INSTITUTO POLITÉCNICO DE BRAGANÇA
ESOLA SUPERIOR DE TECNOLOGIA E GESTÃO

ENGENHARIA INDUSTRIAL RAMO ENGENHARIA ELECTROTÉCNICA

SINTETIZADOR DE FALA DIDÁTICO – MÓDULO ACÚSTICO
MODELO DE FORMANTES

ANILDO PEREIRA FERNANDES

TESE DE MESTRADO

Bragança
Outubro 2012

INSTITUTO POLITÉCNICO DE BRAGANÇA
ESOLA SUPERIOR DE TECNOLOGIA E GESTÃO

SINTETIZADOR DE FALA DIDÁTICO – MÓDULO ACÚSTICO
MODELO DE FORMANTES

Tese realizada sob a orientação do Professor Doutor
João Paulo Ramos Teixeira

Bragança
Outubro 2012

Dedico este trabalho à minha mãe e aos meus irmãos, que tanto contribuíram para a minha formação.

AGRADECIMENTOS

Dirijo ao meu orientador, Professor João Paulo Teixeira, um sincero agradecimento por acreditar e confiar, pelo apoio incansável, pela capacidade de ensinar a pensar, e pela forma amigável com que sempre dispôs a discutir os assuntos relacionados com esta tese.

À minha família, em particular, à minha mãe (Vitalina Pereira Matos Pires) e aos meus irmãos (Vital P. Fernandes, José M. P. Fernandes e Sandra P. Fernandes), pela paciência, compreensão, pelo incentivo e apoio incondicional que sempre me deram até atingir esta fase.

Aos meus amigos, o meu muito obrigado, pela força e pela ajuda que me forneceram, e sobretudo pela amizade.

Por último mas não menos importante, um agradecimento muito especial para a Nicole Fidalga pela ajuda, acompanhamento e contribuição ao longo desta etapa.

RESUMO

Dada a importância da interface homem-máquina para a fala, a função de conversor de texto-fala merece a atenção da engenharia, em particular da eléctrica e de tecnologia de computador. Neste trabalho foram desenvolvidos dois programas, em ambiente Matlab, para a síntese computacional da fala, baseado na excitação glotal artificial e na utilização de filtros que simulam o trato vocal do homem. O primeiro programa consiste numa aplicação didática sobre o modelo de formantes em que que é possível fazer a síntese das vogais em português e também é possível alterar os parâmetros do modelo e ouvir o respetivo som. Na segunda aplicação desenvolveu-se o módulo acústico de um sintetizador de fala em ambiente modular. Foi utilizado o programa Praat para gravar, os sinais de fala e extrair os parâmetros, como as frequências formantes e larguras de banda necessárias para a base de dados de difones. Os resultados são apresentados numa interface gráfica, do Matlab, conhecida por Guide. Na presente tese, é apresentada a constituição de um sistema de conversão texto-fala para o Português, explicando as funções dos vários módulos e descrevendo as técnicas de desenvolvimento utilizando o modelo de formantes.

Palavras-chave: Análise de Fala, Síntese de Fala, Conversão Texto-Fala, Processamento de Sinal, Modelo de Formantes.

ABSTRACT

Given the importance of the human-machine interface for speech, the function of text-to-speech converter deserves the attention of engineering, particularly electrical and computer technology. In this work we developed two programs in Matlab for computational synthesis of speech based on glottal excitation and the use of filters that simulate the human vocal tract. The first program consist in a didactic application about the formants model in which you can do the synthesis of vowels in Portuguese and you can also change the model parameters and hear the respective sound. In the second application was developed the acoustic module of a speech synthesizer in modular environment. We used the Praat program to record speech signals and extracting the parameters such as formant frequencies and bandwidths necessary for the database difones. The results are presented in a graphical interface, in Matlab, known as Guide. In this thesis, we present the constitution of a system text-to-speech conversion for the Portuguese, explaining the functions of the various modules and describing the development techniques using the formants model.

Keywords: Speech Analysis, Speech Synthesis, Text-to-speech Conversion, Signal Processing, Formants Model.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SÍMBOLOS

B1	Primeira largura de banda
B2	Segunda largura de banda
B3	Terceira largura de banda
B4	Quarta largura de banda
Fa	Frequência de amostragem
F0	Frequência fundamental
F1	Primeira frequência formante
F2	Segunda frequência formante
F3	Terceira frequência formante
F4	Quarta frequência formante
G(z)	Função de transferência do impulso glotal
Hz	Hertz
IPA	International Phonetic Alphabet
ms	milisegundos
R(z)	Função de transferência para o efeito de radiação nos lábios
V(z)	Função de transferência do trato vocal
μi	Média do segmento i
σi	Desvio padrão do segmento i

CAPÍTULO 1
INTRODUÇÃO

1 INTRODUÇÃO

1.1 Objectivos e Enquadramento

A síntese de fala é o processo da produção artificial da fala humana, e o seu estudo acompanha a civilização desde há muito tempo. Os primeiros sistemas de produção de fala artificial apareceram no século XVIII (Sproat, 1997). Eram mecânicos, difíceis de operar e não geravam mais do que alguns poucos sons da fala. No entanto, serviram como ferramentas de experimentação para o estudo do mecanismo de produção da fala. Com o desenvolvimento da tecnologia, sistemas electrónicos, a geração da fala a partir de um texto, foi conseguido na década de 1960 (Pacheco, 2010).

A síntese de fala é uma área do processamento digital de fala cujo objectivo básico consiste na geração de voz artificial através de dispositivos conhecidos como sintetizadores. Tais sintetizadores podem ser eléctricos ou digitais e, quando digitais, exigem parâmetros matemáticos de entrada para a geração da voz desejada na saída.

Um sistema de computador utilizado para uma síntese de fala é chamado de um sintetizador de fala, e pode ser implementado em software ou hardware (Saraswathi, 2010). Um sintetizador de fala permite gerar automaticamente o discurso de uma representação simbólica que é fornecido como entrada (Barros, 2002; Thomas, 2007).

A expansão dos sistemas de síntese de fala é cada vez maior e tem uma aplicação

enorme em várias áreas, inclusive nas áreas de Telecomunicações e Multimédia (Klompje, 2006).

Hoje, pode-se notar a extrema importância de um sistema de síntese de fala, pois são cada vez mais comuns as aplicações do nosso quotidiano que integram este sistema, também designados por sistemas de conversão texto-fala, com o objectivo de facilitar o acesso à informação de um número cada vez maior de utilizadores. De entre as principais aplicações, contam-se os sistemas de auxílio à leitura para cegos, os sistemas de auxílio à navegação por GPS, as aplicações para telemóveis usando sistemas de pergunta-resposta, as aplicações industriais do tipo de máquinas com comandos mãos-livres, as aplicações médicas de monitorização de doentes, o software de ensino de línguas e de e-learning em geral e as aplicações que facilitem a acessibilidade na internet (leitura de e-mail, leitura de conteúdos de páginas web, etc.), entre muitas outras (Braga; Mato, 2007).

Neste trabalho desenvolveram-se duas aplicações. A primeira consiste num sintetizador de formantes paramétrico que permite realizar a síntese de vogais recorrendo ao triângulo de vogais, bem como modificar o valor dos parâmetros e realizar a síntese. A segunda aplicação consiste no módulo acústico de um sintetizador de fala em português, baseado no modelo de formantes. O trabalho é realizado inteiramente no software Matlab, apresentando os resultados numa interface gráfica conhecida por Guide.

A segunda aplicação converte uma sequência de fonemas em código SAMPA (acrescentar referência) em som, usando uma base de dados de difones parametrizados em formantes e larguras de banda.

A ideia é fazer uma interacção entre o utilizador e um programa através de uma

janela.

Foi utilizado o programa Praat para extrair os parâmetros dos formantes e respetivas larguras de banda dos difones.

Foram feitas várias experiências para diferentes vogais e palavras, alterando ligeiramente os valores dos parâmetros.

1.2 Organização da Tese

De seguida, apresenta-se a estrutura da Tese, dando ordem aos temas desenvolvidos, permitindo o conhecimento dos assuntos tratados em cada capítulo.

O primeiro capítulo é referente à introdução, onde o trabalho é apresentado de forma geral ao leitor, procurando dar a conhecer o assunto abordado.

O segundo capítulo aborda o estudo de um sistema de conversão texto-fala, conhecido em inglês como Text-To-Speech (representado por TTS), onde é feita a representação deste sistema, explicando as funções de cada bloco constituinte. Ainda neste capítulo é tratado o estudo do arquétipo utilizado para a realização deste sintetizador de fala, que é o modelo de formantes.

No terceiro capítulo é desenvolvido o sintetizador de formantes, explicando o seu objectivo (sintetizar as vogais) e o seu funcionamento por partes, onde se discutem as funções de transferência que simulam o trato vocal V(z), o efeito de radiação nos lábios R(z) e a fonte de excitação glotal G(z). Apresenta-se ainda neste capítulo o funcionamento do sintetizador numa interface gráfica.

Um sistema de conversão de fonemas para fala é apresentado no capítulo quatro, explicando o seu objectivo e o seu desenvolvimento, apresentando as diferenças das funções do sintetizador neste capítulo e no capítulo anterior.

O quinto e último capítulo encerra a presente Tese. Apresentam-se as principais conclusões deste estudo, bem como as suas limitações, apontam-se algumas sugestões com vista a trabalho futuro.

CAPÍTULO 2
SISTEMA DE CONVERSÃO TEXTO-FALA

2 SISTEMA DE CONVERSÃO TEXTO-FALA

Os sistemas de conversão texto-fala (TTS) são, hoje, muito divulgados na comunidade científica. Destes sistemas espera-se uma conversão de um texto escrito de forma eletrónica numa fala sintetizada. Este tipo de sistemas procura, de algum modo, imitar a fala do homem a partir de um texto. A maior vantagem desta técnica será, porventura, a sua flexibilidade, que se traduz na capacidade de gerar um número ilimitado de frases, aliada a um custo de armazenamento relativamente baixo (Oliveira, 2009).

O processo de conversão de texto-fala é bastante complexo e só pode ser resolvido de forma multidisciplinar. De cordo com os estudos, para a resolução deste problema é necessária uma divisão praticamente natural do processo em duas etapas:

- A passagem do domínio texto para um domínio de representação intermediário, baseada em técnicas de processamento de linguagem natural;
- E a passagem do domínio intermediário para o domínio acústico do sinal de fala, baseada em técnicas de processamento de sinais.

Uma primeira e enorme motivação para o estudo e desenvolvimento destes sistemas é a ajuda preciosa que podem dar a pessoas com alguns tipos de incapacidades.

As principais maneiras que uma pessoa cega tem para comunicar com um computador são o terminal "braille", com custos muitas vezes inacessíveis, e o conversor texto-fala que lhe pode proporcionar a saída de som necessário para dar uma "imagem auditiva" do que acontece no écran do computador. Isto permite-lhe ter acesso a livros, mensagens e jornais em forma eletrónica nomeadamente os que estejam disponíveis na rede pública de dados (Teixeira *et al*, 1998). Por outro lado, permite a evolução das suas capacidades profissionais na escrita e desenvolvimento de programação. Para pessoas com deficiência temporária ou definitiva ao nível da fala, permite que o sintetizador de fala associado a um PC se façam ouvir com recurso a um programa específico de onde podem selecionar e compor rapidamente e com facilidade um grande número de mensagens pré-gravadas ou escritas no momento, podendo assim estabelecer comunicação ou telecomunicação (Teixeira *et al*, 1998).

Hoje, há um enorme e crescente interesse destes sistemas nos diversos campos de aplicação, como sejam: serviço de atendimento a clientes em bancos e estações de informação, dicionários multilíngua para turistas, correio de voz, instruções em simuladores, telecomunicações, etc..

Um sistema TTS é constituído por dois módulos claramente distintos, o processamento linguístico-prosódico e o processamento acústico como se representa na figura 1. Na primeira etapa, em que estão relacionados fundamentalmente aspetos linguísticos, o texto é analisado, sendo gerada uma representação fonética associada a informações prosódicas da fala que será sintetizada. Esse estágio de processamento é fortemente dependente do idioma a que se propõe o sistema de conversão. Ao final do processamento linguístico, os sons que devem ser sintetizados estão definidos. A síntese propriamente dita do sinal de fala é realizada no bloco de processamento acústico. Um

modelo de síntese (modelo de formantes, neste trabalho) deve permitir a geração dos sons e a alteração dos parâmetros prosódicos de acordo como que foi prescrito na etapa de análise linguística.

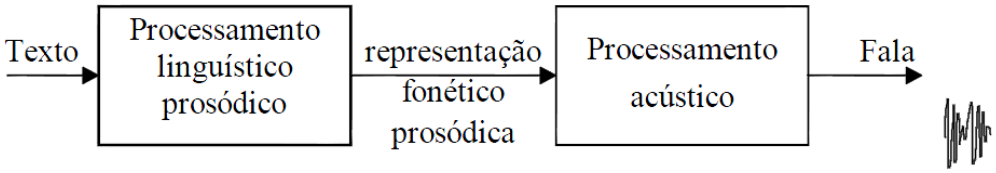

Figura 1 - Diagrama de blocos genérico de um sistema de conversão texto-fala (retirada de (Teixeira *et al*, 1998)).

Diversos autores possuem denominações diferentes para os dois blocos básicos apresentados na figura 1, mas a funcionalidade descrita para cada um deles é praticamente a mesma nas diversas referências. Para uma melhor compreensão deste processo, para cada bloco será feita uma breve descrição.

2.1 Processamento linguístico prosódico

Num sistema de conversão texto-fala, na primeira fase é feita a análise do texto de entrada, a ideia é transformar o texto em uma representação simbólica e estruturada que indique os sons que devem ser sintetizados com seus parâmetros prosódicos associados. A análise de texto é fortemente dependente do idioma a que se propõe o sistema de conversão e é subdividida em módulos.

O objectivo do presente bloco consiste na determinação, a partir de um texto, de dois tipos de informações necessárias para fornecer ao processamento acústicos dados de maneira a poder gerar fala. Estas informações são conhecidas como informações ao nível segmental e informações ao nível supra-segmental (Teixeira *et al*, 2003).

A informação segmental está ligada aos sons básicos que compõem a mensagem.

Cada língua tem um conjunto limitado de sons básicos que permitem produzir fala, quando devidamente combinados, todas as particularidades do discurso nesse idioma. Portanto, ele cria uma série de representações abstratas chamados de fonemas cuja dimensão depende do idioma em questão (Teixeira *et al*, 2003).

A informação supra-segmental está associada com a prosódia. Reflecte os elementos linguísticos (como os tipos de frase, pausas, acentuação e agrupamento de elementos de significado), como elementos não linguísticos. Esta informação é considerada por muitos escritores a chave para alcançar uma naturalidade em fala sintetizada (Teixeira *et al*, 2003). A informação supra-segmental normalmente vêm codificadas através de três parâmetros do sinal acústico de fala:

- A evolução temporal da frequência fundamental (F0);
- Duração de segmentos sonoros que compõe a frase;
- Curva de energia de sinal acústico.

Estes dois tipos de informações, nos conversores texto-fala actuais, são obtidos a partir de uma sequência de tarefas como: Pré-Processamento, Análise Linguística, Análise Morfo Sintáctico-Prosódica, Transcrição Fonética e Processamento Prosódico. Essas tarefas são representadas por blocos, conforme mostra a figura 2, e serão descritas as funções de cada uma.

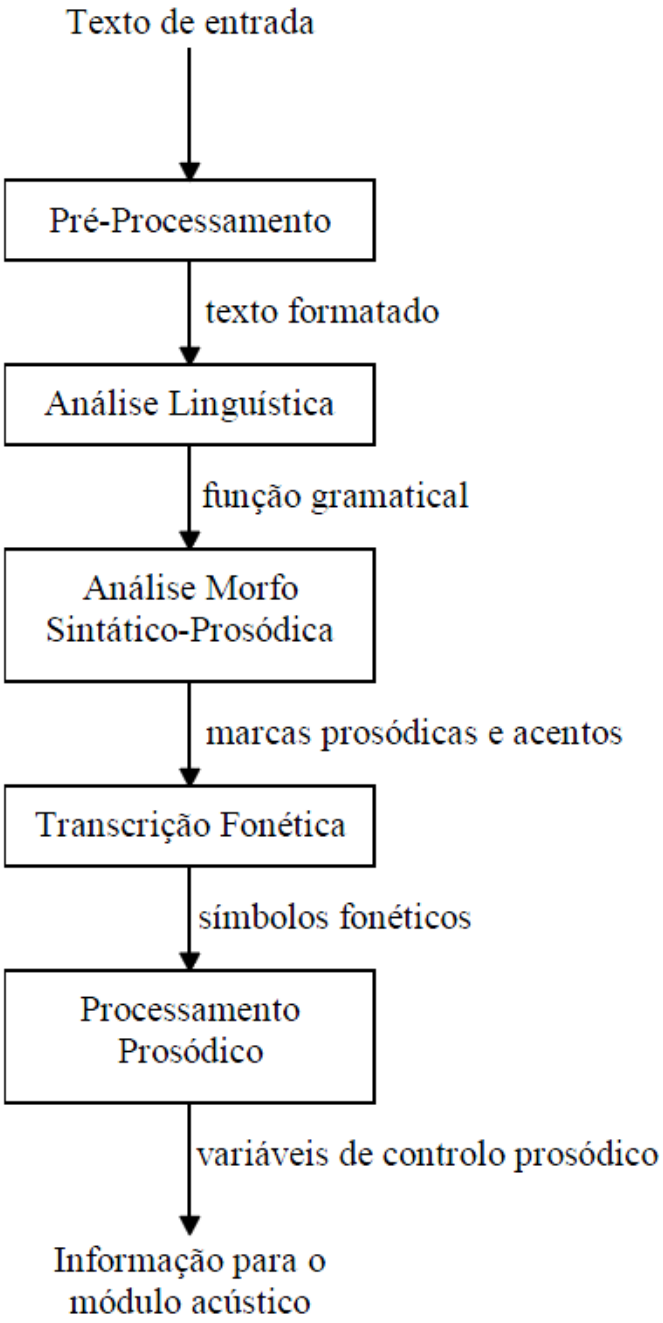

Figura 2 - Diferentes tarefas do processamento linguístico-prosódico (retirada de (Teixeira *et al*, 2003)).

2.1.1 Pré-processamento

A primeira função da etapa de pré-processamento é a formatação do texto, representando adequadamente por extenso, números, abreviaturas, acrónimos e

eventuais caracteres ou conjunto de caracteres que não sejam texto, como por exemplo, o caracter € (Teixeira *et al*, 2003). Neste bloco realiza-se também a separação do texto de entrada em grupos de palavras que facilitem o processo de análise. O grupo que parece mais evidente é a frase, e a maioria dos sistemas separa o texto em frases. Na maioria das línguas, o final das frases é marcado por um conjunto de sinais de pontuação, seguidos de um espaço em branco e de uma palavra em letra maiúscula, pelo que a separação do texto em frases é um processo relativamente simples.

Para a realização destas tarefas recorre-se a tabelas que contém as listas de acrónimos e de abreviaturas e a correspondente forma escrita. No caso dos números o programa que é utilizado é um pouco mais complexo e converte as formas numerais para extenso (Teixeira *et al*, 2003; Pacheco, 2010).

Pode-se considerar a conversão dos números como uma tarefa relativamente simples se for considerado apenas os numerais na sua forma natural. Entretanto, a situação torna-se mais complicada quando se considera que estes podem ser representados de outras formas como decimais, frações, representações na forma exponencial, ordinais, etc. E complica-se ainda mais quando se pretende reproduzir esses números de forma natural. Por exemplo, um número de telefone não é lido da mesma forma que uma data ou uma quantidade monetária. Podemos ver que este caso obriga a considerar as diferentes formas de representar todas as diferentes situações e a identificá-las. A identificação correta, em alguns casos, não depende da forma em que o número está representado, tornando-se então necessário recorrer ao contexto para desambiguar estas situações (Teixeira *et al*, 1998; Teixeira *et al*, 2003).

2.1.2 Análise linguística

Depois do pré-processamento faz-se a análise deste bloco, que abrange tanto uma análise sintática como uma análise semântica, de maneira a encontrar o foco (segmento com maior conteúdo semântico) da oração e tentar modelar aspetos como a ênfase (Teixeira *et al*, 2003).

A fase da análise linguística é muito importante, pois neste bloco pode-se incluir a análise morfológica, identificando a função gramatical de cada palavra, com base nos grandes dicionários de palavras e algoritmos de desambiguação. Esta tarefa é bastante complexa e muito dependente do idioma. A realização da análise gramatical é feita através de um dicionário com um léxico relevante (formas verbais, expressões comuns) e uma tabela de prefixos e sufixos. Normalmente, podem incluir-se regras de conteúdo gramatical para determinar as categorias gramaticais das palavras que não tenham sido encontradas no dicionário.

Ainda a análise deste bloco, são colocadas marcas nas fronteiras de palavras, é feita a separação silábica, marcada a sílaba acentuada, ou eventuais graus de acento. As fronteiras de palavra são facilmente identificadas por haver um caracter de espaço entre palavras. Já não se passa o mesmo para as sílabas, estas podem ser identificadas recorrendo a algoritmos que implementem um conjunto de regras de divisão silábica. Também para a identificação da sílaba tónica é necessário um algoritmo que implementa um conjunto de regras (Teixeira *et al*, 1998; Teixeira *et al*, 2003).

2.1.3 Análise morfo sintático-prosódica

Na análise morfo sintático prosódica pretende-se, a partir da análise anterior, marcar, por um lado, fronteiras sintáticas e/ou prosódicas e, por outro lado, os acentos de palavra. As fronteiras sintático-prosódicas ficam definidas pela sua natureza (relação

lógica entre duas estruturas consecutivas) e força relativa (pausas e alargamento de sílabas). Os acentos obedecem melhor a aspetos rítmicos e de ênfase principal da palavra.

No português muitas palavras são homógrafas mas não homófonas, ou seja, são ortograficamente semelhantes mas tem pronúncias diferentes. O caso dos verbos e substantivos (ou adjetivos) encaixam-se neste perfil. São exemplos de palavras com essa característica "molho" e "seco", no caso da palavra "seco" a ambiguidade é do tipo verbo e adjetivo (Simões, 1999).

Esse procedimento resolve ambiguidades na transcrição fonética para frases como "Eu acordo as seis" e "O governo fechou o acordo". Onde através de uma análise morfológica básica é feita a separação de sujeito, predicado, complementos (Simões, 1999).

De modo a esclarecer o conceito de Morfologia e Sintaxe apresenta-se uma definição de cada uma delas abaixo.

"A Morfologia é à parte da gramática que estuda as palavras de acordo com a classe gramatical a que ela pertence. Quando nos referimos às classes gramaticais, logo sabemos que se refere àquelas dez, que são: substantivos, artigos, pronomes, verbos, adjetivos, conjunções, interjeições, preposições, advérbios e numerais." (Brasil Escola, 2012)

"A Sintaxe é à parte que estuda a função que as palavras desempenham dentro da oração. Agora, referimo-nos a sujeito, adjunto adverbial, objeto direto e indireto, complemento nominal, aposto, vocativo, predicado, entre outros." (Brasil Escola, 2012)

Para saber se a informação de uma palavra é de conteúdo ou funcional será

utilizada no módulo de processamento prosódico para enfatizar ou não esta palavra, com a modificação dos parâmetros prosódicos (Simões, 1999).

Há casos em que a análise gramatical é insuficiente para resolver a ambiguidade, nestes casos uma análise semântica (significado das palavras) e pragmática (intenção do falante) faz-se necessária para a correta pronúncia da palavra, como por exemplo, as frases "A sede do garoto é grande" e a "A sede da empresa fechou". Neste caso a palavra "sede" é substantivo nos dois exemplos e apenas uma análise semântica inferida do contexto da palavra pode resolver esta questão. No entanto poucos são os sistemas hoje que realizam uma análise semântica - pragmática do texto (Simões, 1999).

2.1.4 Transcrição fonética

O objetivo da transcrição fonética consiste na transformação da representação ortográfica em uma representação fonética. Esta tarefa torna-se complicada, uma vez que não existe um mapeamento único no domínio fonético para cada caracter. Entretanto, algumas letras representam mais de um fonema, como a letra "x", que, na língua portuguesa, descreve o fonema [ʃ] em "xaile", [z] em "exame", [s] em "próximo" e os fonemas [ks] em "táxi". Além disso, o processo de transcrição fonética deve ser robusto o suficiente para lidar com nomes próprios, derivados de diferentes idiomas (Teixeira, 2012; Pacheco, 2010).

A transcrição fonética de um texto é realizada mediante regras dependentes do contexto que devem ter em conta, por efeito de coarticulação, também, a existência de pausas e acentuação. Uma das maiores dificuldades na transcrição ortográfico na língua portuguesa, é determinar se as letras "e" e "o" sem acento ortográfico correspondem a vogais abertas ou fechadas. Isto acontece porque nesses casos apenas o contexto lexical não é suficiente para a determinação correta da abertura ou fechamento da vogal. Por

exemplo, para as palavras “bolo” e “bola” não há como desenvolver uma regra que atue apenas pela avaliação do contexto anterior e posterior em que se insere a vogal. A solução é, para esses casos, a inclusão dessas palavras em um dicionário de exceções.

Neste bloco o texto de entrada é transcrito foneticamente para uma sequência de fones (ou códigos de fones). O código SAMPA, alfabeto fonético para leitura por computador, é cada vez mais usado para representar estes fones (SAMPA, 1999-2011).

É muito importante referir a distinção entre representações fonética e fonológica. A representação fonética é o resultado de uma transcrição clássica do texto para fonemas, enquanto que a representação fonológica usa os fones, que são variações dos fonemas, que são efetivamente usados numa determinada realização. Há uma diferença importante, especialmente para o português Europeu, em que a realização fonológica diverge consideravelmente da realização fonética, sendo a mais importante a redução ou elisão de muitas vogais, e de silêncios entre palavras (Teixeira *et al*, 2003).

Este bloco pode ser realizado por um conjunto de regras recorrendo a um dicionário com a transcrição fonética das palavras, ou com máquinas de estados (Teixeira, 1995).

2.1.5 Processamento prosódico

O processamento prosódico, para um sistema de síntese de fala, é considerado como um fator fundamental para que os requisitos de inteligibilidade e naturalidade sejam atendidos.

O termo prosódia diz respeito às características da fala que atuam as nível das sílabas, palavras, orações, parágrafos, etc. O processamento prosódico é, portanto, um processamento de natureza predominantemente supra-segmental (o que não significa, no

entanto, que ele não atue também sobre os segmentos). É através da prosódia que o falante confere estruturação oral à frase, dividindo-a em blocos lógicos menores, que pode ser quebrada mentalmente pelo ouvinte, facilitando-se assim a sua compreensão. A prosódia funciona também como uma portadora da individualidade do falante, pois cada pessoa tem uma maneira particular de enunciar as frases. Muitas características do falante podem ser depreendidas a partir da análise de seus parâmetros prosódicos (as mulheres, por exemplo, possuem um valor de F0 intrinsecamente mais alto que o dos homens) (Simões, 1999).

A prosódia é imposta à fala a partir da variação temporal dos parâmetros prosódicos, frequência fundamental (F0), duração e intensidade.

Esta é a última tarefa a realizar e recolhe a informação supra-segmental e segmental extraída dos últimos passos (marcas prosódicas e transcrição fonética) para as traduzir em variações de duração segmental (ritmo), frequência fundamental (entoação) e inserção de pausas com duração adequada. Por outras palavras, o objetivo de um modelo prosódico é a determinação da evolução temporal dos parâmetros prosódicos, de forma que seja possível identificar na fala sintetizada os atributos linguísticos de acento, ritmo e entoação (Teixeira *et al*, 2003; Pacheco, 2010).

É importante destacar que a modelagem prosódica é fortemente relacionada aos módulos precedentes de análise, principalmente os de determinação de tonicidade e de análise sintática.

Geralmente, a análise prosódica decorre em duas fases:

- Geração de uma representação abstrata/ simbólica de aspetos prosódicos importantes para a síntese de fala;

- Predição dos valores de duração, F0 e intensidade a partir dessa representação abstrata (Oliveira, 2009).

Os resultados deste bloco são comprovadamente determinantes na naturalidade dos sistemas TTS, e a naturalidade é também determinante na aceitação destes sistemas pelos possíveis utilizadores. Define-se a naturalidade como a proximidade à forma natural, ou humana, como os sistemas "falam". Este aspeto é hoje objeto de intensa investigação pela comunidade científica que trabalha com sistemas TTS, por ainda não haver sistemas com naturalidade suficiente. Essa falta deve-se não apenas à insuficiência prosódica, mas também se deve à falta de qualidade segmental bem como às alterações por vezes menos aceitáveis impostas pela modificação prosódica dos segmentos realizadas no bloco acústico (Teixeira *et al*, 2003).

As atitudes do falante (arrogância, humildade, timidez, por exemplo, podem ser expressos através de diferentes estilos de elocução), bem como as suas emoções (alegria, tristeza, surpresa, etc.) e atitudes em relação a si mesmo ou a outrem (ironia, seriedade ou graça). Estes parâmetros definem os aspetos da personalidade do falante (Simões, 1999).

Pode-se dizer que os parâmetros prosódicos são as características do sinal de fala cuja manipulação adequada irá reflectir a estrutura prosódica do enunciado. Estes parâmetros estão associados a cada um dos segmentos fonéticos da frase. Existem três parâmetros prosódicos principais (duração, frequência fundamental e intensidade), os quais serão discutidos a seguir.

2.1.5.1 Duração

Este parâmetro está associado ao intervalo de tempo entre o início e o final de um segmento fonético. Os segmentos fonéticos normalmente possuem durações médias da

ordem de dezenas a centenas de milissegundos, mas deve-se ter em conta que o valor médio e a dispersão são características individuais de cada falante. A duração é um parâmetro prosódico importante, pois varia de acordo com a taxa de elocução do enunciado e reflete também o contexto prosódico em que o segmento fonético está inserido (ambientes prosódicos fortes normalmente ocasionam alongamento dos segmentos fonéticos). Segmentos localizados em fronteiras de constituintes prosódicos também costumam ter a sua duração aumentada (Simões, 1999).

As expressões (1) e (2) são utilizadas para distribuir a duração da sílaba pelos seus fonemas com um valor de z para cada sílaba. Em que μi, e σi são, respectivamente, a média e desvio padrão, do logaritmo das durações, do segmento i (Teixeira, 2004).

$$Dur_i = \exp\left(\mu_i + z\sigma_i\right); \quad (1)$$

$$\sum_i Dur_i = \text{duração da sílaba} \quad (2)$$

Outros autores usam outros modelos para estimar a duração dos próprios fonemas.

2.1.5.2 Frequência fundamental (F0)

Modelar as curvas de frequência fundamental é a questão mais importante para transmitir naturalidade à fala sintética. Diferentes tipos de modelos têm sido utilizados para modelar o parâmetro F0 (Pierrehumbert, 1980).

A frequência fundamental (F0) de um sinal de fala é um valor instantâneo que está diretamente associado à taxa de vibração das cordas vocais, e que se manifesta através da periodicidade da forma de onda nos sinais sonoros. Quando se trata de um segmento de fala não sonoro, não faz sentido falar de F0, uma vez que nesse caso não ocorre vibração das cordas vocais, e a forma de onda tem características aperiódicas (Teixeira

et al, 2003; Simões, 1999).

O significado de “pitch” está associado à frequência fundamental, e no contexto de síntese e reconhecimento de fala os dois termos costumam ser utilizados de forma equivalente. Mas é necessário ter em consideração que a frequência fundamental é um valor numérico real, associado a cada instante do sinal de fala, e corresponde ao inverso do período do sinal sonoro. A frequência fundamental é expresso em Hertz (Hz).

Enquanto que “pitch”, por sua vez, é um conceito meramente perceptual, e diz respeito à sensação de altura (grave/agudo): quanto maior for a frequência fundamental, maior será o “pitch” ou, equivalentemente, mais agudo será o sinal. A relação entre a frequência fundamental e a sensação de altura do sinal é quase logarítmica, e portanto não linear (Simões, 1999).

A frequência fundamental constitui um dos parâmetros mais importantes, juntamente com a duração, a ser considerado durante a etapa de tratamento prosódico.

Existem diversos modelos para modelação de F0, sendo os mais usados os modelos de Tobi e o modelo paramétrico de Fujisaki (Teixeira, 2004).

2.1.5.3 Intensidade

A intensidade tem a ver com a amplitude da forma de onda. Através deste parâmetro pode-se verificar a diferença entre as amplitudes dos sons fortes e dos sons fracos.

A amplitude é considerada como um parâmetro prosódico menos importante que os outros no que à prosódia diz respeito. Na prática, a intensidade do sinal tem uma função de contraste muito menos significativa do que os outros parâmetros prosódicos, como a duração e a frequência fundamental (Simões, 1999).

Com o tempo, passou-se a acreditar que os segmentos fonéticos acentuados se destacavam por meio de um padrão de energia mais alto, no entanto, verifica-se que são as variações de duração e de F0 que determinam, a localização do acento nas frases, mas a contribuição específica de cada parâmetro varia de acordo coma a língua.

A maioria dos sistemas de conversão texto-fala não faz o modelamento de energia durante a etapa de processamento prosódico, focando-se no tratamento dos padrões de duração e de frequência fundamental (Simões, 1999).

2.2 Processamento acústico

Nesta etapa, já são conhecidos os sons que devem ser sintetizados e os parâmetros prosódicos que devem ser aplicados. É realizada, então, a síntese do sinal acústico de fala que é a última tarefa a ser desenvolvida durante o processo de conversão texto-fala, cuja função é de gerar um sinal acústico a partir da sequência de fones determinada pelo módulo de transcrição fonética e das variáveis prosódicas calculadas durante a fase de processamento prosódico (Oliveira, 2009).

O objetivo do processamento acústico consiste, basicamente em converter uma sequência fonética e as variáveis de controlo prosódico na forma de onda associada à voz sintetizada.

Deve-se levar em consideração a existência de um compromisso entre, por um lado, o número de regras de parametrização e concatenação, destinadas a evitar transições bruscas desagradáveis ao ouvido e, por outro lado, o tamanho da base de dados de parâmetros.

A representação do processamento acústico pode ser feita a partir de um diagrama de blocos típico, conforme mostra a figura 3.

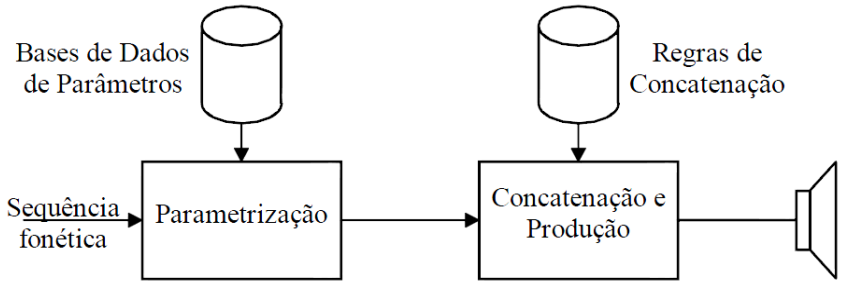

Figura 3 - Diagrama de blocos de processamento acústico (retirada de (Teixeira, 1995)).

Em qualquer caso o modelo de produção de fala deve ser flexível para o controlo prosódico e deve ter uma alta qualidade na geração de fala sintética.

No módulo acústico pode-se utilizar o modelo de Predição Linear (LPC) (Atal B., 1986; Schroeder, M., 1985), o modelo articulatório, muito limitado devido à sua complexidade, o modelo baseado na concatenação de unidades (Paiva, 2005) com muito boa qualidade e melhores resultados, modelo de formantes (Klatt, 1980), Modelo Ocultos de Markov (Hidden Markov Models, HMM), que é o mais recente (Tokuda et al, 2002) entre outros.

Faz-se de seguida uma descrição de 3 modelos de sínteses muito utilizados para a geração de fala sintetizada, nomeadamente a síntese de formantes, a síntese baseada na concatenação de unidades e a síntese mediante modelos articulatórios (Teixeira, 1995).

2.2.1 Síntese de formantes

Este modelo é baseado no modelo fonte-filtro da teoria acústica de produção de fala, onde o filtro é caracterizado por variar suavemente as frequências formantes ao longo do tempo. Para a geração da voz sintética neste modelo é preciso modelar a fonte

de excitação e modelar os filtros capazes de simular o trato vocal através de suas funções de transferência. O trato vocal é descrito através das suas frequências de ressonância (formantes) e respetivas larguras de banda (Teixeira, 1995). Para uma boa qualidade de síntese é necessário usar pelo menos quatro formantes.

Nos sintetizadores de formantes as sequências fonética e prosódica controlam respetivamente as ressonâncias e a excitação do sintetizador de formantes. A configuração mais genérica para o modelo destes filtros é a sua ligação em série e em paralelo. Trata-se de um procedimento, com enorme flexibilidade, que sintetiza a fala com elevada qualidade, mediante ajuste manual dos parâmetros do sintetizador (Teixeira, 1995).

Neste trabalho para a produção de fala é utilizado este modelo. A figura 4 representa um modelo de um sintetizador de formantes. Neste esquema, a partir de uma excitação vozeada e não vozeada, o sinal sonoro é amplificado, através do ganho, e filtrado. Dessa forma, a voz é então produzida (Barros, 2002).

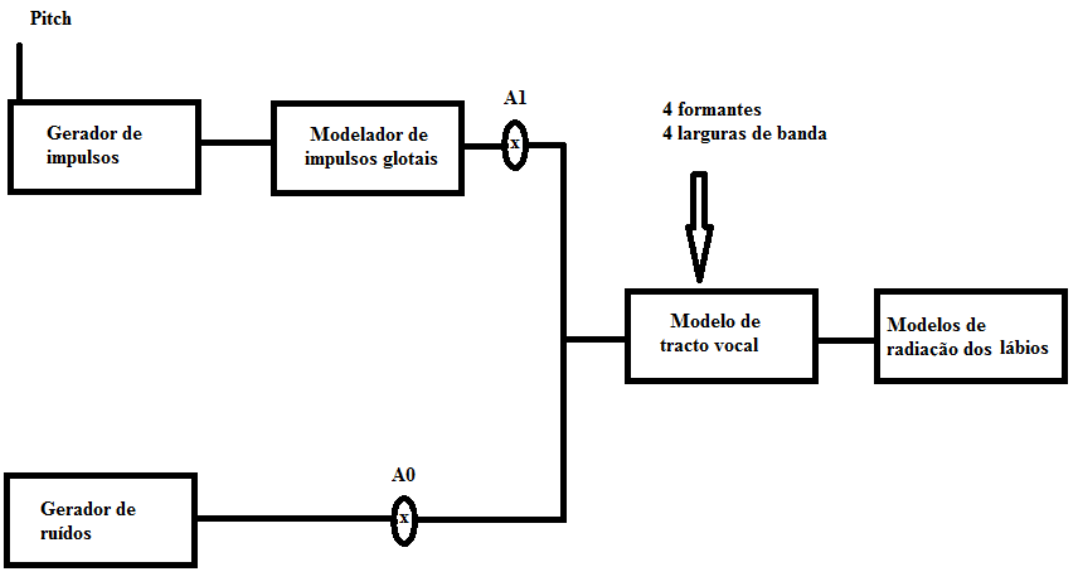

Figura 4 - Esquema de um sintetizador de formantes (retirada de (Barros,

2002))

2.2.2 Síntese baseada na concatenação de unidades

Os sintetizadores concatenativos produzem um sinal de fala através da concatenação de segmentos de fala natural, previamente gravados e armazenados numa base de dados.

A síntese concatenativa é considerada como o método mais simples, pois quando comparado com a síntese por formantes e síntese articulatória não é necessário derivar regras para simular a evolução dos parâmetros entre transições entre os fones. A síntese concatenativa resolve esse problema através da justaposição de segmentos de voz natural pré-gravados, esses segmentos preservam a informação da transição entre fones, e são guardados em um banco de unidades.

Um dos aspectos mais importantes em síntese baseada na concatenação de unidades é determinar o correto comprimento da unidade de fala. A seleção do comprimento da unidade, é normalmente um intercâmbio entre unidades mais longas e unidades mais curtas. Com unidades mais longas, obtém-se alta naturalidade, menos concatenação são efetuadas e bom controlo de coarticulação é alcançado, mas, quanto mais unidades são exigidas, mais memória também é consumida. Com unidades mais curtas, é consumida menos memória, mas a concatenação destas, exige procedimentos mais difíceis e complexos. Nos sistemas atuais essas unidades são normalmente palavras, sílabas, fonemas, difones e às vezes trifones. (Albuquerque, 2001).

Nesta síntese, como referido no primeiro parágrafo, a fala sintética é produzida pela concatenação de segmentos. A escolha dos segmentos necessários para a geração de uma dada elocução baseia-se nas informações obtidas a partir da etapa de processamento linguístico. Com uma etapa de concatenação e alteração de parâmetros

prosódicos, a fala sintetizada é gerada.

Quando comparado com a síntese por formantes, verifica-se que não há necessidade de definição de regras de transição entre sons, pois essas podem estar incorporadas nos segmentos armazenados. Cada segmento é obtido de uma gravação de um locutor, e um resultado de alta qualidade poderia ser esperado. Porém, podem acontecer alguns problemas, fazendo com que os sistemas concatenativos sofram de uma grande variação de qualidade. Numa frase, o resultado é excelente, mas na seguinte, pode ser sofrível. Se a combinação das unidades em uma frase sintética é adequada, o resultado é tão bom quanto o obtido naturalmente em uma gravação (Pacheco, 2010).

2.2.3 Síntese mediante modelos articulatórios

O objectivo da síntese articulatória é reproduzir o sinal de fala, modelando os mecanismos de sua produção natural (Pacheco, 2010). É potencialmente o melhor método para a geração de fala sintética de alta qualidade. Ao mesmo tempo, o de implementação mais complexa, por depender de uma ampla compreensão do processo de produção da fala, e o mais caro computacionalmente (Lemmetty, 1999).

Os articuladores normalmente são modelados como uma área situada entre a glote e a boca. O primeiro modelo de articulador foi baseado em uma área vocal onde funciona a região da laringe e lábios, para cada segmento fonético. Para síntese articulatória baseada em regras, os parâmetros de controlo podem ser por exemplo: abertura dos lábios, protuberância dos lábios, altura da ponta da língua, posição da ponta da língua, altura da língua, posição de língua, etc. Os parâmetros de excitação podem ser abertura da glote, tensão das cordas vocais e pressão pulmonar (Albuquerque, 2001).

Os estudos feitos a respeito desta síntese são recentes, e os experimentos

realizados até agora apenas conseguiram a geração de segmentos curtos de fala, mas as pesquisas até agora mostram o potencial promissor deste tipo de síntese (Simões, 1999).

Basicamente a síntese mediante modelos articulatórios simula a propagação das ondas acústicas no trato vocal. Os segmentos e as variáveis prosódicas traduzem-se em parâmetros de um modelo simplificado do aparelho fonador humano que explicitamente modelam a dinâmica do sistema, podendo produzir voz da mais alta qualidade (Teixeira *et al*, 2003).

CAPÍTULO 3
DESENVOLVIMENTO DA SÍNTESE PARA VOGAIS

3 DESENVOLVIMENTO DA SÍNTESE PARA VOGAIS

3.1 Objectivo

Neste capítulo, o objectivo consiste em sintetizar as vogais da Língua Portuguesa, a partir dos valores da frequência de formantes e de larguras de banda (extraído do Praat). Com a frequência fundamental (F0), pretende-se a variação da entoação dos vogais. Os valores de F0 variam entre 100Hz a 250Hz. O sinal excitador é gerado a partir de seguintes formas: um trem de impulsos, onda rectangular, onda dente de serra, onda sinusoidal e impulsos glotais sintetizados.

Com isto, pretende-se gerar o som de uma vogal e que este sinal seja armazenado no computador para que seja posteriormente analisado.

Após o desenvolvimento desta síntese, foi utilizado a interface gráfica (Guide) para representar os sinais numa janela.

3.2 Classificação das vogais a sintetizar

As vogais podem ser classificadas quanto à base articulatória, a base acústica, ao grau de elevação da língua e a dimensão de abertura do trato vocal.

3.2.1 Classificação quanto à base articulatória

Quanto à base articulatória, as vogais podem ser caracterizadas de acordo com a região de articulação e qualidade vocal.

3.2.1.1 Região de articulação

A classificação relacionada à região de articulação está voltada para a parte ou o ponto em que ocorre o contacto da língua e do palato para a produção do som vocálico.

Deste modo, classifica-se como vogal média aquela em que a língua permanece baixa e a boca fica entreaberta. A vogal anterior ocorre quando a abertura da boca é diminuída e a parte anterior da língua é elevada em direcção ao palato duro. Por fim, a vogal posterior é aquela em que a parte posterior da língua é elevada em direcção ao palato mole enquanto os lábios são arredondados (Chbane, 1994; Cunha, 2011).

3.2.1.2 Qualidade vocal (timbre)

A classificação quanto à qualidade vocal, está diretamente relacionada à abertura dos lábios. Assim, classifica-se como abertura máxima a vogal que exige abertura quase que total do trato vocal para ser emitida.

De modo oposto, classifica-se como abertura mínima a vogal que exige pouca abertura do trato vocal e abertura média para a vogal de uma abertura média. A tabela 1 apresenta a classificação quanto à base articulatória (Chbane, 1994; Cunha, 2011).

Tabela 1 - Classificação quanto à base articulatória (retirada de (Chbane, 1994; Cunha, 2011)).

Região de articulação			**Qualidade vocal (timbre)**		
Vogal Média	Vogal Anterior	Vogal Posterior	Abertura Máxima	Abertura Média	Abertura Mínima
a	e,i	o,u	a	e,o	i,u

3.2.2 Classificação quanto à base acústica

A classificação quanto a base acústica está relacionada com a intensidade da força do ar que foi expirado pelos pulmões e faz vibrar as cordas vocais.

Classifica-se então uma vogal tónica aquela que exige maior energia para ser pronunciada, e a vogal átona, que exige menor energia. Na primeira classificação encontram-se as vogais que estão localizadas nas sílabas que necessitam de maior força para serem pronunciadas. Opostamente na segunda classificação encontram-se as sílabas que pedem menos força ao serem articuladas (Chbane, 1994; Cunha, 2011).

Tabela 2 - Classificação quanto à base acústica (retirada de (Chbane, 1994; Cunha, 2011)).

Vogal tónica	Vogal átona
á,é,í,ó,ú	ã,õ

3.2.3 Classificação quanto ao grau de elevação da língua

Aqui as vogais podem ser classificadas como altas, ou seja, a língua encontra-se em posição elevada no trato vocal, médias, posição média, e baixas, posição baixa.

Tabela 3 - Classificação quanto ao grau de elevação da língua (retirada de (Chbane, 1994; Cunha, 2011)).

Vogal alta	Vogal média	Vogal baixa
i,u	e,o	a

3.2.4 Classificação quanto á dimensão de abertura do tracto vocal

As vogais também podem ser classificadas quanto a dimensão de abertura do trato vocal. Para este caso, têm-se as vogais abertas e as fechadas. As vogais abertas são aquelas em que a dimensão é maior, pois exigem maior energia para serem pronunciadas, e as vogais fechadas, que, por necessitarem de menos energia, possuem

menor dimensão (Chbane, 1994; Cunha, 2011).

Tabela 4 - Classificação quanto a dimensão do trato vocálico (retirada de (Chbane, 1994; Cunha, 2011)).

Vogal aberta	Vogal fechada
a,e,o	i,u

3.3 Funcionalidade do sistema

Aqui, começa-se por explicar o funcionamento geral deste sistema, ou seja faz-se uma descrição de como foi implementada a síntese de formantes, desde a produção de fala até a sua representação, procurando explicar a funcionalidade de cada etapa.

3.3.1 Produção de fala e o modelo utilizado

O processo de produção da fala está ligado aos órgãos e sistema de respiração.

Quando se expira são permitidas maiores variações de pressão do que no processo de inspiração, tornando-se audível, pela produção de ondas sonoras que, modeladas pela laringe e as cavidades superiores orais e nasais, darão as características da voz.

Para que seja possível produzir uma voz e estabelecer comunicação, o organismo humano faz uso de alguns órgãos, conforme mostra a figura 5. Na laringe encontram-se as cordas vocais, que possuem um papel de fundamental importância neste processo, e a glote. A esse conjunto chama-se, de aparelho fonador.

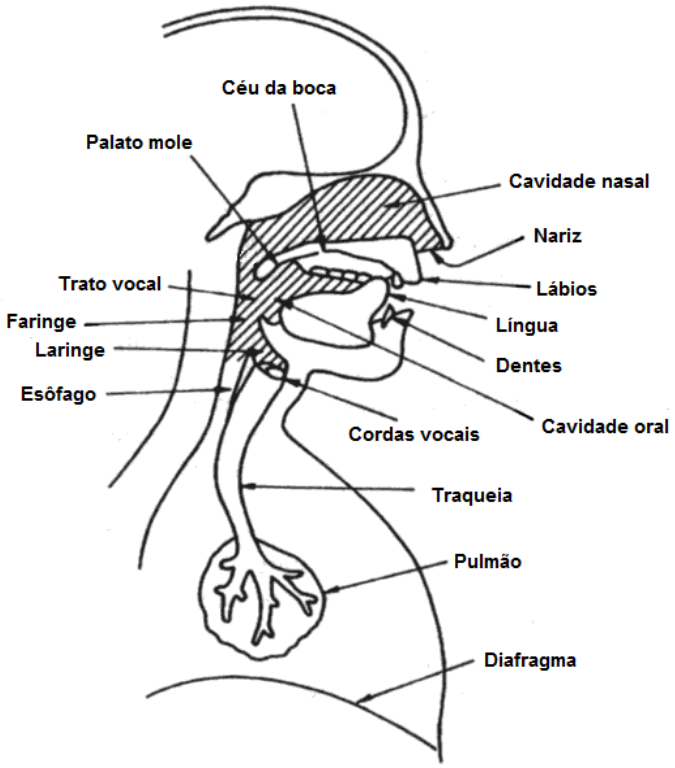

Figura 5 - Aparelho fonador humano (retirada de (Meneses, 2008)).

No processo da produção da fala, o aparelho fonador transforma o ar saído dos pulmões em som articulado. Este processo tem início quando esta corrente de ar percorre os brônquios, penetra na traqueia e atinge a laringe. Aqui poderá encontrar o primeiro obstáculo.

Após atravessar a glote encontrará as cordas vocais, que são duas cordas musculares, poderão estar abertas ou fechadas. Estando abertas, esta corrente não possuirá barreiras neste trecho do percurso. Se estiverem fechadas, o ar forçará a passagem. Tal esforço causará vibração nas cordas e reproduz em som.

Este som manterá o percurso e encontrará o segundo obstáculo. Ao entrar na faringe, encontrará duas vias de acesso ao meio externo: cavidade bucal e nasal. Quem determinará o destino deste som será a úvula. De acordo com a posição que adotar, o som irá atravessar só o canal bucal ou ambos os canais. Assim, estando a úvula

levantada, isto é, unida a parede posterior da faringe, o canal nasal será obstruído e o som sairá só para o canal bucal. Estando abaixada, a corrente de ar irá se dividir e ressoará por ambos os canais. Os órgãos encontrados nestes canais serão responsáveis por dar forma ao som, isto é, transformá-lo em voz humana (Teixeira, 1995).

As cavidades supraglotais são formadas por faringe e cavidades nasais e oral. Estas têm um papel fundamental na fonação de diferentes sons.

As cavidades supraglotais formam um conjunto de ressoadores que favorecem a passagem de algumas frequências e a atenuação de outras consoante as suas formas e dimensões. Às frequências favorecidas pelas cavidades supraglotais dá-se o nome de frequências formantes ou simplesmente formantes, e ao conjunto das formas tomadas pelas cavidades supraglotais chama-se trato vocal (Teixeira, 1995).

Um modelo de produção da fala em que se separa a fonte sonora (excitação acústica que pode ser causada pela vibração das cordas vocais ou pelo simples fluxo de ar) da operação de filtragem realizada pelo trato vocal dá-se normalmente o nome de modelo de engenharia e pode ser representado como na figura 6.

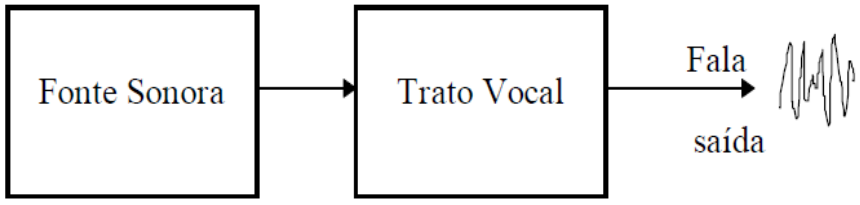

Figura 6 - Modelo de engenharia (retirada de (Teixeira, 1995)).

A fonte sonora no modelo de engenharia é um sinal excitador que pode ser periódico ou um sinal de ruído aleatório. O caso do sinal excitador ser periódico acontece nos sons vocalizados pelas cordas vocais com o abrir e fechar da glote. A frequência deste sinal é o inverso do tempo de duração do impulso glotal e é

denominada por frequência fundamental (F0). O ruído aleatório, como sinal excitador, é produzido com a passagem de ar pela glote completamente aberta (Teixeira, 1995).

A gama de valores da frequência fundamental (F0) varia com as pessoas, ou seja, há valores de F0 para falantes masculinos, femininos e crianças.

A gama típica de valores desta frequência para os homens é dos 80 aos 200 Hz, para as mulheres entre os 200 e 300 Hz e para as crianças dos 400 aos 500 Hz (Teixeira, 1995). Os valores de F0 variam com outros parâmetros como por exemplo: o período do dia (manhã, tarde e noite); estado nervoso, etc. o valor da F0 ainda varia com o modo de expressar, dando a entoação pretendida à frase que pode ser do tipo interrogativa, declarativa, exclamativa…

Aqui chama-se de formantes as zonas ou gamas de frequências mais favorecidas pelas cavidades supraglotais, as frequências de ressonância do trato vocal. O termo formante é um adjetivo atribuído às frequências que "formam" a fala (Teixeira, 1995). A frequência da formante será assim a frequência central da região de amplitudes mais pronunciadas e a largura de banda será a sua banda a -3 dB da mesma zona no espectro quando analisado um intervalo de tempo curto do sinal de fala.

Cada som é caracterizado por um conjunto de formantes. Uma sequência específica de sons dá origem a um fonema que é percebido pelo cérebro humano. Assim um fonema terá sempre ao longo do tempo uma variação idêntica de uma série de formantes. A figura 7 apresenta o espectro alisado típico de um som da vogal [a] onde são visíveis os 4 formantes.

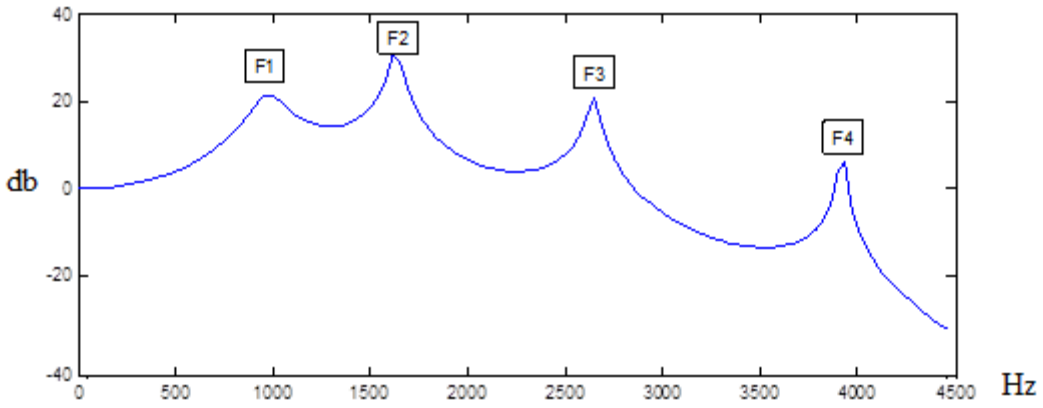

Figura 7 - Espectro alisado de um segmento de som da vogal [a].

Quando se usam formantes para parametrizar o trato vocal, é necessário guardar as frequências destes e as respetivas larguras de banda.

Nesta tese refere-se a F1 como a primeira formante, F2 a segunda formante, F3 a terceira formante e F4 a quarta formante, e B1 como primeira largura de banda, B2 a segunda largura de banda, B3 a terceira largura de banda e B4 a quarta largura de banda.

Uma questão importante que se coloca é o número de formantes a utilizar (neste trabalho são 4 como referido anteriormente). Deste modo, sabe-se que com os valores do primeiro e segundo formantes consegue-se caracterizar uma vogal (Teixeira, 1995). Com muito estudo os autores chegaram a conclusão que utilizando quatro formantes para modelizar o trato vocal, consegue-se uma qualidade muito boa. Este resultado foi verificado experimentalmente.

Foi utilizado o modelo de engenharia como o modelo de produção de fala.

3.3.2 Modelo de engenharia

Segundo (Teixeira, 1995), o modelo de terminais análogos (apresentado na figura 8) é o modelo mais utilizado para representar o processo de produção de fala.

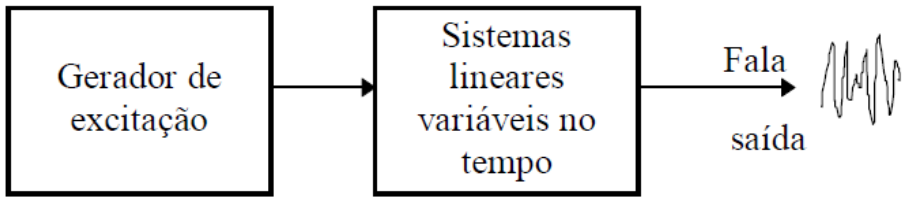

Figura 8 - Modelo de engenharia (retirada de (Teixeira, 1995)).

Este é um modelo linear que quando é feito o controlo dos parâmetros da produção de fala, na saída obtém-se as propriedades desejadas da fala.

De modo a produzir sinais parecidos com a fala o modo de excitação e as propriedades do sistema linear devem variar com o tempo.

Quando se trata de um sistema linear excitado por um sinal cuja natureza básica pode ser um trem de impulsos quase periódicos para a fala vocalizada, um ruído aleatório para a fala não vocalizada ou um sinal misto para fricativas vocalizadas e zonas de transição, o modelo de engenharia apresenta uma variação lenta no tempo (Teixeira, 1995).

Num sistema linear, a relação entre a entrada e a saída é dado pela função de transferência V(z), mostrada na equação 3 (Rabiner; Schafer, 1978).

$$V(z) = \frac{G}{1 - \sum_{K=1}^{N} \alpha_K z^{-1}} \tag{3}$$

Onde, G e α_K dependem da função da área.

O modelo completo de engenharia, além do trato vocal, inclui a variação da função de excitação e dos efeitos da radiação do fluxo de ar nos lábios (Teixeira, 1995).

3.3.2.1 Trato vocal

As ressonâncias do trato vocal, modelizadas por formantes, correspondem aos polos da função de transferência V(z). A maioria dos sons da fala pode ser modelada com um modelo só com polos, visto que é considerado como um boa representação desses efeitos. Embora, sabe-se, da teoria acústica, que os sons nasais e fricativos requerem ressonâncias e anti-ressonâncias (polos e zeros) para a sua correta representação. Nestes casos deve-se incluir zeros na função de transferência ou aumentar o número de polos que simula o efeito de um zero na função. Em muitos casos é preferível esta segunda aproximação (Rabiner; Schafer, 1978).

Pode-se representar V(z) como uma cascata de sistemas ressoadores de segunda ordem (Rabiner; Schafer, 1978).

$$V(z) = \prod_{K=1}^{M} V_K(z) \tag{4}$$

Onde M é o número de pares de polos complexos conjugados, e V(z) é a função de transferência para um par de polos conjugados (um formante) do trato vocal e é dado pela eq. seguinte que tem apenas polos e não tem zeros (Rabiner; Schafer, 1978):

$$V_K(z) = \frac{(1 - 2|z_K|\cos(2\pi F_K T) + |z_K|^2)}{(1 - 2|z_K|\cos(2\pi F_K T)z^{-1} + |z_K|^2 z^{-2})} \tag{5}$$

Note-se que à frequência zero (z=1), $V_K(1) = 1$

$|z_K|$ é a distância do polo à origem e F_K a frequência do polo.

$$|z_K| = e^{-\sigma_K T} = e^{-\frac{B_K}{2}T} \tag{6}$$

Em que B_k é a largura de banda do formante k.

Este é o trato vocal usado na realização do trabalho. No Matlab o código é dado por:

```
F=FORM(i,j);
B=BAND(i,j);
BB=1-2*abs(exp(-B*T/2))*cos(2*pi*F*T)+abs(exp(-B*T));
A=[1 -2*abs(exp(-B*T/2))*cos(2*pi*F*T) abs(exp(-B*T))];
sinal=filter(BB,A,sinal)
```

Em que F representa os formantes e B as larguras de banda. Os vetores F e B têm a dimensão de 4 linhas para os 4 formantes/larguras de banda e comprimentos igual ao número de segmentos de fala. A variável BB (com apenas um termo) contém os coeficientes do polinómio numerador e a variável A (com 3 termos) os coeficientes do polinómio denominador da função de transferência de V(z).

Inicialmente, definem-se os valores da frequência de amostragem e da frequência fundamental. Os valores das frequências dos formantes e larguras de banda são os parâmetros da entrada da função do trato vocal. Para cada vogal a sintetizar usa-se um vetor com os valores de 4 formantes (F1, F2, F3 e F4) e outro vetor com os valores de 4 larguras de bandas (B1, B2, B3 e B4). Depois é usado a equação (5) para filtrar o sinal do impulso glotal para cada formante.

3.3.2.2 Efeito de radiação dos lábios

Até aqui foi considerada a função de transferência V(z) como a velocidade do volume de ar desde a fonte até aos lábios. Os efeitos de radiação devem ser incluídos quando se pretende um modelo para a pressão de ar nos lábios (Rabiner; Schafer, 1978).

De acordo com (Rabiner; Schafer, 1978), a pressão está relacionada com a velocidade volumétrica do ar por uma operação de filtragem passa alto.

Muitos autores concluem que uma boa aproximação dos efeitos de radiação dos lábios é conseguida pela seguinte expressão (Teixeira, 1995):

$$R(z) = R_0 - z^{-1} \quad (7)$$

Este modelo de radiação pode ser considerado em cascata com o modelo do trato vocal como na figura 9.

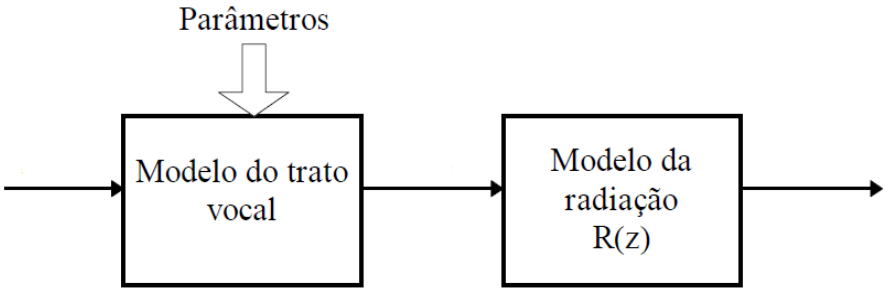

Figura 9 - Modelo de engenharia incluindo os efeitos de radiação (retirada de (Teixeira, 1995)).

No Matlab, depois de simular o trato vocal, e resultado é armazenado num vetor de forma sequencial. Depois usa-se o modelo de radiação de lábios para filtrar o resultado do sinal concatenado.

```
AR=1e-2;
BR=[1 -1];
sinalfinal=filter(BR,AR,sinal);
```

3.3.2.3 Excitação

Os sons de fala podem ser classificados como vocalizados ou não vocalizados. É necessária uma fonte geradora de excitação (como mostra a figura 10) capaz de produzir formas de onda de impulsos (representada na figura 11) e ruído aleatório (representado na figura 12). Pode-se fazer a combinação destes dois tipos de forma de onda para a obtenção de uma excitação mista (representado na figura 13).

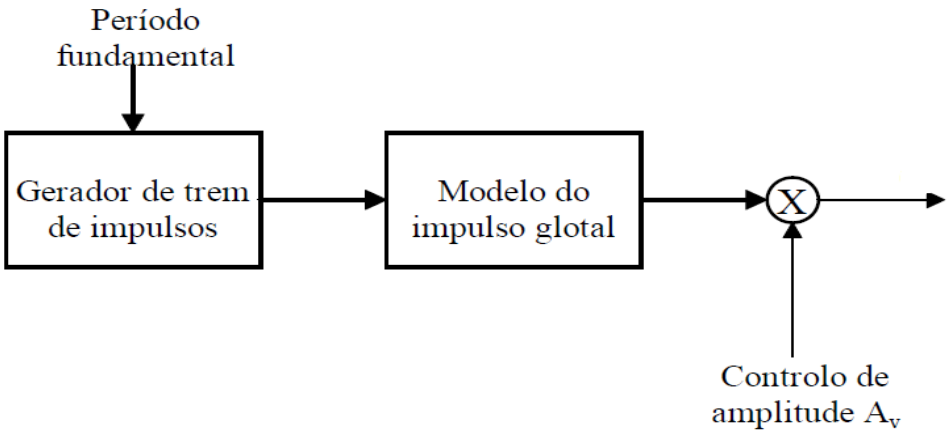

Figura 10 - Gerador do sinal de excitação para a fala vocalizada (retirada de (Teixeira, 1995)).

O gerador de trem de impulsos produz sequências de impulsos unitários espaçados pelo período fundamental desejado. Por sua vez esta sequência de impulsos excita um sistema linear com resposta impulsional g(n) desejada para a forma de onda glotal. O controlo do ganho Av controla a amplitude de excitação e pode ser usado como o controlo físico da amplitude ou de energia do sinal acústico (parâmetro prosódico).

O impulso glotal pode ser conseguido a partir da seguinte função de transferência (Teixeira, 1995):

$$G(z) = \frac{-ae\ln(a)z^{-1}}{(1-az^{-1})^2} \tag{8}$$

Onde a está relacionado com o timbre da voz.

A figura 11 apresenta a forma de onda do impulso glotal sintético obtido a partir da expressão (8).

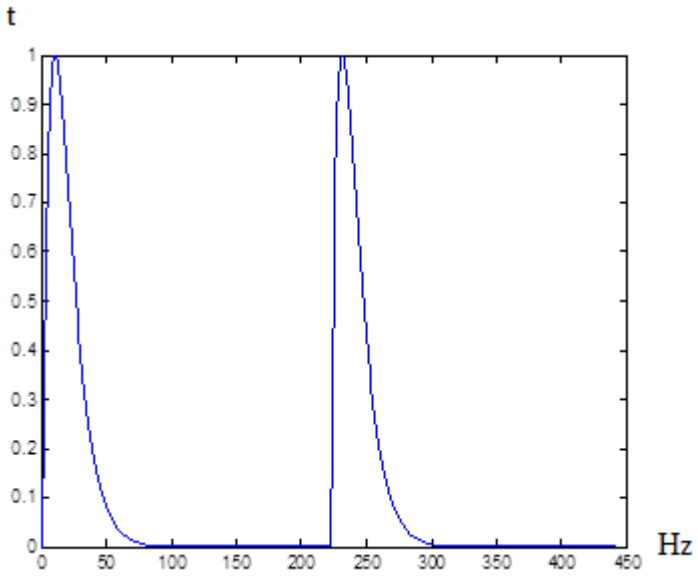

Figura 11 - Forma de onda do impulso glotal sintético G(z), com a=0.90

O modelo de excitação para sons não vocalizados é mais simples, é necessário uma fonte de ruído aleatório e um parâmetro de ganho para controlar a intensidade da excitação não vocalizada.

Para se obter uma excitação mista, utiliza-se um somador que some o ruído aleatório com o sinal periódico sendo o resultado a excitação mista.

Na prática, usando o Matlab, o impulso glotal foi definido por uma função cujos parâmetros de entrada são a frequência de amostragem (Fa) e a frequência fundamental (F0). Nesta função foi gerado segmento de duração aproximadamente 20ms, com um trem de impulsos com o período fundamental (T0), e para o parâmetro a, que representa a timbre de voz o valor de 0.90. Foi realizado vários testes com diferentes valores do parâmetro a, até que se chegou a conclusão de que o valor 0.9 é o ideal para esta síntese, ou seja, com este valor o sintetizador apresenta um sinal de fala mais clara e compreensível para ouvido humano.

O resultado deste sinal é filtrado usando a equação (8), conforme mostra o código

seguinte:

```
T=1/Fs;
T0=1/f0;
N=round(20e-3/T0);
salto =round(Fs/f0);
a=0.9;
B=[0 -a*exp(1)*log(a)];
A=[1 -2*a a^2];
sinal=zeros(1,round(N*T0*Fs));
for i=1:salto:round(N*T0*Fs),
    sinal(i)=1;
end
 sinal=filter(B,A,sinal);
```

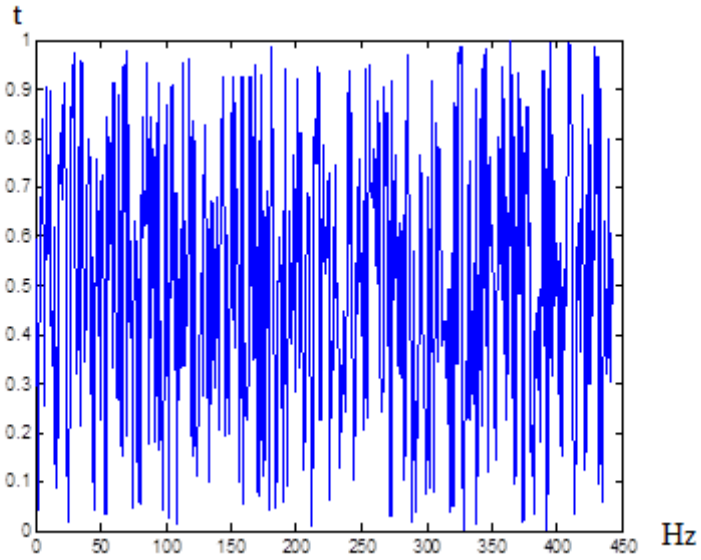

Figura 12 - Geração de um ruido aleatório.

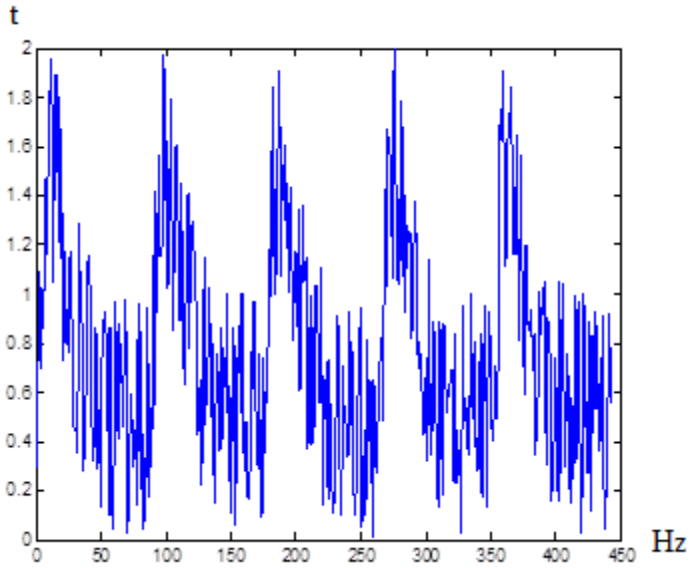

Figura 13 - Excitação mista.

3.3.2.4 Síntese completa

O impulso glotal foi implementado com uma função cujos parâmetros de entrada são a frequência de amostragem e frequência fundamental e a saída é um vetor com um número inteiro de impulsos glotais completos. O número de impulsos glotais de cada segmento é o inteiro mais próximo de 20 ms, então gera-se um trem de impulsos, que por sua vez é filtrado usando o sistema definido pela equação (8).

Os parâmetros dos formantes e larguras de banda para cada vogal foram previamente gravados. A síntese foi realizada utilizando estes parâmetros gravados e suavizada no seu início e fim com uma janela de hanning em que a metade dessa janela foi aplicada no início do sinal de fala e a outra metade no fim do sinal de fala a fim de evitar o início e o final abrupto do som sintetizado.

Como referido anteriormente, o filtro do trato vocal foi implementado com a equação (5) para cada formante e respetiva largura de banda. Inicialmente a função

glotal é chamada para criar um impulso glotal. Então este sinal é filtrado num ciclo de 4 iterações com a correspondente resposta em frequência do filtro da equação (5) e utilizando os valores do primeiro formante e da largura de banda na primeira iteração e os formantes, segundo, terceiro e quarto / largura de banda pares nas iterações seguintes.

O resultado é armazenado num vetor para a concatenação. Para implementar a radiação dos lábios o sinal foi filtrado através do filtro da equação (7).

Em seguida, as duas metades da janela de hanning foram aplicados para o início e o fim do sinal. Por fim, o sinal está representado na parte inferior da janela e o som da fala é reproduzido.

O banco de dados com os parâmetros dos formantes e largura de banda para as respetivas vogais foi construído usando a ajuda do programa Praat (Boersma, Paulo e Weenink, David), da forma explicada adiante.

A figura 14 apresenta um modelo genérico utilizado para a produção de fala

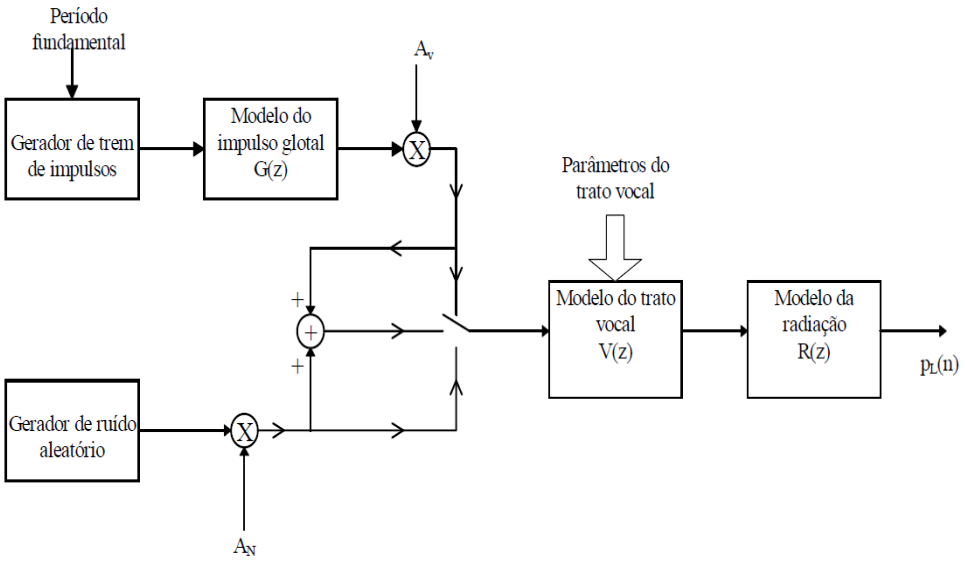

Figura 14 - Modelo genérico para a produção da fala (retirada de (Teixeira, 1995)).

Quando se comuta o gerador de excitação vocalizada ou não vocalizada e excitação mista pode-se modelar a alteração do modo de excitação.

Segundo (Teixeira, 1995) devido à variação temporal dos parâmetros do modelo e do respetivo som, para o caso dos sons contínuos, os parâmetros variam lentamente e o modelo funciona perfeitamente bem, que é o caso dos vogais. Com os sons de transições mais rápidas como as oclusivas, o modelo não é perfeito mas continua a produzir resultados bons ao nível da sua inteligibilidade e aceitáveis ao nível da sua naturalidade.

3.3.2.5 Praat

Para extrair os parâmetros do modelo de formantes para as vogais, foi utilizado o programa Praat, onde foram gravados os sinais de voz, a uma frequência de amostragem de valor igual a 22050Hz, com o objetivo de recolher as frequências de formantes e larguras de banda de cada vogal.

O Praat é um programa de software científico livre e gratuito. Como dito anteriormente no capítulo da introdução, o Praat é um programa para a análise de discurso em fonética. Foi projetado e continuamente desenvolvido por Paul Boersma e Weenink David, da Universidade de Amesterdão (Fonseca, 2009).

É um programa de fácil acesso, constantemente atualizado sendo que uma nova versão é publicada com muita frequência. O programa permite trabalhar com arquivo de som longo e curto, bem como com arquivos mono e stereo, guarda e lê vários formatos de sons.

A figura abaixo (figura 15) mostra a janela de edição na qual são realizadas as principais medidas acústicas de um sinal de fala.

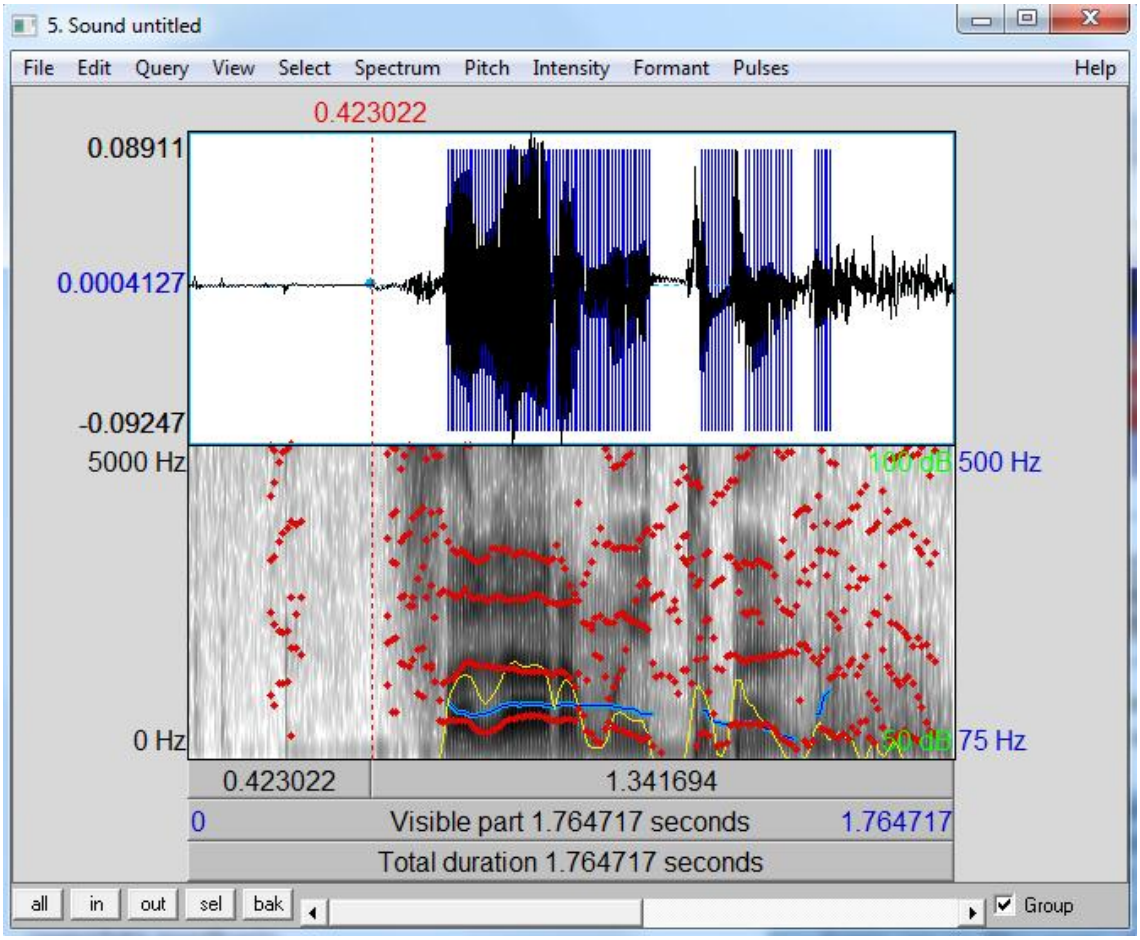

Figura 15 - Janela onde são feitas as principais medidas acústicas de um sinal de fala.

De seguida, será feita uma breve descrição do comando considerado mais importante quando se trata de um sintetizador de formantes, cujo nome é *formant*.

Este comando reúne as funções correspondentes aos formantes, como: torná-los visíveis sobre o espectrograma na forma de pontos/ linhas vermelhas, e ajustar os parâmetros para cada análise. Quando estão visíveis, tem-se a possibilidade de se obter os valores tanto das frequências dos formantes quanto de suas larguras de banda.

Antes de tudo, para se obter os resultados dos valores das frequências dos formantes e dos valores das larguras de banda, deve-se escolher a vogal ou a palavra que se pretende sintetizar. Depois, deve-se gravar o sinal ou a palavra escolhida,

utilizando a opção *record mono sound*, com a frequência da amostragem pretendida (neste trabalho foi utilizado uma frequência de amostragem de valor 22050Hz).

Depois de gravar é feita uma análise cuidadosa para que possamos ter informações corretas dos valores das frequências dos formantes e das larguras de banda.. Esses valores foram utilizados no programa de síntese feito no Matlab, e obteve-se um resultado considerado satisfatório.

3.3.2.6 Interface gráfica (Guide)

Esta é a última etapa do desenvolvimento do sintetizador de formantes para vogais. Depois de terminar o programa e após a realização de vários testes, foi utilizado a interface gráfica do Matlab (Guide), de modo a poder fazer uma interação entre o utilizador e o programa através de uma janela. Basicamente nesta janela foram criados botões e imagens que representam diferentes funções, ou seja foram criados botões que representam: vogais, frequências formantes, largura de banda, frequência fundamental, os tipos de ondas que geram o impulso, gráficos que representam a forma de sinal do trato vocal e a forma do sinal de saída da vogal escolhida.

Atribui-se assim a cada símbolo (botão) uma função particular que, ao fazer um "clique" executa a função desejada. Em seguida, é mostrado como foi criada a interface (Guide) no Matlab para esta aplicação.

Para a criação desta interface gráfica (Guide), escreve-se “guide” na linha de comando do Matlab seguido de um enter e aparece a seguinte janela:

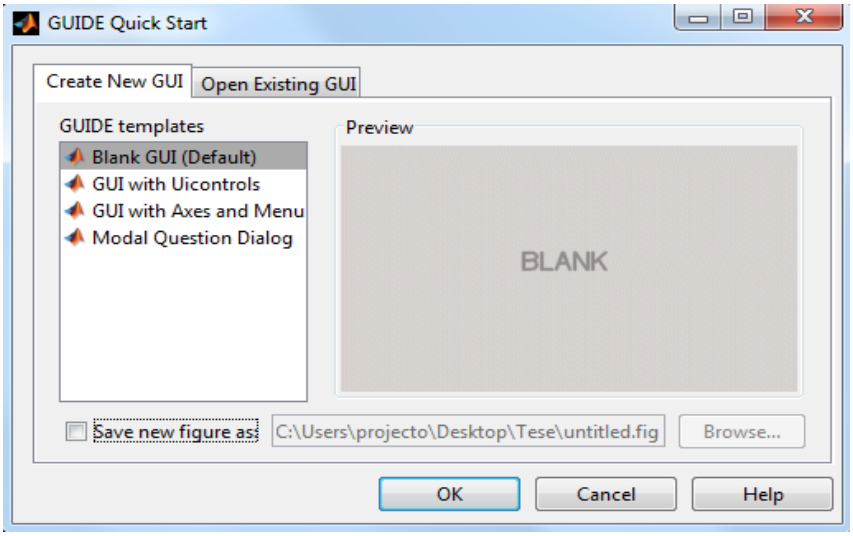

Figura 16 - Como criar a interface gráfica.

Depois escolhe-se a opção "*Blank GUI (Default)*" e pressiona-se o Botão "ok", e vai aparecer:

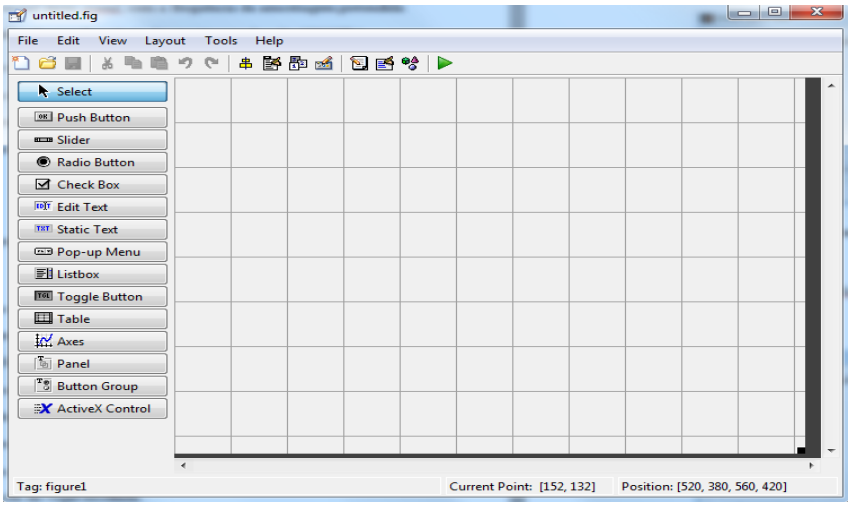

Figura 17 - Como criar a interface gráfica (cont).

Como se pode observar na figura 17, é apresentado uma janela em que no lado esquerdo apresenta uma série de botões para as configurações pretendidas. A seguir explica-se as configurações de botões e as suas funções escolhidas.

Slider – Este botão funciona como um controlador, em que os seus valores podem

variar de acordo com a necessidade do programador. O programador pode escolher o intervalo que ele pode variar. Esta configuração é feita clicando o botão direito encima do botão e escolher a opção “Property inspector”. Este botão foi utilizado para representar os valores da frequência fundamental, frequências formantes e larguras de bandas, visto que estes não possuem um valor único, os seus valores variam dentro de um intervalo. Neste trabalho a frequência fundamental (F0) varia de 100Hz a 250Hz, as frequências formantes variam de 0Hz a 4500Hz, e as larguras de banda variam entre 0Hz a 1500Hz. É de ter em consideração conforme dito anteriormente, que existem 4 formantes e 4 larguras de banda, e com a combinação dos seus valores consegue-se sintetizar as vogais.

Para interligar este botão com o programa da síntese, é preciso fazer um click com o botao direito e escolher a opção “Callback” e colocar o seguinte linha de código:

```
valor=get(handles.slider2,'value');
```

Este código faz com que seja possível atribuir aos parâmetros escolhidos do programa os valores do botão *slider*, ou seja, os botões *sliders* passam a ser os parâmetros do programa. Quando se faz a alteração do valor deste botão (clicando no botão) muda o resultado do programa automaticamente. O valor deste botão é apresentado através de um outro botão conhecido por *static text*.

Static text - Este serve para apresentar dados e textos produzidos pelo usuário, também é possível inserir numa variável o que está escrito nele.

A linha de código que permite apresentar os valores guardados no botão *slider* neste botão é:

```
set(handles.text25,'string',valor);
```

Os valores são apresentados no botão *static text* cujo nome é *text25*.

Figura 18 - Botão *Slider* com a representação do seu valor usando *Static text*.

Push button - é um botão onde a sua ação será determinada de acordo com a necessidade do utilizador, depois de selecionado por um clic produz a acção do programa. Neste caso será usado para reproduzir a leitura das vogais. Cada vogal tem um botão para ser reproduzida.

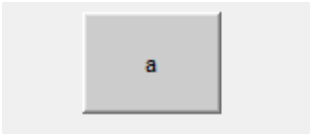

Figura 19 - Botão *Push button* representando vogal [a].

Este botão é configurado com o código do sintetizador de formantes, mais precisamente na parte da reprodução da vogal [a], na reprodução da outra vogal será configurado o botão correspondente.

Axes - O axes é uma ferramenta utilizadas para a construção de gráficos, o programador pode inserir vários gráficos ao mesmo tempo, mas para se referir a cada axes basta digitar a seguinte linha de código no M-file gerado pela interface:

```
axes(handles.nome do gráfico);
```

No presente trabalho, foram utilizados dois destes botões, um foi utilizado para representar a função de transferência do trato vocal, e o outro para representar o sinal de

saída.

Pop-up Menu - São normalmente utilizados para apresentar uma lista de opções mutuamente exclusivas para o usuário. O *Pop-up Menu* exibe sempre a opção escolhida.

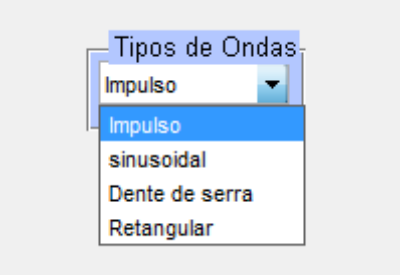

Figura 20 - Botão *Pop-up Menu*.

Para produzir um som, foram utilizadas as 4 funções, conforme mostra a figura 20, função de geração de impulsos, função que representa uma onda sinusoidal, função que representa uma onda retangular e uma função que representa uma onda dente de serra. Foram testados as quatros funções e obtiveram-se resultados diferentes.

O resultado final da interface gráfica é apresentado na figura 22. Procurou-se utilizar um interface didáctica semelhante à apresentado no software KTH referido na figura 21.

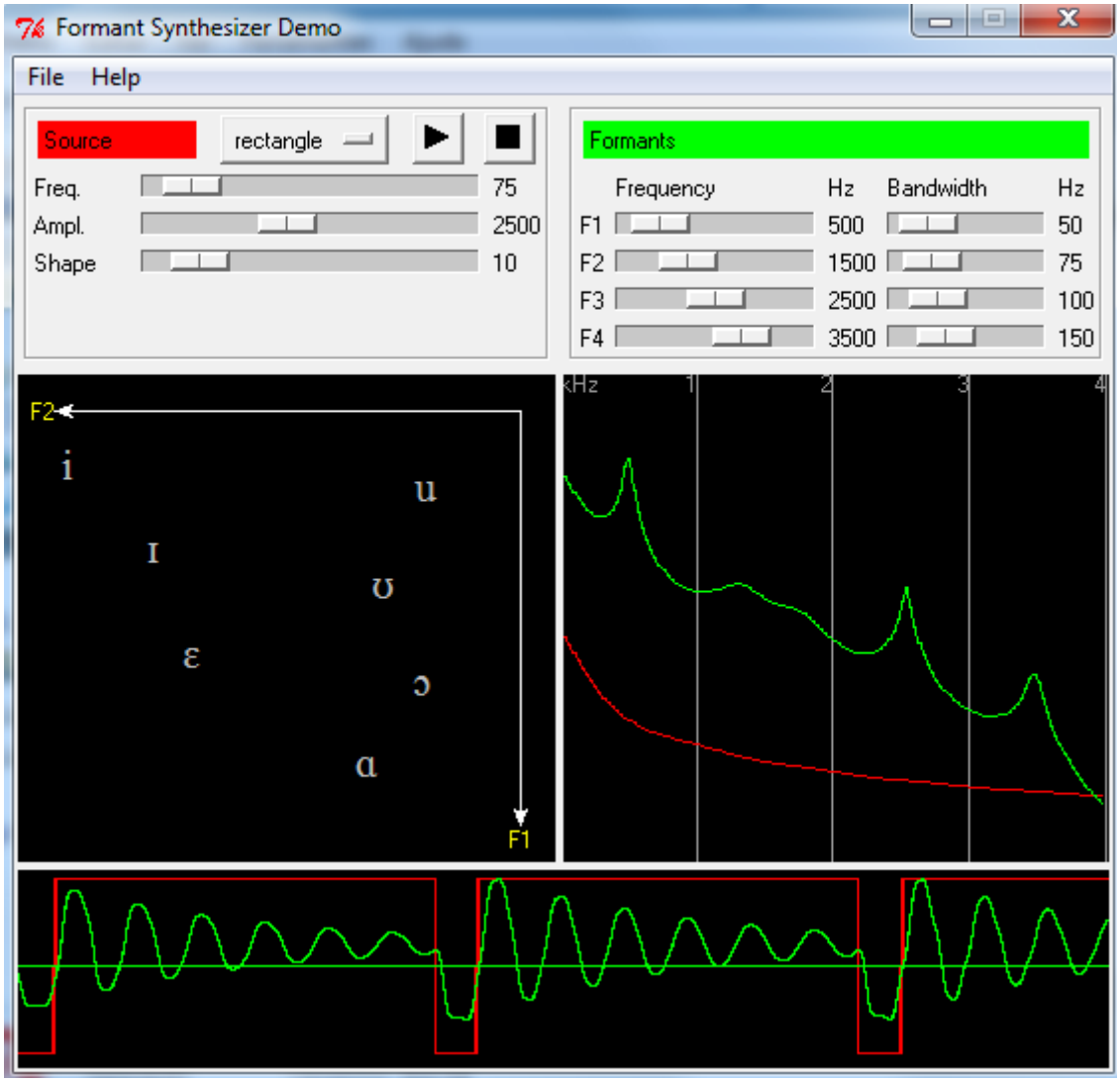

Figura 21 - Software de KTH

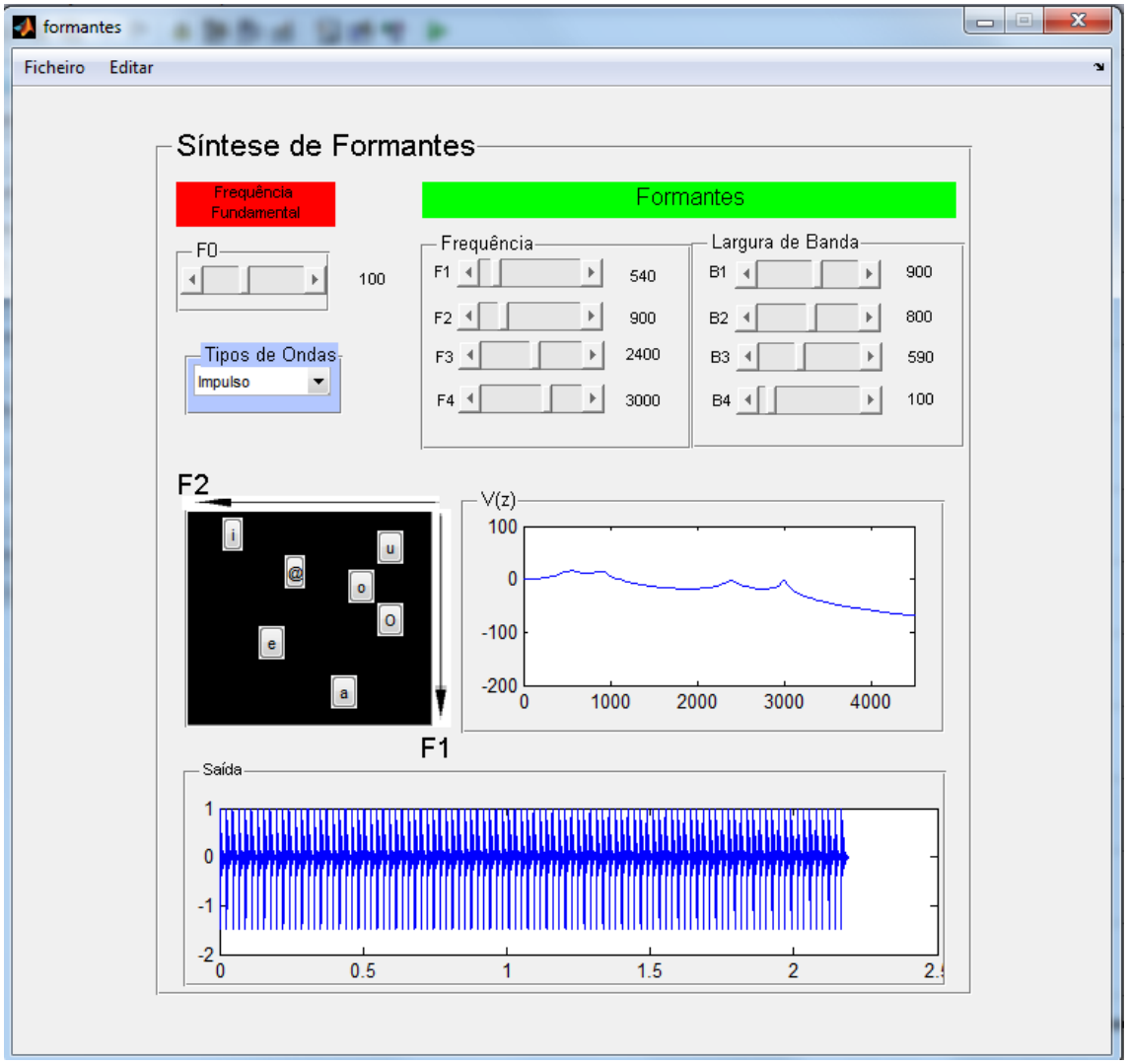

Figura 22 - Interface gráfica criada no Guide. A representação da vogal [i].

A janela acima representa a informação relativa à fonte (F0) e do tipo de sinal de excitação na parte superior lado esquerdo, abaixo plano dos formantes F1-F2 para a vogais. Ali, pode ser visto o conhecido triângulo dos vogais. No lado superior direito está a informação sobre o trato vocal ou seja, os formantes e as suas respostas em frequência.

Fez-se a experiência de vários vogais, alterando os valores levemente dos formantes e larguras de bandas, verificando assim a alteração no trato vocal. Com a alteração dos valores da frequência fundamental (F0) nota-se uma entoação diferente

para cada vogal, visto que para a F0 não foi utilizado um valor fixo, mas sim um intervalo de valores que variam de 100Hz a 250Hz, com a ideia de reproduzir as vogais com diferentes valores de frequência fundamental.

Como se pode observar na figura 21, a representação das vogais foi feita utilizando o código SAMPA. De seguida apresenta-se uma tabela com exemplos de alguns exemplos deste código.

Tabela 5 – Exemplo de código SAMPA (retirada de (Teixeira, 95)).

Identificação de vogais, consoantes	Símbolo sampa	Palavras portuguesas	Tanscrição em sampa
Vogais	6	cama	k6m6
	a	cara	kar6
	e	pêra	per6
	E	sete	sEt@
	@	que	k@
	i	Fita	fit6
	o	dou	do
	O	corda	kOrd6
	u	mudo	mudu
	6~	manta	m6~t6
	e~	menta	me~t6
	i~	pinta	pi~t6
	o~	ponta	po~t6
	u~	mundo	mu~du
Consoantes	w	pau	paw
	j	pai	paj
	w~	cão	k6~w~
	j~	mãe	m6~j~
	p0,p	pai	p0paj
	t0,t	tia	t0ti6
	k0,k	casa	k0k6za
	b0,b	bar	b0bar
	d0,d	data	d0dat6
	g0,g	gato	g0gatu
	f	ferias	fErj6s
	s	selo	selu
	S	chave	Sav@
	v	vaca	vak6
	z	azul	azul~
	Z	agir	aZir

Para a pausa foi utilizado o símbolo "_".

SAMPA (Speech Assessment Methods Phonetic Alphabet) é um sistema de escrita fonética legível por computadores que usa os caracteres ASCII de 7 bits. É baseado no alfabeto fonético internacional (IPA), criado como uma alternativa para resolver a incapacidade de codificações de texto para representar símbolos IPA (SAMPA, 1999-2011).

3.4 Resultados

Com base no programa implementado para análise e síntese das vogais, alcançou-se um dos objetivos do trabalho, a sintetização das vogais. A síntese foi obtida a partir dos parâmetros encontrados através da análise acústica do sinal original, previamente gravado. Apresenta-se de seguida os gráficos com a resposta em frequência do trato vocal para cada vogal.

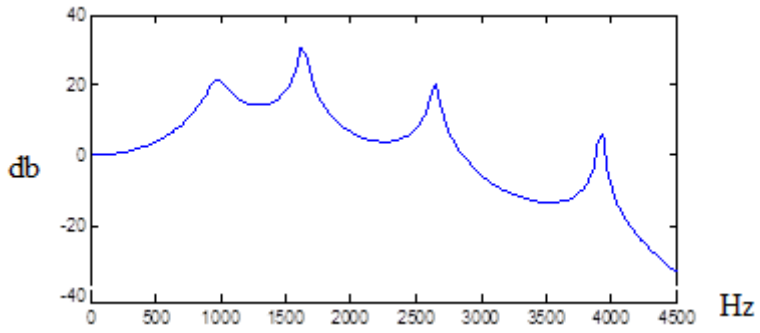

Figura 23 - Função de transferência do trato vocal da vogal [a].

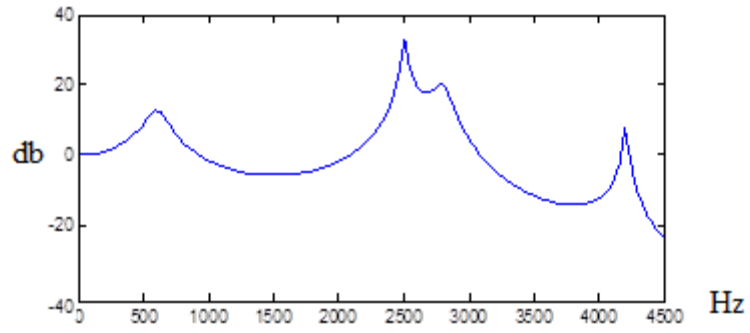

Figura 24 - Função de transferência do trato vocal da vogal [e].

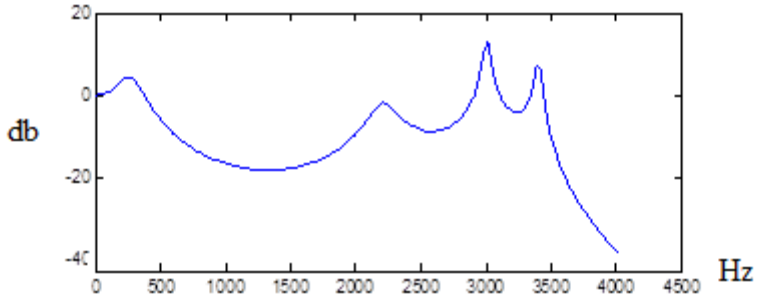

Figura 25 - Função de transferência do trato vocal da vogal [i].

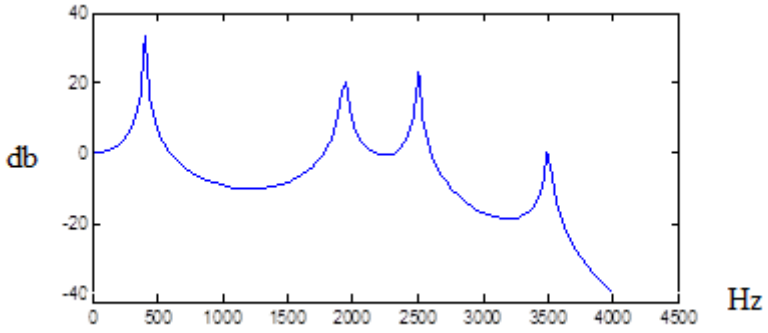

Figura 26 - Função de transferência do trato vocal da vogal [@].

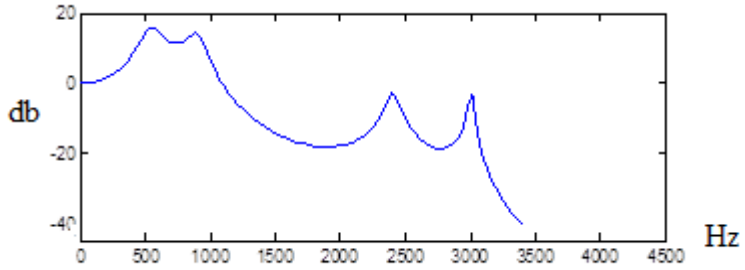

Figura 27 - Função de transferência do trato vocal da vogal [O].

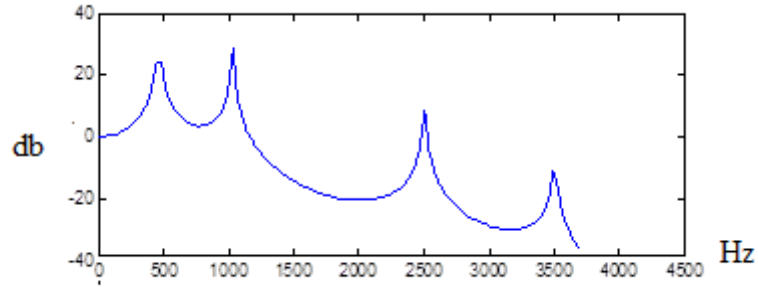

Figura 28 - Função de transferência do trato vocal da vogal [o].

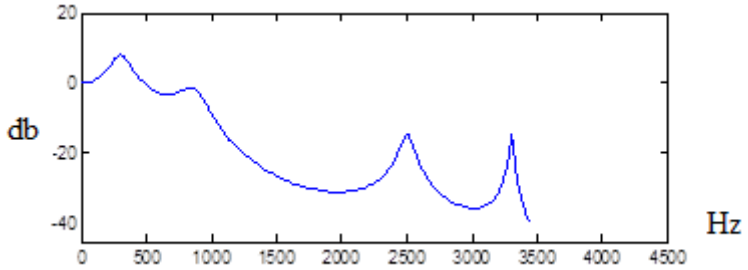

Figura 29 - Função de transferência do trato vocal da vogal [u].

Os valores das frequências formantes e das larguras de banda (expresso em Hz) utilizados para sintetizar os vogais são apresentados na tabela 7. Com estes resultados notou-se, na prática, que as vogais sintetizadas apresentaram uma proximidade satisfatória em comparação com as vogais originais (vogais pronunciadas por homem).

Tabela 6 - Valores das frequências de formantes e das larguras de bandas utilizadas para sintetizar as vogais.

Vogais	F1	F2	F3	F4	B1	B2	B3	B4
a	967	1631	2644	3918	1000	300	328	100
e	600	2500	2800	4200	1010	190	730	100
i	274	2600	3000	3400	1200	1300	200	100
@	402	1932	2500	3500	50	75	100	150
O	540	900	2400	3000	900	800	590	100
o	460	1028	2500	3500	50	75	100	150
u	305	867	2500	3300	900	1200	500	100

CAPÍTULO 4
CONVERSÃO FONEMA-FALA

4 CONVERSÃO FONEMA-FALA

4.1 Introdução

Neste capítulo descreve-se o programa desenvolvido para satisfazer o segundo objetivo do trabalho que consiste na conversão fonema-fala, ou seja um programa capaz de fazer uma leitura a partir da informação contida na base de dados. Estas informações não são mais do que as sequências de frequências formantes e larguras de bandas, visto que, esta conversão baseia-se na síntese de formantes referida no capítulo 3. A grande diferença entre o capítulo 3 e este capítulo, ou por melhor dizer, entre utilizar as informações das frequências formantes e larguras de banda é que para sintetizar as vogais utiliza-se um vector com 4 valores de frequências de formantes e um vector com 4 valores de larguras de banda, já na conversão fonema-fala utiliza-se uma matriz que consiste numa sequência de linhas com esses 8 parâmetros (4 formantes e respetivas larguras de banda. Cada linha corresponde a um segmento com uma duração de cerca de 10 ms. O tamanho desta matriz depende da duração de cada difone. Estes parâmetros também são retidas do Praat após a gravação de um sinal de fala (com uma frequência de amostragem de 22050Hz).

É muito importante ter em atenção que aqui, o trato vocal é considerado invariante

no tempo a cada 10ms, ao contrário do que foi usado na síntese dos vogais (sempre invariante no tempo). Utilizou-se uma frequência fundamental (F0) com um valor fixo igual a 100Hz, contudo posteriormente será possível usar uma curva de variação de F0 para produzir fala sintetizada com entoação. O sinal é obtido a partir da geração de um trem de impulsos.

Para esta conversão também foi utilizado a interface gráfica (Guide), com o intuito de fazer uma interação entre o utilizador e o programa através de uma janela.

Nesta janela o utilizador introduz uma palavra/frase e o sintetizador por sua vez faz a leitura e apresenta um gráfico que representa o respectivo sinal de fala e o seu espectrograma.

4.2 Desenvolvimento

A implementação da síntese de formantes para a conversão fonema-fala baseou-se na utilização do mesmo modelo de geração de fala sintética referido no capítulo 3, utilizando as mesmas funções de transferência.

Os parâmetros (representados em Matrizes) das frequências formantes e larguras de bandas retiradas do Praat variam de difone para difone. Uma das formas de obter esses parâmetros é gravar um sinal de fala (neste caso um frase) e dividir os difones cuidadosamente de maneira a poder concatena-los da melhora maneira possível uns com os outros no sintetizador (ou seja, os valores terminais de um difone serão próximos dos iniciais do difone seguinte). A obtenção da correta matriz dos difones exige muita atenção e muito cuidado e nem sempre os valores obtidos num difone são adequados, sendo assim deve-se fazer inúmeras gravações e divisões até que apresente um resultado mais aceitável na síntese. A análise do sinal de fala realizada no programa Praat produz bons resultados para os parâmetros das frequências formantes, já para os valores das

larguras de banda (B1, B2, B3 e B4) apresenta apenas um valor da média para cada largura de banda. Assim, deve ser feito um estudo do tamanho da matriz, visto que o comprimento da matriz de largura de banda deve ser igual ao comprimento da matriz das frequências formantes.

De seguida, por passos, apresenta-se um exemplo de uma conversação fonema-fala para a palavra "Aula" reassentado em código sampa como (Aul6).

Grava-se então um sinal de fala "*A ultima fala da laura*" no Praat, depois faz-se a divisão dos difones, conforme mostra a figura 30.

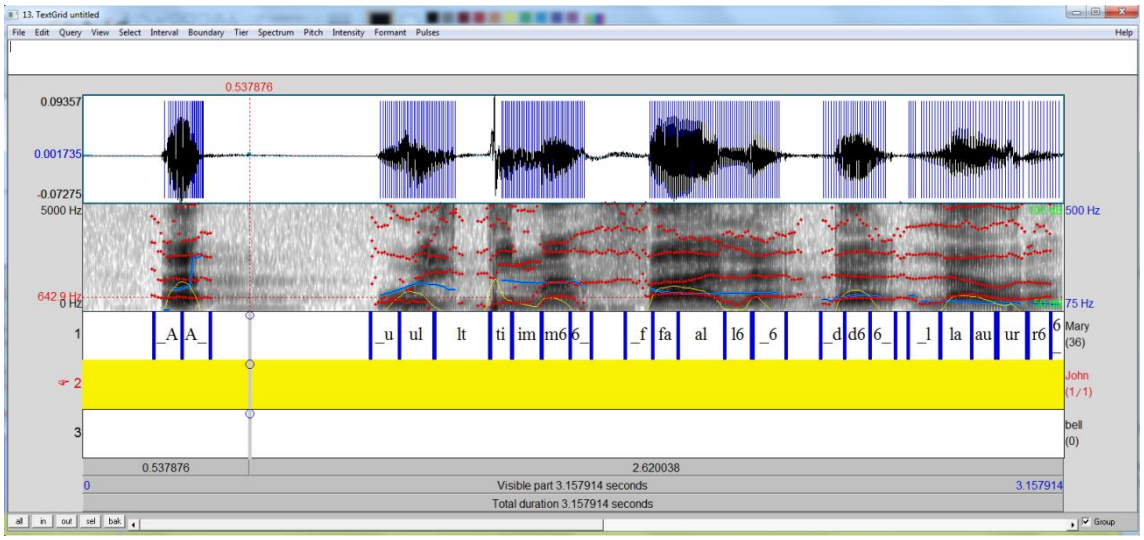

Figura 30 - Divisão dos difones

A função geradora de impulsos (fgerimpulso2.m) gere impulsos em segmentos de comprimento finitos de maneira que o próximo segmento comece no ponto seguinte ao que terminou o segmento anterior para não causar descontinuidades na amplitude do sinal do impulso ao longo da sequência de segmentos. Para a palavra aul6 segue-se a seguinte sequência de difones: _a (pausa seguida de a aberto), au, ul, l6 e 6_ (6 seguido de pausa). De acordo com várias pesquisas e após vários testes realizados concluiu-se que a melhor duração de segmentos para este caso é de 10ms. Para cada difone temos

uma matriz de formantes e outra de larguras de banda. Para um completo funcionamento do sistema é necessário criar uma base de dados dos difones. Neste exemplo, para cada difone é guardado nesta base de dados os parâmetros das frequências formantes e das larguras de banda. Por exemplo para o difone "_A", os seguintes parâmetros são guardados e apresentam-se nas tabelas e 8 e 9.

Tabela 7 - Valores das frequências formantes do difone _A (pausa seguida de a aberto).

Frequências formantes do difone _A				
Tempo	F1	F2	F3	F4
0.202	843	1276	2637	3472
0.212	809	1289	2676	3422
0.222	804	1293	2661	3414
0.232	807	1281	2640	3416
0.242	802	1266	2603	3339
0.252	805	1259	2610	3331
0.262	815	1243	2626	3346

Tabela 8 - Valores das larguras de banda do difone _A.

Larguras de banda do difone _A				
Tempo	B1	B2	B3	B4
0.202	158	241	175	412
0.212	158	241	175	412
0.222	158	241	175	412
0.232	158	241	175	412
0.242	158	241	175	412
0.252	158	241	175	412
0.262	158	241	175	412

Para gerar o código, cria-se uma outra função no Matlab de modo a poder fazer a

leitura das informações dos difones. Para tal usou-se a função fgets do Matlab.

No Guide, criou-se um janela de interface onde temos, um campo para introduzir uma palavra ou uma frase utilizando o botão *Edit Text* , configurado pelo seguinte código:

```
s=get(handles.edit1,'string');
```

E outros dois campos de representação dos sinais, em que uma representa um sinal de saída e outro representa o espectrograma deste sinal. Usou-se o comando *axes* para fazer estas representações:

```
axes(handles.axes1);
plot(sinalfinal);
axes(handles.axes2);
NFFT=512;
spectrogram(sinalfinal,NFFT,NFF
T/2,NFFT,Fa,'yaxis');
V=axis;
axis([V(1) V(2) 0 4000]);
```

Para que um utilizador venha utilizar este trabalho, e de modo a simplificar a sua compreensão cria-se uma base de dados dos difones onde permite verificar os difones disponíveis para a conversão fonema-fala. A representação dos difones é feita usando o código sampa.

4.2.1 Base de dados

A criação de uma base de dados para um sistema de síntese de fala é uma tarefa muito complicado, visto que é preciso abranger todas as palavras do dicionário da língua em causa.

A base de dados será constituída por 37 unidades base.

Neste trabalho a unidade utilizada para a criação da base de dados é o difone.

A base de dados será uma matriz de 1369 difones (37^2).

Cada difone contém uma matriz de frequência formantes e outra de largura de banda guardada como um ficheiro de extensão .txt que será chamado depois por uma função.

Como trabalho futuro recomenda-se completar esta base de dados, visto que contém apenas 5% dos difones, falta completar os restantes 95%. Em anexo apresenta-se a matriz de difones, de modo a que o utilizador possa saber quais difones estão prontos a serem utilizados.

De seguida é apresentado o resultado final desta aplicação numa janela de interface gráfica conforme mostra a figura 31.

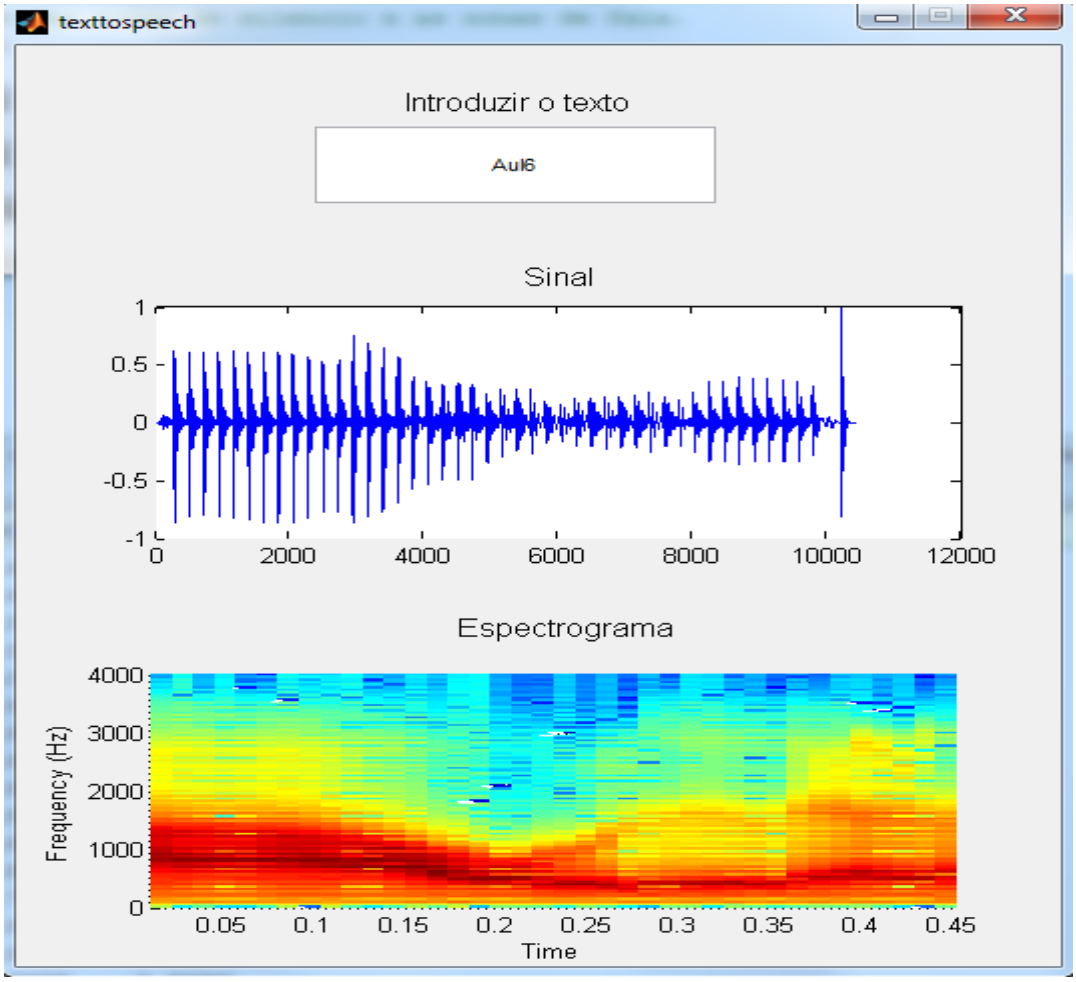

Figura 31 - Representação final da interface gráfica da conversão fonema-fala com exemplo da palavra Aul6.

CAPÍTULO 5
CONCLUSÕES E DESENVOLVIMENTOS FUTUROS

5 CONCLUSÕES E DESENVOLVIMENTOS FUTUROS

Apresenta neste capítulo de forma resumida as conclusões do trabalho desenvolvido. Pretende-se também deixar um alerta para as linhas de investigação que ficam em aberto neste trabalho e fazer um apontamento de algumas aplicações ligadas ao tema desenvolvido.

5.1 Conclusões

Esta tese teve como objetivo fazer um estudo da análise dos sinais de fala e a extração dos parâmetros para o desenvolvimento de um sintetizador de formantes para duas aplicações (sintetização das vogais e conversão fonema-fala).

Através da realização desta tese foi possível compreender o mecanismo básico de produção de fala humana. Após entender o processo de gerar um sinal de fala pelo ser humano, foi feito um estudo da síntese de fala e quais as técnicas actuais utilizadas. Dessas técnicas optou-se pela síntese de formantes (que foi o objetivo), que segundo vários estudos é considerada mais rápida quando comparada com algumas outras, porém o som produzido é mais metálico e robótico.

O modelo fonte-filtro utilizado possibilitou o desenvolvimento de um algoritmo capaz de produzir um sinal de excitação, que foi posteriormente equalizado com o uso de filtros digitais.

A fonte de excitação consiste num sinal periódico de impulsos glotais (gerados através de uma função) para os sinais vocalizados, um sinal de ruído para os sinais não vocalizados e um sinal composto pela sobreposição dos dois anteriores para sinais de excitação mista (exemplo: consoantes fricativas sonoras – v, j, z). O modelo do trato vocal toma também configurações diferentes consoante o tipo de sinal. Para um sinal vocalizado apresenta um comportamento diferente de sinal não vocalizado.

Na primeira parte, a do desenvolvimentos de um sintetizador de formantes para vogais, o resultado da tese foi o desenvolvimento de um algoritmo que permite alterar os valores das frequências formantes larguras de banda, frequência fundamental acordo com o sinal final que se deseja reproduzir.

Na parte da conversão fonema-fala, o sintetizador exige que a extração dos parâmetros seja muito bem feita para um bom funcionamento. Visto que os parâmetros das larguras de banda retiradas do Praat não eram completos teve-se que trabalhar manualmente estes parâmetros porque num sintetizador de formantes apresentam uma enorme importância na produção da fala.

Os dois sistemas desenvolvidos tiveram uma completa realização estando a funcionar corretamente e com resultados de qualidade razoáveis e que podem ser ouvidos.

No caso do sintetizador fonema-fala, a base de dados de formantes está incompleta e necessita de ser completada para que o sintetizador possa sintetizar qualquer som do Português.

5.2 Desenvolvimentos futuros

Como sugestão para trabalhos futuros pode ser citado o desenvolvimento de um sintetizador para as consoantes. Pode-se ainda completar o trabalho com mais

aplicações na interface gráfica.

Na parte de conversão fonema-fala sugere-se a expansão da base de dados incluindo todos os difones do português de Portugal (os restantes 95%) não só em termos de números de unidades mas também no número de realização de cada unidade.

Propõe-se ainda à base de dados a realização em sílabas tónicas, sempre que possível em sílabas atónicas e com diferentes durações.

Também pode ser desenvolvida neste projeto todo o processamento linguístico-prosódico para permitir que a entrada seja texto e não fonema, tornando o sistema num sintetizador de fala.

Sugere-se em geral um aperfeiçoamento do presente trabalho, como o ajuste dos filtros ressonadores, a inclusão de ganhos no sinal sintetizado para que a atenuação sofrida pelo sinal seja compensada, etc.

REFERÊNCIAS BIBLIOGRÁFICAS

6 REFERÊNCIAS BIBLIOGRÁFICAS

Albuquerque, A. S., (2001). "*Análise Comparativa dos Métodos de Sintetização de Voz*". Dissertação de Mestrado, Universidade Federal de Santa Catarina, Florianópolis.

Atal, B.. (1986). "*High-quality speech at low bit rates: multi-pulse and stochastically excited linear predictive coders - Acoustics, Speech, and Signal Processing*", IEEE International Conference on ICASSP'86, pages: 1681-1684.

Barros, M. J., (2002). "*Estudo Comparativo e Técnicas de Geração de sinal para Síntese de Fala* ." Dissertação de Mestrado, Faculdade de Engenharia da Universidade do Porto.

Braga, D.; Mato, X., (2007). "*Algoritmos de Conversão Grafema-Fonema em Galego para Sistemas de Conversão Texto-Fala.*" Facultade de Filoloxía da Universidad da Coruña. Campus de Zapateira.

Brasil Escola., (2012). "*Análise Sintática e Análise Morfológica*." Acedido em 24 de Outubro de 2012, em:http://www.brasilescola.com/gramatica/analise-sintatica-analise-morfologica.htm

Chbane, D., (1994). "*Desenvolvimento de Sistema paraConversão de Textos em Fonemasno Idioma Português*". Dissertação de Mestrado, Universidade de São Paulo – São Paulo.

Cunha, M., (2011). "*Variação Acústica das Vogais Orais de Crianças no Português Europeu*". Dissertação de Mestrado, Universidade de Aveiro – Aveiro.

Fonseca, A., (2009). "*Análise do Tutorial do programa de análises acústicas Praat*." Universidade Estadual de Campinas, Faculdade de Letras, SP, Brasil. vol.1 N.o 2 vol. 1, 2009.

Klatt, Dennis H. - *"Software for a cascade/ parallel formant Synthesizer"*. J. Acoust. Soc. Am. 67, Massachussets of Technology, Cambridge, Massachussetts 02139, March 1980, pp. 971 a 995.

Klompje, G., (2006). "*A Parametric Monophone Speech Synthesis System*." In LREC 2006.

Lemmetty, S., (1999). "*Review of Speech Synthesis Technology*." MAster's Thesis, Helsinki University of Technology - Espoo.

Meneses, C., (2008). "*Tecnologias de fala*". Departamento de Engenharia de Electrónica e Telecomunicações e Computadores.

Oliveira C., (2009). "*Contributos Linguísticos para um Sistema de Síntese de Base Articulatória*." Dissertação do Doutoramento, Universidade de Aveiro.

Pacheco, F., (2010). "*Sistemas de Síntese de Fala*." Revista Ilha Digital, ISSN 2177-2649, vol. 2, páginas 3 – 17, 2010.

Paiva, S., (2005). "*Síntese por concatenação de variantes regionais: falar do Porto*." Dissertação de Mestrado, Universidade de Aveiro – Aveiro.

Pierrehumbert, J. B.. (1980). "*The Phonology and Phonetics of English Intonation*." PhD thesis, Massachusetts Institute of Technology.

Rabiner, L. R.; Schafer, R. W., (1978). "*Digital Processing of Speech Signal*." Prentice-Hall Signal Processing Series 1978.

SAMPA - Computer Readable Phonetic Alphabet - http://www.phon.ucl.ac.uk/home/sampa/. University College London, 1999-2011.

Saraswathi, S., (2010). "*Design of Multilingual Speech Synthesis System*." Academic journal article from Intelligent Information Management, Vol. 2, No. 1.

Schroeder, M.; Atal, B. (1985). "*Code-excited linear prediction (CELP): High-quality speech at very low bit rates - Acoustics, Speech, and Signal Processing*". IEEE International Conference on ICASSP '85, pages: 937-940.

Simões, F.O., (1999). "*Implementação de um sistema de conversão texto-fala para português do Brasi*l." Dissertação de Mestrado, Faculdade de Engenharia Elétrica e de Computação - Campinas.

Sproat, Richard W. (1997). "*Multilingual Text-to-Speech Synthesis*": The Bell Labs Approach. Springer.

Teixeira, J. P., (1995). "*Modelização Paramétrica de Sinais para Aplicação em Sistemas de Conversão Texto-Fala*." Dissertação de Mestrado, FEUP – Porto.

Teixeira, J. P., (2004). "*A Prosody Model to TTS Systems*." Ph.D. Dissertation FEUP – Porto.

Teixeira, J. P., 2012. "*Prosody Generation Model for TTS Systems – Segmental Durations and F0 Contours with Fujisaki Model*". LAMBERT Academic Publishers.

Teixeira, J. P.; Barros, M. J. and Freitas, D., (2003). "*Sistemas de Conversão Texto-Fala*." Procedings of CLME, Maputo.

Teixeira, J. P.; Freitas, D.; Gouveia, Paulo D.F.; Olaszy, Gabor; Nemeth, G. (1998). "*Multivox: conversor texto fala para português*". In III Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada. Porto Alegre, Brasil. p.88-98

Thomas, S., (2007). "*Natural Sounding Text-To-Speech Synthesis Based on Syllable-like Units*." Master Thesis, Department of Computer Science and Engineering Indian Institute of Technology Madras.

Tokuda, K.; Heiga Zen; Black, A.W (2002). "*An HMM-based speech synthesis system applied to English*." Proceedings of 2002 IEEE Workshop on Speech Synthesis. Pages 227-230.

ANEXO

MATRIZ DE DIFONES

	_	6	a	e	E	@	i	o	O	u	6~	e~	i~	o~	u~	W	j	W~	j~	p_p	t_t	k_k	b_b	d_d	g_g	f	s	S	v	z	Z	m	n	j	l	l~	L	R	r
_			_a		_E				_O																										_l				
6																																						
a																																						
a_							ai			au																						ne		al				
6_																																						
6																																						
E										E_u																							E_l					
@						@																																
@_																																						
i																																						
i_							ii_																									in						
o																																						
o_																																						
O																																on		ol				
u			u_																													un		ul				
u_																																						
6~																																						
e~																																						
i~																																						
o~																																						
u~																																						
W																																						
j																																						
W~																																						
j~																																						
p_p																																						
t_t																																						
k_k																																						
b_b																																						
d_d																																						
g_g																																						
f																																						
s																																						
S																																						
v																																						
z																																						
Z																																						
m																																						
n			n6_		nE_	n@_																																
j																																						
l			l_l				l_E			l_u																												
l~																																						
L																																						
R																																						
r																																						