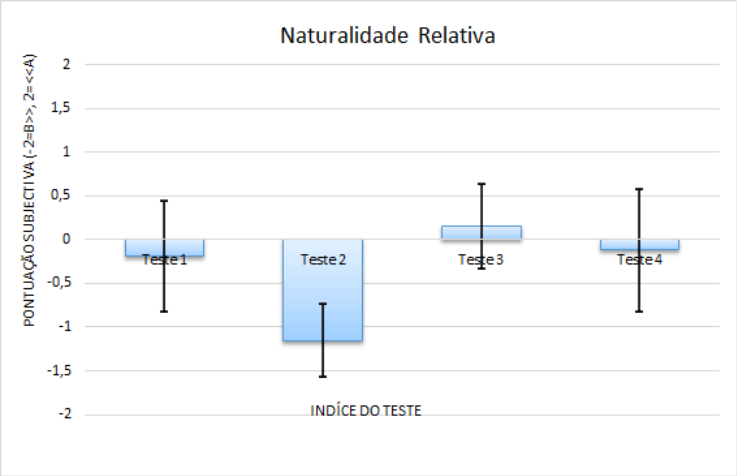


Figura 7.9: Média e intervalos de confiança dos testes subjetivos de naturalidade

ID de ficheiro	Média	Intervalo de Confiança (±)
Elaringe 1	0,115	0,594
Vogais_V1	0,192	0,502
Elaringe 2	0,385	0,526
Vogais_Vnew	1,192	0,419
Frase1_VA	0,269	0,612
Frase1_VB	0,385	0,464
Frase2_Va	0,462	0,530
Frase2_Vb	0,423	0,639

Tabela 7.1: Média e intervalos de confiança para os testes de inteligibilidade

ID de Teste	Média	Intervalo de Confiança (±)
Cenário 1	-0,192	0,635
Cenário 2	-1,154	0,416
Cenário 3	0,154	0,484
Cenário 4	-0,115	0,699

Tabela 7.2: Média e intervalos de confiança para os testes de naturalidade

7.4 Discussão dos Resultados

Analisando os resultados tabelados na secção anterior torna-se evidente que existe uma grande dispersão ao longo dos dados, o que se reflete na dilatação dos intervalos de confiança. Esta variação grande pode, em parte, estar relacionada com a dificuldade em estabelecer, através da descrição dos objetivos do problema, um referencial subjetivo forte o suficiente para garantir uma maior coerência nos dados.

Contudo, comprova-se que os métodos mais simples de vozeamento por coloração espectral tendem, face aos resultados pouco definidos dos testes da EL, a ter valores positivos. Mais estatisticamente relevante é a evidência da qualidade do algoritmo melhorado face a aos restantes mé-

todos, tanto na melhoria obtida em termos de inteligibilidade como em questões de naturalidade, sendo que no cenário 2, onde este ficheiro é implicado como versão modificada B, a pontuação média é de, aproximadamente, -1.15 e, em termos de inteligibilidade, tem a média mais alta do conjunto, de 1.19. Este é um indício forte do peso que a regularidade do sistema produção de fala tem na percepção do objeto sonoro produzido.

7.5 Resumo do capítulo

Neste capítulo foram expostos dois métodos de vozeamento baseados na modulação de um sinal periódico sintetizado em frequência, dado por uma forma de onda arbitrária, pela envolvente espectral do sinal de voz sussurrada.

O processo de desenvolvimento do algoritmo permitiu extrair dois dados importantes no tocante ao peso que o carácter ruidoso do sinal de sussurro tem na reconstituição de voz:

1. Que o balanço entre a amplitude do sinal de sussurro original e a componente periódica sintetizada deve ser ajustado, reduzindo o contributo da onda original, para que o sinal resultante tenha um carácter mais melódico e natural.

2. Que o regime não-estacionário sob o qual é produzida a voz sussurrada interfere na análise e reconstrução do sinal, o que leva à degradação dos formantes pela falta de coerência na sobreposição (neste caso, de 50%) entre *frames* consecutivas.

Apesar de contarem com uma dispersão pouco favorável, os resultados dos testes subjetivos permitem concluir acerca da importância destes pontos e da adequação dos métodos adotados, de balanço de mistura entre os sinais nas frequências, e de modelização espectral de vogais, para os responder.

Capítulo 8

Conclusões e Trabalho Futuro

As conclusões a extrair deste trabalho de investigação podem ser divididas entre o sucesso do desempenho do plataforma de processamento adotada, e os dados experimentais que decorrem da investigação realizada a nível da algoritmia num contexto mais abrangente.

- Verificou-se que, dentro da estrutura adotada para a análise e processamento, o *kit* foi capaz de cumprir os requisitos funcionais propostos, tendo-se verificado experimentalmente que o atraso introduzido estava dentro do intervalo estipulado para uma fusão entre os sinais visuais e sinais de áudio, e que a síntese digital, e atribuição de uma estrutura harmónica, nas frequências, a um sinal acústico captado em tempo-real era viável. Contudo, a pouca flexibilidade para a incorporação de modelos pré-calculados é limitante no sentido de se obter uma caracterização mais refinada do perfil espectral do áudio produzido.

- A investigação que resultou no desenvolvimento da variante final do algoritmo de vozeamento por modelo de vogal, que pode ser vista como uma aplicação ilustrativa, mas muito simplificada, do método já utilizado no DyNaVoicer para a classificação de vogais, ajuda a compreender como os pequenos artefactos, introduzidos pelo perfil ruidoso do sinal de voz, podem ter um impacto significativo na caracterização de um som de voz devido à alta sensibilidade do sistema auditivo humano à regularidade de padrões sonoros. Nesse sentido, o algoritmo criado para abordar diretamente o problema, e os resultados dos testes subjetivos, sucedem em aplacar estes efeitos e favorecem, consequentemente, a hipótese colocada.

Deste modo, o trabalho futuro passa, em primeiro lugar, por um aprimoramento de técnicas para a adaptação de módulos do DyNaVoiceR que consigam incorporar modelos de produção de voz no *kit*, e controlar o padrão de evolução de F0, e, em seguida, uma aplicação mais abrangente da regulação do balanço entre a amplitude do sinal ruidoso e estruturas harmónicas re-sintetizadas, por exemplo, no contexto de plataformas de processamento anteriores do DyNaVoiceR.

Anexo A

Verificação da implementação da estrutura Análise/Síntese

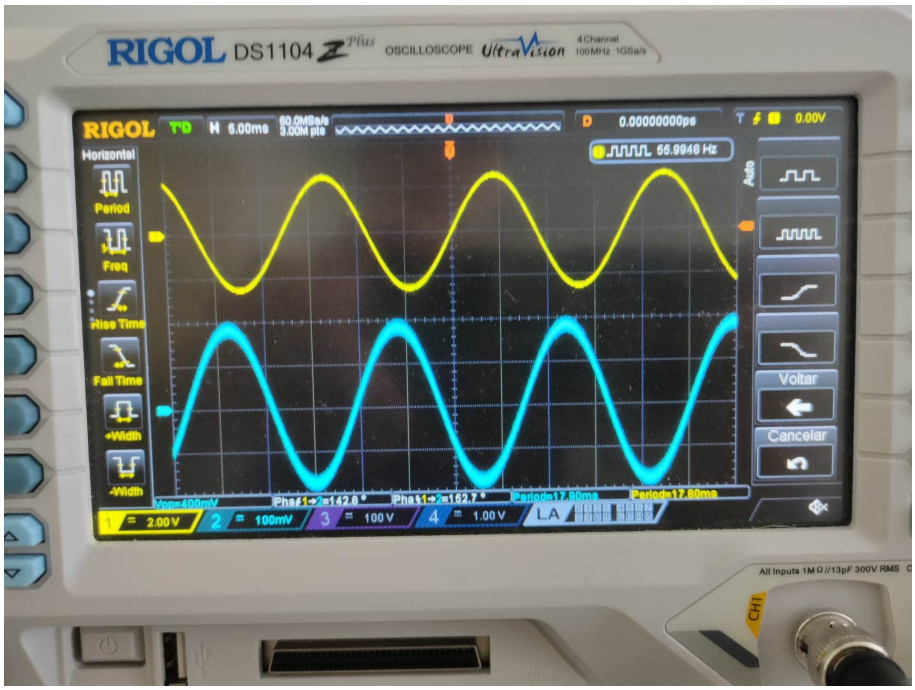


Figura A.1: Verificação de onda sinusoidal (traçado inferior)

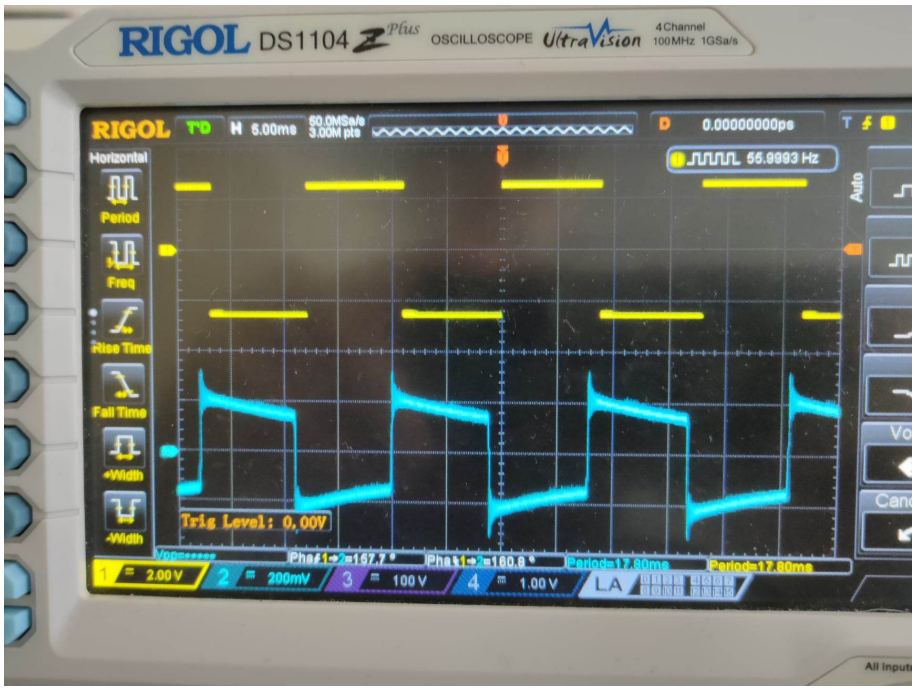


Figura A.2: Verificação de onda quadrada (traçado inferior)

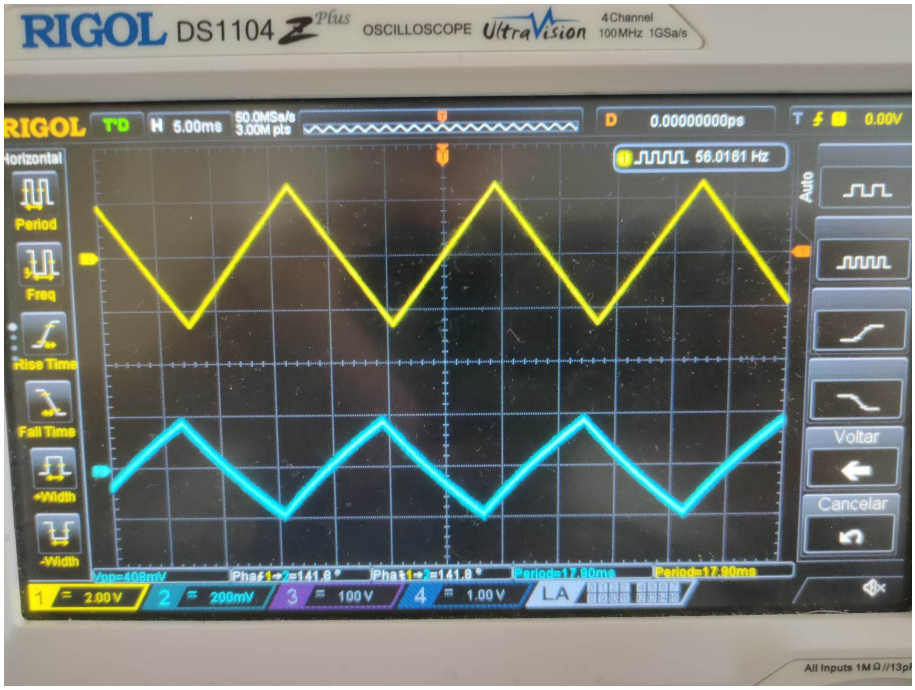


Figura A.3: Verificação de onda triangular (traçado inferior)

Figura A.4: Verificação de onda dente-de-serra (traçado inferior)

Referências

[1] Diamantino Freitas e Vitor Pêra. Audição e produção de fala, 2010.

[2] Denis Butler Fry. *The Physics of Speech*. Cambridge University Press, 1979.

[3] Aníbal Ferreira, Fernando Pereira, Carlos Salema, Sérgio Faria, Pedro Assunção, Isabel Trancoso, e Paulo Correia. *Comunicações audiovisuais : tecnologias, normas e aplicações*. IST Press, 2009.

[4] A.J.S. Ferreira. Accurate estimation in the odft domain of the frequency, phase and magnitude of stationary sinusoids. Em *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, páginas 47–50, 2001. doi:10.1109/ASPAA.2001.969539.

[5] A.P. Mendes, S. Ibrahim, e I. Vaz. *A voz no fado: vocologia*. Prime Books, 2021. URL: https://books.google.pt/books?id=48rCzgEACAAJ.

[6] Ewelina Sielska-Badurek e Ewa Osuch-Wójcikiewicz. Combined functional voice therapy in singers with muscle tension dysphonia in singing. *Journal of Voice*, 31:509.e23–509.e31, 7 2017. doi:10.1016/j.jvoice.2016.10.026.

[7] Gerald L. Culton e John M. Gerwin. Current trends in laryngectomy rehabilitation: A survey of speech-language pathologists. *Otolaryngology–Head and Neck Surgery*, 118:458–463, 4 1998. doi:10.1177/019459989811800405.

[8] Susanne Singer, Dorit Wollbrück, Andreas Dietz, Juliane Schock, Friedemann Pabst, Hans Joachim Vogel, Jens Oeken, Annett Sandner, Sven Koscielny, Karl Hormes, Kerstin Breitenstein, Heike Richter, Andreas Deckelmann, Sarah Cook, Michael Fuchs, e Sylvia Meuret. Speech rehabilitation during the first year after total laryngectomy. *Head and Neck*, 35:1583–1590, 11 2013. doi:10.1002/hed.23183.

[9] Dynavoicer, tasks: Quick overview. https://paginas.fe.up.pt/~voicestudies/dynavoicer/?page_id=341. Accessed: 2023-01-04.

[10] Robert W. Morris e Mark A. Clements. Reconstruction of speech from whispers. *Medical Engineering Physics*, 24(7):515–520, 2002. Models Analysis of Vocal Emissions. URL: https://www.sciencedirect.com/science/article/pii/S1350453302000607, doi:https://doi.org/10.1016/S1350-4533(02)00060-7.

[11] P.B. Denes, P. Denes, e E. Pinson. *The Speech Chain*. 1998.

[12] Ian Vince McLoughlin. *Speech and Audio Processing*. Cambridge University Press, 2016.

[13] Xuedong Huang, Alex Acero, e Hsiao-Wuen Hon. *Spoken Language Processing*. Prentice-Hall, 2001.

[14] Aníbal J. S. Ferreira. Static features in real-time recognition of isolated vowels at high pitch. *Journal of the Acoustical Society of America*, 122(4):2389 – 2404, 2007. doi:10.1121/1.2772228.

[15] Marco Oliveira. Modelização de filtro de trato vocal para reconstrução de voz disfónica. Tese de mestrado, 2020.

[16] Wai Chu. *Speech Coding Algorithms: Foundation and Evolution of Standardized Codecs*. 04 2003. doi:10.1002/0471668850.

[17] Hideaki Konno, Mineichi Kudo, Hideyuki Imai, e Masanori Sugimoto. Whisper to normal speech conversion using pitch estimated from spectrum. *Speech Communication*, 83:10–20, 2016. URL: https://www.sciencedirect.com/science/article/pii/S016763931530090X, doi:https://doi.org/10.1016/j.specom.2016.07.001.

[18] Hamid Reza Sharifzadeh, Ian V. McLoughlin, e Farzaneh Ahmadi. Reconstruction of normal sounding speech for laryngectomy patients through a modified celp codec. *IEEE Transactions on Biomedical Engineering*, 57(10):2448–2458, 2010. doi:10.1109/TBME.2010.2053369.

[19] Aníbal Ferreira. Implantation of voicing on whispered speech using frequency-domain parametric modelling of source and filter information. Em *2016 International Symposium on Signal, Image, Video and Communications (ISIVC)*, páginas 159–166, 2016. doi:10.1109/ISIVC.2016.7893980.

[20] Patrícia Cristina Ramalho de Oliveira. Artificial voicing of whispered speech. Tese de mestrado, 2015.

[21] Aníbal Ferreira. Dysphonic to natural voice reconstruction, first-year report (2018-2019). 8 2019.

[22] Aníbal Ferreira. First experiments on speaker identification combining a new shift-invariant phase-related feature (nrd), mfccs and f0 information. Em *ICETE (1)*, páginas 513–524, 2018.

[23] Ruth Y Litovsky, H Steven Colburn, William A Yost, e Sandra J Guzman. The precedence effect. *The Journal of the Acoustical Society of America*, 106(4):1633–1654, 1999.

[24] Quentin Summerfield. Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 335(1273):71–78, 1992.

[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, 2005.*, páginas 203–206, 2005. doi:10.1109/ASPAA.2005.1540205.

[26] John Princen e Alan Bradley. Analysis/synthesis filter bank design based on time domain aliasing cancellation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(5):1153–1161, 1986.

[27] R. Rowlands. The odd discrete fourier transform. Em *ICASSP '76. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, páginas 130–133, 1976. doi:10.1109/ICASSP.1976.1170122.

[28] Aníbal Ferreira. *Spectral Coding and Post-Processing of High Quality Audio*. Tese de doutoramento, 11 1998.

[29] Complex fft functions. URL: https://arm-software.github.io/CMSIS_5/DSP/html/group__ComplexFFT.html#gade0f9c4ff157b6b9c72a1eafd86ebf80.

[30] Aníbal Ferreira, João Silva, Francisca Brito, e Deepen Sinha. Impact of a shift-invariant harmonic phase model in fully parametric harmonic voice representation and time/frequency synthesis. Em *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, páginas 701–705, 2020. doi:10.1109/ICASSP40776.2020.9054496.

[31] Gunnar Fant. Glottal flow: models and interaction. *Journal of Phonetics*, 14(3-4):393–399, 1986.