Universidade Federal da Paraíba
Centro de Ciências e Tecnologia
Curso de Pós-Graduação em Engenharia Elétrica

Processamento Linguístico para um Conversor Texto-Fala para a Língua Portuguesa

Fabrícia Abrantes de Figueiredo

Campina Grande – PB
Setembro – 1998

Fabrícia Abrantes de Figueiredo

Processamento Linguístico para um Conversor Texto-Fala para a Língua Portuguesa

Dissertação submetida ao corpo docente da Coordenação dos Cursos de Pós-Graduação em Engenharia Elétrica da Universidade Federal da Paraíba – Campus II como parte dos requisitos necessários para obtenção do grau de *Mestre em Engenharia Elétrica.*

ÁREA DE CONCENTRAÇÃO: ***Processamento Digital de Sinais***

Benedito Guimarães Aguiar Neto
Orientador

Campina Grande, Paraíba, Brasil
Fabrícia Abrantes de Figueiredo, Setembro/1998

F475p Figueiredo, Fabrícia Abrantes de.
Processamento linguístico para um conversor texto-fala para a língua portuguesa / Fabrícia Abrantes de Figueiredo. - Campina Grande, 1998.
95 f.

Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal da Paraíba, Centro de Ciências e Tecnologia, 1998.
"Orientação : Prof. Dr. Benedito Guimarães Aguiar Neto".
Referências.

1. Processamento de Sinais. 2. Processamento Linguístico - Conversor Texto. 3. Fala - Língua Portuguesa. 4. Dissertação - Engenharia Elétrica. I. Aguiar Neto, Benedito Guimarães. II. Universidade Federal da Paraíba - Campina Grande (PB). III. Título

CDU 621.391(043)

PROCESSAMENTO LINGUÍSTICO PARA UM CONVERSOR TEXTO-FALA PARA A LÍNGUA PORTUGUESA

FABRÍCIA ABRANTES DE FIGUEIREDO

Dissertação Aprovada em 01.09.1998

PROF. BENEDITO GUIMARÃES AGUIAR NETO, Dr.-Ing., UFPB
Orientador

PROFA. MARIA DE FÁTIMA QUEIROZ VIEIRA TURNELL, Ph.D., UFPB
Componente da Banca

PROF. EDILSON FERNEDA, Dr., UFPB
Componente da Banca

PROFA. LÍRIDA ALVES DE BARROS NAVINER, Dr., ENST-França
Componente da Banca

CAMPINA GRANDE - PB
Setembro - 1998

Dedico este trabalho a

Vicente Francisco de Figueiredo e

Maria Abrantes de Lima Figueiredo

Agradecimentos

A realização deste trabalho recebeu o apoio de muitas pessoas. É chegado a hora de agradecer, muito embora citar nomes seja uma tarefa difícil: é quase impossível lembrar, imediatamente, de todos. Assim, de uma forma mais particular, devo os meus sinceros agradecimentos:

A ***Deus***, que sempre me deu forças para concluir esta grande caminhada;

Ao professor ***Benedito Guimarães Aguiar Neto***, pela orientação e compreensão prestados;

À professora ***Lírida Alves Barros Naviner***, pelo estímulo, esforço e orientação, fundamentais ao desenvolvimento deste estudo;

A ***Alexandro Vladno***, meu noivo, presente na etapa de conclusão, pelo incentivo, compreensão e sugestões;

Aos irmãos ***Giovannini (Evânia)*** *e* ***Jason (Adriana)*** pelo incentivo. Em especial, a ***Jorge (Adriana)*** pelas suas idéias e sugestões;

Aos "amigos de verdade" – ***Adelmar*** *e* ***Edmar*** – que demonstraram, acima de tudo, o real significado da amizade;

A ***Leonel***, ou melhor, ao meu "co-orientador", que tanto me ensinou quanto me incentivou no decorrer deste trabalho;

A ***Carminha***, que sempre acreditou em mim;

A ***Ângela***, sempre incansável no desempenho de suas responsabilidades;

A ***Pedro***, pela atenção, ajuda e paciência;

A amiga ***Márcia Mary*** e ***Isabella Mendonça***, pelos comentários, compreensão e paciência;

Aos colegas ***Luciana***, ***Joseana***, ***Suely***, ***Madeiro***, ***Rosângela***, ***Hélcio***, ***Roberto***, ***Ricardo***, ***Paulo Márcio***, ***Rinaldo*** e demais colegas do *LAPS*;

Aos colegas da *TELERN*, ***Afrânio***, ***Eduardo*** e ***Nildeives***, pela contribuição nos testes do Sistema.

Resumo

Devido as suas caraterísticas e aplicações, os *Sistemas de Resposta por Voz* (***SRV***) têm se tornado importantes nos últimos anos. Os ***SRV*** podem ser classificados em *Sistema de Voz Armazenada* (***SVA***) e *Sistemas de Conversão Texto-Fala* (***SCTF***).

Os ***SVA*** e os ***SCTF*** apresentam uma diferença principal. Os primeiros originam a fala a partir de palavras previamente gravadas e armazenadas. Os últimos empregam uma técnica que produz a voz a partir da concatenação de unidades acústicas - polifones-, viabilizando, portanto, a implementação de todas as possíveis palavras de uma língua.

O ***SCTF*** apresenta dois módulos principais, ou seja, *processamento linguístico* - responsável por caracterizar o idioma - e *processamento do sinal* - responsável pela síntese sonora. O principal objetivo deste trabalho consiste em determinar os módulos integrantes do processamento linguístico capazes de reproduzir uma fala natural e inteligível para o português falado no Brasil.

Abstract

Due to its characteristics and applications *Voice Response Systems* (***VRS***) have grown in importance in the past. ***SRV*** can be classified as *Stored Voice Systems* (***SVS***) *and Text-to-Speech Conversion Systems* (***TSCS***).

SVS and ***TSCS*** pose a major difference: the former originates speech from words previously recorded and stored. The latter employs a technique to produce voice by concatenation of acoustic units, allowing the implementation of the entire set of words of a language. ***TSCS*** present two main modules: linguistic processing that characterizes an idiom and a signal processing unit responsible for sound synthesis.

The main objective of this work is to formulate a ***TSCS*** structure, considering the Portuguese language. This structure must be able to reproduce a natural and comprehensible speech for the Portuguese language spoke in Brazil.

Sumário

Lista de Quadros

Lista de Figuras

Lista de Símbolos

1. f_o – freqüência fundamental;
2. $x_m(n)$ – sinais elementares de análise;
3. t_m – marcas de análise;
4. $h_m(n)$ – janela de Hamming;
5. $x_q(n)$ – sinais elementares de síntese;
6. t_q – marcas de síntese;
7. $x(n)$ – sinal de síntese.

Lista de Abreviaturas

1. SRV – Sistemas de Resposta por Voz,
2. SVA – Sistema de Voz Armazenada;
3. SCTF – Sistema de Conversão Texto-Fala;
4. LEIA – Leitura Eficiente Inteligível e Automática
5. LPC – Linear Predictive Coding (Codificação por Predição Linear);
6. TD-PSOLA – Time-Domain Pitch Synchronous OverLap-Add;
7. MBR-PSOLA – Multi-Band Re-Synthesis Pitch-Synchronous OverLap-Add;
8. PCM – Pulse Code Modulation;
9. ADPCM – Adaptative Differential Pulse Code Modulation;
10. SBC – Sub-Band Coding;
11. V-SELP – Vector-Sum Excited Linear Predictive;
12. AFI – Alfabeto Fonético Internacional;
13. C – Consoante;
14. V – Vogal.

CAPÍTULO 1

INTRODUÇÃO

Este capítulo apresenta, os ***Sistemas de Resposta por Voz****, em especial os* ***Sistemas de Conversão Texto-Fala****, destacando suas aplicações e sua estrutura geral. São citados também os objetivos e os fundamentos deste trabalho.*

1.1 Sistemas de Reprodução de Mensagens da Fala

A linguagem escrita é a maneira mais eficiente e confiável de se transmitir e armazenar informações, além de ser a mais econômica. Por outro lado, vale ressaltar que, através da fala as informações podem ser obtidas e fornecidas mais rapidamente e de forma bastante flexível

e confortável [1]. Entretanto, a escolha do tipo de linguagem a se utilizar dependerá da aplicação, pois em muitas situações a escrita não é o meio mais adequado para acessar informações.

Algumas vezes, no processo de comunicação homem/máquina, a exemplo de mensagens de alarme ou de advertência, a forma falada pode ser bem mais interessante do que a escrita. São situações em que se deseja que as mãos e os olhos estejam livres para que outras tarefas sejam realizadas. Além disso, mensagens faladas possibilitam o acesso a sistemas automatizados de informação, através da rede de telefonia pública, sem que seja preciso o uso de equipamentos especiais.

Portanto, percebe-se que a utilização de mensagens faladas apresenta algumas vantagens com relação às mensagens escritas. Entretanto, sistemas de comunicação vocal homem/máquina ainda não são amplamente usados, pois dependem, sobretudo, das técnicas digitais de armazenamento e reprodução de voz, as quais encontram-se em desenvolvimento.

Dentre os sistemas de reprodução de mensagens da fala os ***Sistemas de Resposta por Voz (SRV)*** [2, 3] apresentam várias vantagens, pela flexibilidade e maior capacidade de compressão. Esses sistemas podem ser divididos em dois grupos: ***Sistema de Voz Armazenada (SVA)*** e ***Sistema de Conversão Texto-Fala (SCTF)*** [4, 5, 6].

O ***SVA*** permite um melhor controle da qualidade e inteligibilidade da voz. Mas, em compensação, trabalha com um vocabulário restrito, tendo em vista que é impossível armazenar todas as palavras possíveis de serem originadas numa determinada língua. Além do mais, o custo de tal sistema se torna extremamente elevado a medida que novas palavras vão sendo incorporadas ao vocabulário existente. Logo, funciona do seguinte modo: todas as possibilidades de mensagens são gravadas e armazenadas para, posteriormente, serem reproduzidas. Para tanto, conta com as etapas de codificação, armazenamento e reconstrução da fala. Desse modo, são importantes a escolha adequada da técnica de codificação, a maneira mais econômica possível de armazenamento e o uso de formas eficientes de acesso.

Em muitas aplicações, onde se exige um vocabulário limitado, a utilização de ***SVA*** é indicada. É o caso, por exemplo, do fornecimento de saldo bancário automático por telefone. Entretanto, vale salientar que as palavras constituintes do dicionário são gravadas por um

locutor e quando da necessidade de se aumentar o vocabulário seria necessário convocar a mesma pessoa para realizar as novas gravações. Isso nem sempre pode ser possível.

O ***SCTF*** é mais flexível e proporciona uma melhor utilização do espaço de memória disponível. Ele trabalha com vocabulário não restrito sem precisar armazenar todas as possíveis palavras de uma língua. Para tanto, manipula unidades acústicas, através das quais as palavras são originadas, possibilitando a obtenção de um número ilimitado de palavras, sem a necessidade de definição de um dicionário. Entretanto, apesar das vantagens inerentes, esse sistema apresenta desvantagens, já que exige um *processamento linguístico* de alto nível para que, desse modo, seja obtida uma fala inteligível e natural.

1.2 Objetivo do Trabalho

O objetivo deste trabalho é originar uma estrutura destinada a realizar o processamento linguístico para um ***Conversor Texto-Fala para a Língua Portuguesa***, pois as referências bibliográficas seguidas evidenciam que estudos desenvolvidos não apresentam um tratamento totalmente voltado para tal idioma. Verifica-se que o processamento do texto em português obedece um conjunto de regras, que, em sua maioria, são adaptações daquelas desenvolvidas para os sistemas de conversão para a língua inglesa e alemã - fato este que ocasiona uma fala não natural, distinguindo-se bastante de certos atributos típicos de cada língua, tais como entoação, tonicidade, etc. Assim, a definição de um tratamento adequado é fundamental para a obtenção de um ***SCTF*** de alta qualidade. A originação de uma fala natural e inteligível não implica apenas em se ter um sintetizador considerável. Antes de tudo, é essencial fornecer, ao processamento do sinal, informações representativas do texto a ser convertido dentro de um formato adequado, ou seja, informações capazes de simular a pronúncia correta dos vocábulos e, consequentemente, do texto. Logo, deve-se pensar nas etapas necessárias para se determinar o modo como os vocábulos são formados, bem como na forma de sua evolução quando são lidos.

Nesse sentido, a produção de uma fala com as características desejadas está, de início, diretamente ligada à análises básicas de linguística. É imprescindível definir uma estrutura para

o sistema na qual seja considerada a evolução ou disposição das palavras. No idioma português, um vocábulo exerce influência na pronúncia do vocábulo seguinte. Por outro lado, apresenta particularidades na sua própria constituição: em um vocábulo as sílabas apresentam diferenças quanto as suas proeminências. Além do mais, uma sílaba, dependendo da palavra, pode assumir várias pronúncias. Todas essas ressalvas permitem afirmar que o ***SCTF*** deve apresentar um tratamento mais representativo. As considerações estabelecidas para o *processamento linguístico* devem dar subsídios à *prosódia.* É primordial se ter um direcionamento para o português, destacando alguns pontos essenciais, tais como: identificação dos limites de sentença, merecendo maior atenção às pausas mentais, determinação da pronúncia correta para vogais - no caso de *palavras homógrafas heterofônicas*[1] - e as várias entoações assumidas por uma unidade acústica.

Em paralelo a estruturação do *processamento linguístico* será desenvolvido um *ambiente de trabalho* de forma a possibilitar a concepção do ***SCTF.***

1.3 Aplicações de Sistemas de Síntese de Voz

A seguir são apresentadas algumas aplicações importantes, presentes no dia a dia das pessoas [7, 8], que justificam a utilização do ***SRV.***

- ***Sistema de Reprodução de Escritas:*** Nesse sistema, informações na forma escrita são armazenadas em um computador, podendo serem recuperadas quando solicitadas e reproduzidas auditivamente. Pode-se destacar, por exemplo, os dicionários falados, as enciclopédias, etc.

- ***Sistema de Instruções:*** Instruções são armazenadas e podem ser usadas em situações em que os olhos e as mãos precisam estar livres para desempenharem outra tarefa. Essas instruções, em geral, referem-se à maneira de usar um

[1] Essas palavras também são denominadas de *homônimas homógrafas.*

determinado equipamento ou mesmo como deve ser feita a sua manutenção. Como exemplo tem-se a função "help" de algum software de computador.

- ***Sistema de Fornecimento de Saldo Bancário por Telefone:*** Esse sistema encontra-se muito presente na vida das pessoas, já que é uma maneira prática e rápida de se obter informações, pois evita os desagradáveis problemas de se deslocar e perder tempo em fila de banco.

- ***Sistema de Informações em Linha:*** Pode-se dizer que dispor de informações em linha é o equivalente a se ter um jornal na sua forma falada. Usualmente, tais informações são obtidas através do telefone. Por exemplo, notícias sobre a chegada/partida de vôos, previsão meteorológica, programação de teatro e cinema, horóscopo, etc.

- ***Auxílio a Deficientes Visuais:*** O auxílio a deficientes visuais é uma das mais importantes aplicações do ***SRV***, a exemplo da máquina de leitura para cegos, a qual consiste na reprodução, de maneira falada, de um texto qualquer, impresso em braile. Outro exemplo é a monitoração das atividades desenvolvidas por cegos em um escritório através de um ***Conversor Texto-Fala*** acoplado a uma máquina de escrever ou a um computador, permitindo ao indivíduo verificar o que está sendo digitado.

- ***Auxílio de Fala a Deficientes Vocais:*** O ***SRV*** também pode ser utilizado por pessoas que apresentam deficiência vocal. A idéia geral desse tipo de aplicação é fornecer algumas palavras ao sistema para que, a partir delas, seja gerado um conjunto de frases gramaticalmente corretas. Para tanto, esse sistema exige todo um *tratamento prosódico* - procedimento essencial para a obtenção de uma fala natural e inteligível e que será enfocado nos próximos capítulos, pois a entoação das frases é fator principal em um processo de comunicação.

Assim, de uma forma rápida e geral, foram abordadas algumas aplicações dos ***SRV***. Em certas situações onde se trabalha com um vocabulário restrito - saldo bancário por telefone, por exemplo - é mais viável a utilização do ***SVA***. Por outro lado, em outros casos,

onde a mensagem não é previamente estabelecida ou se utiliza um vocabulário muito grande de palavras - Sistemas de Leitura para Cegos ou de Reprodução de Informação, respectivamente -, o ***SCTF*** é mais adequado. No caso do auxílio a deficientes visuais, ainda convém citar que o texto pode ser introduzido no sistema através de um scanner. Para tanto, é necessário a definição de um algoritmo de reconhecimento de caracteres para realizar a sua transformação.

1.4 Considerações Iniciais

Como o objetivo principal desse estudo é trabalhar com textos de vocabulário não restrito, a abordagem enfocará o ***SCTF***, que, em linhas gerais, pode ser ilustrado através da *figura 1.1*. Posteriormente será apresentada a estrutura a partir da qual será desenvolvido todo o trabalho.

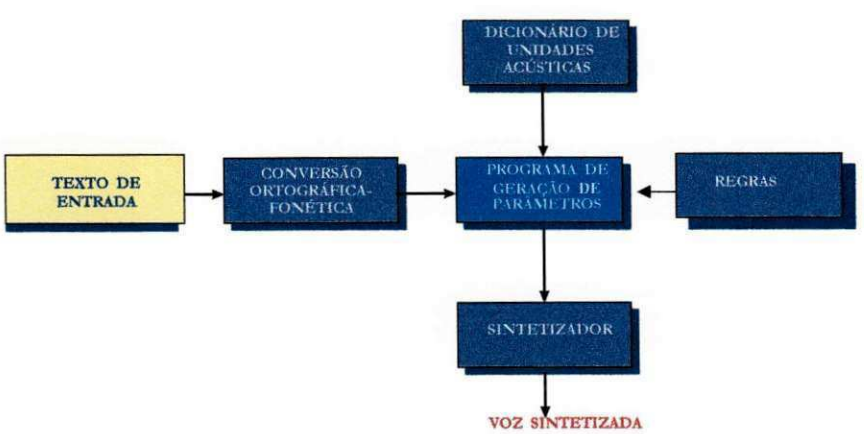

Figura 1.1: Esquema Básico de um Sistema de Conversão Texto-Fala (SCTF)

Na entrada do Sistema tem-se um texto na sua forma ortográfica, o qual pode conter símbolos, abreviações, algarismos, siglas e sinais de pontuação. Assim, esse texto recebe um tratamento - que será abordado em detalhes nas seções subsequentes - para que a sua

composição *fonética* seja obtida. A partir dessa representação *fonética*, de um conjunto de regras *fonéticas* e *fonológicas* e de um dicionário de unidades acústicas é gerado um conjunto de parâmetros que controla um sintetizador de voz. Este, por sua vez, tenta modelar o processo de produção de voz do ponto de vista acústico. A produção da fala, portanto, é possibilitada através do conjunto de parâmetros de entrada e da variação destes no tempo.

O conjunto de regras citado na estrutura - *figura 1.1* - objetiva simular o aspecto dinâmico da fala, isto é, ele é responsável por modelar as variações dos parâmetros da fala ao longo do tempo para que, desse modo, possa ser obtida a pronúncia correta das palavras, juntamente com a entoação adequada das frases. As unidades acústicas são armazenadas diretamente como fones, difones, etc. ou seus parâmetros representativos, a partir de sons isolados da fala, no dicionário.

1.5 O Processo de Leitura Realizado pelo Homem

Fisiologicamente falando, o processo de leitura realizado pelo homem pode ser representado pelo esquema exibido na *figura 1.2*. Assim, numa primeira observação, já se pode evidenciar uma razão que justifica a complexidade de estudo de um sistema originador da fala: a fala natural é um fenômeno de compensação constante, que adapta os músculos articulatórios em termos das atividades dos neurônios motores, a partir de estímulos percebidos pelo ouvido e enviados ao córtex [9].

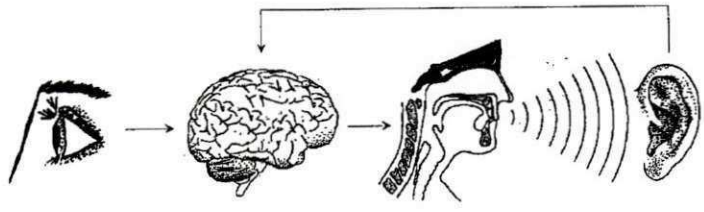

Figura 1.2: Diagrama Esquemático do Processo de Leitura Realizado pelo Homem

De uma forma mais detalhada, pode-se dizer que o processo de leitura tem início com a captura da imagem através dos neurônios sensitivos oculares, sendo transmitida na forma de estímulos elétricos para o cérebro, onde é processada pelos neurônios motores, responsáveis pela ativação correta dos músculos articulatórios, dos pulmões e das cordas vocais. Isto, então, produz a fala que é permanentemente monitorada pelo cérebro e pelos órgãos de escuta.

1.6 O Processo de Leitura Realizado pela Máquina

O processo de leitura concebido pela máquina deve apresentar uma estrutura que represente, com naturalidade, aquele realizado pelo homem. A *figura 1.3* mostra um esquema para o ***SCTF***.

Figura 1.3: Diagrama Esquemático de um SCTF

O ***SCTF*** compreende dois módulos principais. O primeiro, ou seja, o *Processamento Linguístico*, deve ser capaz de produzir uma transcrição *fonética* do texto a ser convertido em fala, além de originar informações contendo as características de entonação e ritmo, fundamentais a obtenção de uma fala mais natural. O segundo consiste no *Processamento do Sinal*, responsável pela conversão da informação simbólica - concebida no *Processamento Linguístico* - na fala propriamente dita.

1.7 A Leitura Realizada pelo Homem e a Realizada pela Máquina

Fazendo uma comparação entre a leitura realizada pelo homem e a realizada pela máquina, pode-se chegar a certas considerações que justificam a complexidade do ***SCTF*** e a preocupação em se investir cada vez mais em estudos direcionados a particularidades que influenciam a naturalidade e inteligibilidade da fala.

A leitura é uma forma de comunicação bastante usada pelo homem. Assim, devido a esse uso intenso ao logo de sua vida, adquire uma competência linguística que é impossível de ser traduzida para uma máquina. Na realidade, o processo da leitura não implica apenas na conversão da palavra em sua representação *fonológica.* O texto pode ter uma diversidade de enunciados. Os aspectos produzidos depende, portanto, do contexto, do leitor e do efeito pretendido.

Mais claramente falando, o homem, quando está aprendendo a ler, soletra as palavras. Com o passar do tempo, ele vai armazenando a pronúncia dos vocábulos, o que permite a leitura de forma contínua. Ou seja, diante do texto, inconscientemente, ele realiza uma identificação das palavras já conhecidas e promove a leitura de forma desembaraçada. O processo só sofre uma descontinuidade quando uma palavra desconhecida é encontrada: nesse caso, em geral, o locutor retorna a fase de aprendizado soletrando o vocábulo.

Logo, promover esta identificação automática entre palavra/pronúncia para uma máquina não é trivial. Seria preciso então armazenar todos os possíveis vocábulos de um idioma - o que corresponderia ao ***SVA*** . Entretanto, este procedimento não é uma alternativa viável para a conversão de textos de vocabulário ilimitado. Por outro lado, mesmo que essa alternativa fosse usada, se teria o problema de determinar, por exemplo, a entoação da palavra. Então, tudo leva a concluir que um ***SCTF***, por mais estruturado que seja, jamais originará uma leitura igual à produzida pelo homem. Na realidade, é muito difícil duas pessoas diferentes realizarem a leitura de um texto com a mesma entoação, pausas, ...

Por tudo isso, o desejado é a concepção de um sistema capaz de originar uma leitura dentro dos padrões aceitáveis, com uma qualidade de fala considerável. Esses fatores, entretanto, estão diretamente relacionados a sua complexidade, permitindo afirmar que o

desenvolvimento de arquiteturas cada vez mais completas culmina na dificuldade de se determinar os pontos que devem ser melhorados e quais os procedimentos necessários para promover esta otimização.

1.8 A Implementação dos SCTF em Tempo Real

A consideração sobre tempo real, quando direcionada a *Sistemas Texto-Fala*, pode assumir enfoques bem particulares que merecem ser citados e questionados. Pode-se dizer que a resposta imediata à entrada, ou seja, a conversão instantânea texto-fala, como uma fala natural, é impossível. Analisando a própria estrutura do Sistema – *figura 1.3* -, tem-se antes da geração do sinal de fala todo um processo de formatação do texto e, além do mais, as etapas que serão necessárias evidenciarão que os tratamentos/procedimentos levam em conta uma análise da palavra e de suas vizinhanças. Tudo isso permite afirmar que o tempo real, nesse caso, pode ser considerado como uma medida subjetiva. Logo, o tempo de resposta satisfatório será dependente da tolerância imposta pelas aplicações e pelos usuários.

Todas essas considerações realizadas até o momento são fundamentais para justificar os caminhos seguidos ao longo desse estudo. Dessa forma, em capítulos posteriores, os processamentos integrantes do ***SRV*** serão abordados mais detalhadamente. Mas, desde já, pode-se afirmar que eles devem contar com formalismos e algoritmos linguísticos, que originam restrições na pronúncia do texto.

1.9 Organização da Dissertação

Esta dissertação conta com sete capítulos. A seguir, de uma forma resumida, tem-se o conteúdo de cada um.

Capítulo 2: apresenta conceitos básicos de linguística, fundamentais para o entendimento da estrutura proposta para o ***SCTF***.

Capítulo 3: aborda os métodos existentes para se realizar a síntese sonora a partir das informações fonéticas e prosódicas extraídas do texto a ser convertido.

Capítulo 4: apresenta a estrutura proposta para o ***SCTF***, denominada de ***Sistema LEIA - Sistema de Leitura Eficiente Inteligível e Automática***..

Capítulo 5: aborda, de forma detalhada, os vários módulos integrantes do ***Sistema LEIA***. Dessa forma, são descritos os procedimentos que devem ser considerados e utilizados para permitir a originação da fala dentro de padrões aceitáveis de naturalidade e inteligibilidade.

Capítulo 6: refere-se à implementação de alguns módulos do ***Sistema LEIA*** e os resultados obtidos.

Capítulo 7: apresenta as conclusões do trabalho e as perspectivas para o desenvolvimento de novos estudos na área.

CAPÍTULO 2

CONCEITOS DE LINGUÍSTICA

*A estrutura geral para os **SCTF**, apresentada no capítulo anterior, mostra que a etapa inicial do sistema refere-se ao processamento linguístico. Assim, para permitir o desenvolvimento das suas etapas integrantes, é de fundamental importância o entendimento de determinados conceitos básicos de linguística, os quais serão abordado neste capítulo.*

2.1 Fonologia e Fonética

Considerando-se que a concepção de um ***SCTF*** deve levar em conta um *Processamento Linguístico*, é, portanto, essencial abordar conceitos básicos de *fonologia* e *fonética* (*fonemas*, *fones*,

alofones, etc.), que serão exigidos para a estruturação desta etapa, bem como realizar algumas considerações a respeito das características ou níveis de análise que serão utilizados no estudo.

Em uma primeira consideração a *fonologia* é o estudo dos sons isolados (*fonética*) ou combinados na pronúncia (*prosódia*) e na escrita (ortografia) [10]. Sob outro ponto de vista, pode-se definir *fonologia* como sendo o estudo da realização dos sons significativos de uma determinada língua, enquanto que a *fonética* consiste no estudo e classificação dos sons básicos de uma língua, definindo-os quanto a sua capacidade distintiva, ou seja, trata das particularidades na produção dos *fonemas* próprios de uma língua [11].

2.1.1 Fonemas

Os *fonemas* são sons elementares e distintivos[2] que o homem produz quando, pela voz, exprime seus pensamentos e emoções [11, 12]. Não são *letras*: *fonema* é uma realidade acústica que nosso ouvido registra, enquanto que *letra* é o sinal empregado para representar na escrita o sistema sonoro de uma língua. Assim, verifica-se que não existe uma identidade perfeita entre os *fonemas* e as *letras*. Isso pode ser facilmente exemplificado através das vogais, pois, de início, oito pronúncias diferentes podem ser associadas às vogais[3], mas só existem cinco símbolos gráficos. Por outro lado, existem *letras* que se escrevem mas que não são pronunciadas e, assim, não apresentam um *fonema* representativo.

Então, para o desenvolvimento de um ***SCTF*** pode-se destacar, inicialmente, o seguinte fato: é fundamental que exista uma correta *transcrição ortográfico-fonética* para que a fala a ser originada seja correta.

---

[2] O *fonema* é dito ser distintivo porque, ao ser substituído por outro, estabelece distinção de significado entre os vocábulos de uma língua.

[3] Os sons assumidos pelas vogais são: á, ã, é, ê, i, ó, ô *e* u.

2.2 Alofones e Fones

Há uma série de definições ligadas aos *fonemas* que precisa ser considerada quando da construção de um ***SCTF***. Inicialmente tem-se os *alofones*, que constituem a realização física dos *fonemas*: são os variantes de enunciação que cada *fonema* pode apresentar, ou seja, é a maneira como realmente são pronunciados. Como exemplo, considere duas pessoas, uma cearense e outra paraibana, pronunciando as palavras *dia* e *tia*. Os *fonemas* proferidos pelos falantes assumem sons diferentes: na realidade, cada um impõe características próprias à sua pronúncia. O paraibano produz sons que soam de forma mais dura. O cearense, por sua vez, pronuncia os sons de maneira mais "chiada", falando algo parecido como "*djia*" e "*tchia*". Mas, convém ressaltar que embora existam pronúncias diferentes, o significado dos vocábulos permanece o mesmo. Na realidade, os *fonemas* são os mesmos; o que se tem são suas variações.

Por outro lado, devem ser considerados os *fones*, unidades que formarão o dicionário utilizado no processo de *síntese*. Um *fone* consiste na realização acústica de um *fonema*, ou seja, não é uma classe de som: é um sinal sonoro.

Um mesmo *fonema* pode produzir diferentes *fones*, isto é, um falante ao pronunciar um mesmo *fonema* diversas vezes produzirá sinais sonoros distintos a cada pronúncia. Entretanto, estas distinções apresentarão um grau de semelhança que será suficiente para classificá-los com realizações acústicas de um mesmo *fonema*.

Os conceitos citados anteriormente - *fonemas*, *alofones* e *fones* - são bem próximos um dos outros e, portanto, muitas vezes são confundidos. Mas, será essencial identificá-los corretamente, pois são conceitos básicos para o Sistema proposto.

2.3 Aparelho Fonador

Não existe um aparelho especial para a produção da fala. Na realidade, os *fonemas* são produzidos através dos órgãos do aparelho respiratório e da parte superior do aparelho

digestivo. A esses órgãos da fala, constitutivos do aparelho fonador, pertencem, além de músculos e nervos, os brônquios, a traquéia, a laringe (com as cordas vocais), a faringe, as fossas nasais e a boca com a língua, as bochechas, o palato duro, o palato mole com a úvula, os dentes com os alvéolos e os lábios. A *figura 2.1* mostra uma vista da seção transversal do trato vocal onde são indicados os articuladores da fala.

Os *fonemas* são produzidos graças a modificação que esses órgãos da fala impõem à corrente de ar que sai dos pulmões. Essa corrente de ar passa pela traquéia e chega à sua parte superior, chamada laringe. Na laringe, por sua vez, se acham, horizontalmente, duas membranas mucosas elásticas - cordas vocais - por cujo estreito intervalo - glote - a corrente de ar tem de passar para ganhar a faringe, e daí, ou totalmente pela boca - *fonemas* orais - ou parte pela boca e parte pelas fossas nasais - *fonemas* nasais -, chegar à atmosfera.

Além dessa classificação dos *fonemas* citada anteriormente - orais e nasais - eles ainda podem ser surdos, sonoros e explosivos. Quando a corrente de ar se dirige à glote, esta pode encontrar-se aberta, fechada ou quase fechada. Então, um *fonema* é dito surdo quando a corrente de ar passa livremente, sem provocar a vibração das cordas vocais. Um *fonema* é considerado sonoro quando a corrente de ar encontra a glote fechada ou quase fechada e ao forçar a passagem provoca a vibração das cordas vocais.

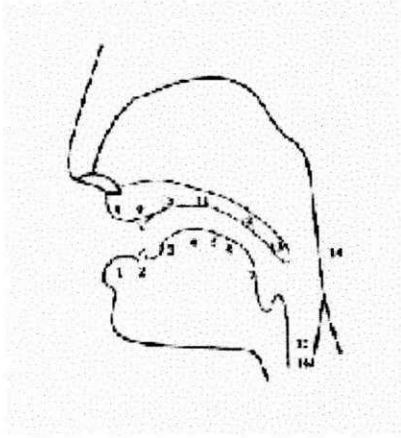

Figura 2.1: Vista da seção transversal do trato vocal[4]

[4] (1) lábio inferior, (2) incisivo inferior, (3) ponta da língua, (4) dorso, (5) frente, (6) costas, (7) raiz, (8) lábio superior, (9) incisivo superior, (10) alvéolo, (11) palato duro, (12) velum, (13) úvula, (14) faringe, (15) laringe, (16) cordas vocais e glote.

2.4 Nível de Análise

Antes de continuar com as abordagens de lingüística, é interessante considerar os níveis de análise - que estão ligados à transcrição fonética -, os quais serão usados para o desenvolvimento dos processamentos necessários a obtenção de uma fala natural e inteligível. O fato de realizar esta citação, nesse ponto, é porque consiste num enfoque que introduzirá as seções seguintes.

O nível *segmental* consiste na possibilidade de segmentar a fala em unidades mínimas - sons distintivos da língua -, formando o sistema primário da língua [1]. O nível *suprasegmental* - também denominado de *prosódia* - se relaciona com os efeitos comunicativos, ou seja, engloba processos que incluem o acento, o ritmo e a entonação.

2.5 Prosódia

Com base nas considerações feitas até o momento, pode-se dizer que a fala nada mais é do que uma seqüência de *fonemas*. Entretanto, na realidade, a fala não é um processo tão simples; não é apenas uma seqüência de sons.

Uma pessoa expressa várias informações através da fala - estado emocional, personalidade, naturalidade, etc. -, mais precisamente através da entoação que dá às palavras e às sentenças. Assim, surge a necessidade de definir para o ***SCTF*** uma etapa responsável em realizar um tratamento prosódico.

A *prosódia* é a parte da *fonética* que trata da correta acentuação e entonação dos *fonemas*. Assim, a sua preocupação principal é conhecer a sílaba predominante, denominada de sílaba tônica, e a melodia quando da produção de uma frase.

A acentuação, portanto, é importante e merece uma atenção especial. Ela representa o realce que um falante dá as palavras. Através dela o ouvinte sabe em que partes do enunciado ele deve concentrar sua atenção.

As funções prosódicas são desempenhadas através de variações de certos parâmetros, tais como *frequência fundamental*, *duração* e *energia*. Essa afirmação pode ser explicada considerando cada parâmetro separadamente. Além disso, um outro fenômeno denominado coarticulação[5], - fenômeno de mudanças na articulação e acústica de um *fonema* devido ao seu contexto fonético - desempenha papel relevante na *prosódia*.

2.5.1 Frequência Fundamental

A *frequência fundamental* (f_o) - parâmetro prosódico mais importante - corresponde à taxa de vibração das cordas vocais de uma pessoa enquanto ela está falando. Do ponto de vista acústico, corresponde à periodicidade do sinal de fala: o número de vezes que as cordas vocais se abrem e se fecham por segundo é igual ao número de repetições que a forma de onda apresenta por segundo.

2.5.2 Duração

A *duração* é definida como sendo o intervalo de tempo de um *fone*. É por isso que muitas vezes também é chamada de comprimento de um *fone*.

Com relação a este parâmetro é interessante relembrar que o sinal de fala é contínuo e, devido a isso, não é trivial determinar o início e o fim de um *fone*. Além do mais, existe o problema da coarticulação que aumenta a dificuldade de identificação das fronteiras dos *fones*.

---

[5] Melhor explicando o fenômeno da coarticulação, a produção da voz envolve uma sequência de movimentos de articuladores, que implica numa sucessão de configurações do trato vocal ao longo do tempo. Entretanto, os movimentos para sucessivos fonemas sobrepõem-se no tempo, o que torna as configurações fortemente dependentes das variações devidas a fonemas adjacentes.

2.5.3 Energia

A *energia* é outro parâmetro que merece ser citado, embora nesse estudo não seja tratado, já que, para o objetivo pretendido, a *frequência fundamental* e *duração* apresentam maior importância. Relaciona-se com a amplitude do sinal de fala, variando, portanto, segundo a pressão do fluxo de ar vindo dos pulmões.

2.5.4 Considerações sobre os Parâmetros Prosódicos

Para a produção da fala, será utilizado, no processamento do sinal, um dicionário onde estarão armazenadas as unidades acústicas (*fones*). Estas unidades, de início, apresentam valores para os parâmetros *frequência fundamental* e *duração*. Mas, é sabido que, dependendo do contexto e da sílaba, os valores devem ser alterados, pois os *fones* não soam com a mesma proeminência nem com a mesma entoação. Por exemplo, uma palavra pode ser proferida mais forte ou mais fraca; ela pode ser pronunciada variando a *duração* das vogais. É por tudo isso que surge a necessidade de utilizar algum procedimento para fazer esses acertos.

2.5.5 Acentuação e Ritmo

A *acentuação* é o modo de proferir um som ou grupo de sons com mais relevo que outros. Assim, para o português é verificado um acento de intensidade, que implica em um maior esforço respiratório. Ele se manifesta de duas maneiras: no vocábulo considerado isoladamente (acento vocabular) ou ligado à enunciação da frase (acento frásico).

Após essas considerações de *linguística* e dos *parâmetros prosódicos*, a seguir serão abordadas as maneiras de se realizar a síntese sonora propriamente dita.

CAPÍTULO 3

CONCEITOS DE SÍNTESE

*No **SCTF**, após o processamento linguístico há o processamento do sinal. Este capítulo apresenta as maneiras de se realizar o processo de síntese propriamente dito.*

3.1 Introdução

No primeiro capítulo foram classificados os dois grupos de ***SRV***. A partir dessas considerações apresentadas, pode-se perceber que eles o armazenamento das palavras/unidades que originarão a fala. Em ambos os casos, há a necessidade de contar com uma etapa preliminar de codificação, na qual o sinal analógico é amostrado e digitalizado,

podendo incorporar distintas taxas de compressão ao sinal digital em função do grau de complexidade/custo permitido. Quanto maior a taxa de compressão desejada, maior a complexidade presente no processo de codificação e, consequentemente, maior o custo.

Apesar deste trabalho ser direcionado para o processamento linguístico, as afirmações anteriormente feitas justificam a importância de se abordar, de uma forma geral, os métodos básicos de codificação digital, além das técnicas existentes utilizadas para implementação do sintetizador.

3.2 Métodos Básicos de Codificação Digital de Voz

Em se tratando de codificação de voz, três métodos básicos podem ser citados: *codificação de forma de onda*, *codificação paramétrica* e *codificação híbrida* [13]. A seguir, são apresentadas, de forma resumida, suas descrições.

3.2.1 Codificação de Forma de Onda

Os *Codificadores de Forma de Onda* apresentam-se definidos tanto no domínio do tempo quanto no domínio da frequência. Como característica principal, o processo de quantização é realizado diretamente na forma de onda do sinal. O grande objetivo consiste em reproduzir a forma de onda amostra por amostra de maneira eficiente. Assim, podem ser citadas várias classes representativas deste método:

- Codificador *PCM* (Pulse Code Modulation) [14];
- Codificador *ADPCM* (Adaptative Differential Pulse Code Modulation) [15];
- Codificador *SBC* (Sub-Band Coding) [14].

3.2.2 Codificação Paramétrica

A *Codificação Paramétrica*, como o próprio nome sugere, consiste em representar os segmentos de fala a serem sintetizados por parâmetros. Assim, pode-se evidenciar o *VOCODER-LPC* [15, 16], que representa um exemplo clássico deste modelo, ilustrado na *figura 3.1*. Pode-se ver que ele consiste basicamente de uma excitação e um filtro digital (modelo fonte-filtro).

3.2.3 Codificação Híbrida

A *Codificação Híbrida* combina os dois modelos anteriormente abordados. Na realidade une o potencial de alta qualidade dos *Codificadores de Forma de Onda* com a baixa taxa de bits dos *Codificadores Paramétricos* do tipo *VOCODER*.. Os *Codificadores Híbridos* são baseados no modelo fonte-filtro, entretanto utilizam tipos de excitação mais complexos, objetivando impor uma maior naturalidade a voz sintetizada. Esse tipo de codificação ocupa lugar de destaque por ter originado uma série de codificadores padronizados para telefonia móvel celular, a exemplo do utilizado no sistema americano *V-SELP* (Vector-Sum Excited Linear Predictive) [14].

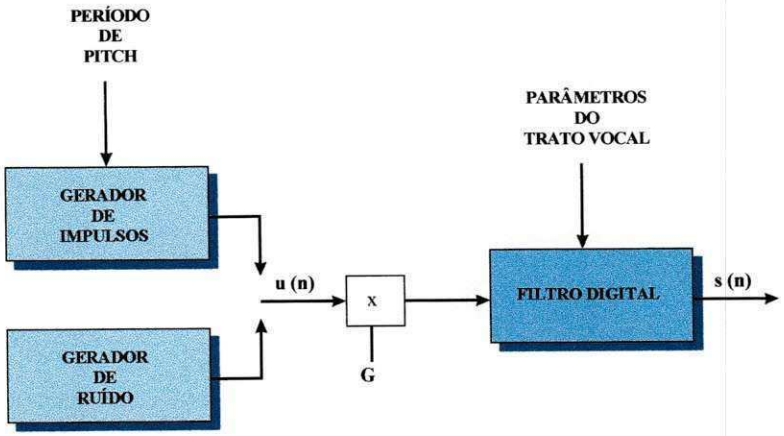

Figura 3.1: Codificação Preditiva Linear (LPC)

3.3 Síntese Sonora

Existem várias maneiras de se realizar a *Síntese Sonora.* Dentre elas, três categorias principais podem ser evidenciadas: *Síntese por Regras*, *Síntese Articulatória* e *Síntese por Concatenação de Unidades Acústicas.* A seguir será apresentada uma rápida discussão para cada uma, embora este trabalho utilize um sintetizador concatenativo, fato este que justifica o detalhamento das técnicas existentes e suas devidas comparações, possibilitando a determinação de suas vantagens e desvantagens.

3.3.1 Síntese por Regras

A *Síntese por Regras* baseia-se num modelamento paramétrico do sinal de voz e num conjunto de regras que regem a evolução temporal dos parâmetros [9, 12]. A geração da fala não é realizada através da solução de equações físicas para o aparelho vocal: ela é obtida a partir do modelamento das características acústicas principais do sinal de fala. A seguir pode-se ver a ilustração da citada categoria de síntese - *figura 3.2* -, bem como uma breve explicação das suas partes constituintes.

Seguindo o diagrama esquemático da *Síntese por Regras*, pode-se verificar que a etapa de síntese do sinal apresenta-se dependente de uma série de etapas anteriores responsáveis por determinar informações adequadas para alimentar o processamento sonoro propriamente dito. Assim, de início, é concebido um dicionário de fala, que conterá unidades representativas das transições e coarticulações. Para tanto, um locutor profissional realiza a leitura de palavras, que são gravadas e posteriormente armazenadas. Em seguida, através de uma análise da fala, é originado um dicionário paramétrico, a partir do qual serão determinadas as regras, que ficarão acondicionadas em um arquivo de regras. As referidas regras combinam-se com as informações prosódicas e fonéticas e um sinal de fala paramétrico é produzido, o qual é, posteriormente, transformado em fala digital pelo sintetizador.

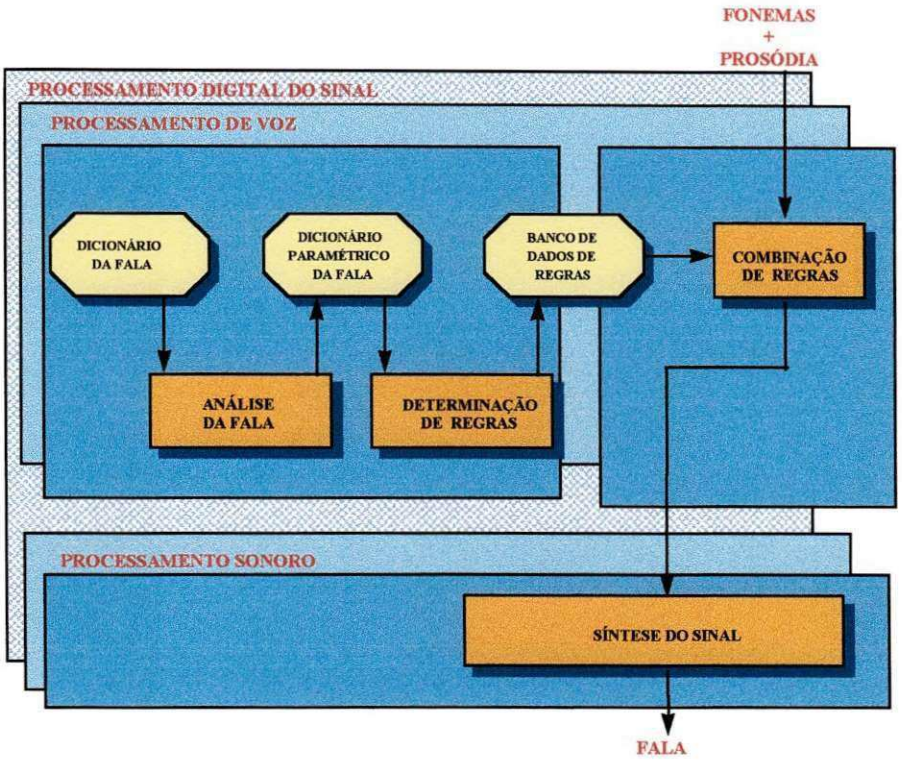

Figura 3.2: Diagrama Esquemático da Síntese por Regras (Dutoit et. al., 1993. [9])

É interessante citar que a determinação dos parâmetros está ligada a um estudo minucioso das propriedades espectrais da fala natural, a partir do qual serão determinadas aquelas propriedades que são relevantes no domínio acústico, ou seja, considerando o modelo acústico básico fonte/filtro, tem-se que o filtro - descrito por um conjunto de formantes - representa a articulação na fala, sendo responsável por modelar o espectro, o qual representa a posição e o movimento dos articuladores. A fonte representa a fonação e modela o fluxo glotal ou ruído de excitação. O controle da fonte/filtro se dá através de um conjunto de regras fonéticas.

3.3.2 Síntese Articulatória

Para facilitar a apresentação da *Síntese Articulatória* [1, 9, 17] considere o processo de produção da fala. Resumidamente, este processo pode ser explicado da seguinte forma:

inicialmente um indivíduo manifesta o desejo de produzir um enunciado; esse desejo se materializa enviando comandos aos músculos formadores do trato vocal; tais músculos, por sua vez, fazem com que os articuladores se movam e atinjam uma posição que possibilitará a formação do pretendido som. Assim, esta categoria de síntese tenta modelar as configurações que o trato vocal humano assume à medida que a fala vai sendo produzida.

As considerações realizadas acima permitem afirmar que a utilização da *Síntese Articulatória* exige um estudo apurado a cerca dos movimentos do trato vocal humano enquanto uma pessoa está falando, com o objetivo de conhecer a evolução dos articuladores ao longo do tempo. Essa afirmativa só vem reforçar um fato já esperado: a implementação deste modelo se constitui em tarefa consideravelmente complexa, tendo em vista o grau de dificuldade, senão a impossibilidade, de simular todos os possíveis movimentos dos articuladores.

Os *Sintetizadores Articulatórios*, portanto, são modelos físicos baseados na descrição detalhada da fisiologia da produção da fala e na física da geração do som no aparelho vocal. Assim, pode-se concluir que este tipo não é viável: um bom sintetizador deve ser capaz de simular a posição dos articuladores , o que não é trivial.

3.3.3 Síntese por Concatenação de Unidades Acústicas

Esta categoria realiza o processo de S*íntese* através da C*oncatenação de Unidades Acústicas* previamente gravadas e armazenadas. Então, para se utilizar esta técnica é exigida a concepção de um dicionário com todas as unidades acústicas necessárias para se gerar o sinal de fala desejado.

O diagrama esquemático da *Síntese por Concatenação de Unidades* [9] é mostrado na *figura 3.3.* Inicialmente, são gravadas palavras a partir das quais serão identificadas e extraídas as unidades. Uma inconveniência principal, e que merece ser abordada, logo surge: as transições entre as unidades, já que as descontinuidades espectrais nas junções produzem efeitos não desejados. Como contornar tal problema?

Uma possível solução, e que será aplicada neste estudo, se baseia na seguinte consideração: quanto maior for a unidade em número de *fones*, menor será o número de junções, o que implica em uma melhor qualidade da fala produzida. Entretanto, à medida que as unidades vão se tornando maiores, se faz necessário também aumentar o número de unidades acústicas do dicionário, necessárias para gerar todas as possíveis combinações de palavras que um idioma pode originar. Assim sendo, o ***SCTF*** tende a se aproximar do ***SVA***, perdendo suas vantagens peculiares. Então, como determinar as unidades a serem isoladas e armazenadas?

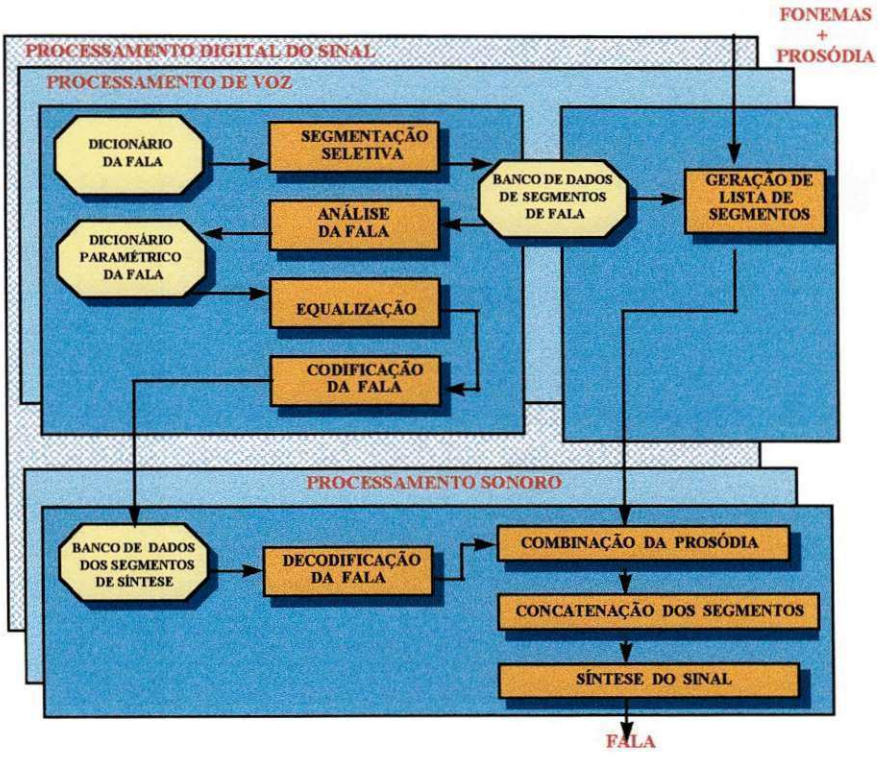

Figura 3.3: Diagrama Esquemático da Síntese por Concatenação de Unidades (Dutoit et. al., 1993. [9])

A determinação das unidades acústicas que irão compor o dicionário exige uma avaliação individual e, também, uma que relacione cada unidade e suas possíveis junções. Portanto, em geral, são utilizadas as seguintes classes de unidades: *sílabas*, *semi-sílabas* e *fones* (*difones, trifones e polifones*).

Para o sistema em desenvolvimento será considerado, principalmente, o *difone*, que é definido como sendo o segmento de voz que tem início no centro da região espectralmente estável de um *fone* e termina no centro da região estável do próximo *fone*, contendo a transição completa entre os dois *fones*. Entretanto, também serão utilizadas unidades maiores quando a inconveniência da transição ainda se faz presente, ou seja, *trifones* e *polifones*.

Assim, através de uma segmentação seletiva é originado um banco de segmentos, o qual passará por uma análise [9]. O resultado forma um outro arquivo contendo informações paramétricas dos segmentos. Mas, como os segmentos são obtidos a partir de diferentes palavras, que se encontram em diferentes contextos, surgem desacordos relacionados à amplitude e ao timbre. Um equalizador, responsável por impor amplitude espectral similar nos extremos dos segmentos - sendo a diferença distribuída nas suas vizinhanças - contorna parte do problema. Os conflitos de timbre são melhor resolvidos suavizando individualmente acoplamentos de segmentos, quando se fizer necessário. Após estes tratamentos, tem-se a construção de um arquivo com os segmentos de síntese.

Para a produção da fala, um conjunto de etapas ainda se faz necessário. Primeiramente, acontece uma combinação *prosódica*, que se baseia em duas informações: a primeira, refere-se a uma lista de segmentos, obtida a partir do banco de segmentos da fala e das considerações referentes aos *fonemas* e a *prosódia*; a segunda provêm do banco de segmentos de síntese. Em seguida, é verificada a concatenação dos segmentos e, por fim, tem-se a síntese do sinal.

3.4 Técnicas para Implementação do Sintetizador

Considerando que a fala será originada a partir de unidades previamente gravadas e armazenadas, várias técnicas podem ser seguidas para a implementação do sintetizador. Assim, as mais utilizadas [18] são: *LPC* (Linear Predictive Coding), o *Codificador Híbrido*, o *TD-PSOLA* (Time-Domain Pitch-Synchronous OverLap-Add) e o *MBR-PSOLA* (Mullti-Band Re-Synthesis Pitch-Synchronous OverLap-Add).

3.4.1 Técnica LPC

Na realidade, esta técnica já foi citada na seção 3.2.2. Entretanto, algumas colocações pertinentes ainda podem ser feitas. O *LPC* constitui um modelo paramétrico de codificação bastante conhecido e utilizado, tendo em vista que permite uma representação precisa e compacta de parâmetros espectrais de voz. Além do mais, em se tratando de *síntese de voz*,, possibilita modificações nas unidades, que serão concatenadas para originar os vocábulos, pela simples alteração de parâmetros, facilitando o tratamento prosódico.

No *VOCODER-LPC*, ilustrado na *figura 3.1*, o sinal de fala é segmentado em quadros. A cada quadro, um detetor de pitch classifica o sinal em sonoro e não sonoro. Quando o sinal é dito sonoro o período de pitch é calculado. A excitação sonora é representada por um trem de impulsos com período igual ao período de pitch. A excitação não sonora, por sua vez, é representada por um ruído branco de distribuição uniforme.

O filtro *LPC* é calculado através de uma análise *LPC* [16]. Os coeficientes *LPC*, o ganho, a classificação sonoro/não sonoro e o período de pitch, são parâmetros associados à produção do sinal no quadro de análise. Eles são quantizados antes da sua transmissão/armazenagem.

Em primeira análise, pode-se afirmar que este modelo seria o mais indicado. Entretanto, conta com uma desvantagem que justifica o estudo de outros modelos de codificação. Por ser paramétrico, a forma de onda não é conservada, o que prejudica a qualidade de voz.

3.4.2 Técnica Híbrida

A *técnica Híbrida* também foi abordada na seção 3.2. Entretanto, é interessante tecer mais alguns comentários. Neste modelo, o sinal de fala também é segmentado em quadros. O filtro pode ser o mesmo utilizado no *modelo LPC*. Entretanto, trabalha com tipos de excitação mais complexos. Em geral, um processo análise-por-síntese é utilizado para selecionar dentre

um conjunto de possíveis excitações - denominado de *codebook* - aquela a ser aplicada num dado quadro.

Assim, pode-se dizer que o nível de processamento desta técnica apresenta-se bem mais complexo do que o verificado nos outros tipos de codificadores. Entretanto, origina, para taxas de 4 a 16 Kbits/s [13], uma qualidade superior, quando comparada com os resultados produzidos a partir dos codificadores de forma de onda com taxas mais elevadas.

Apesar dessa técnica apresentar fundamentações mais complexas do que a *LPC*, ela vem sendo estudada e utilizada, devido aos bons resultados alcançados, relacionados à qualidade da produção da fala [9].

3.4.3 Técnica TD-PSOLA

A *técnica TD-PSOLA* usa uma representação não paramétrica do sinal de fala, objetivando, dessa forma, conservar melhor a forma de onda das unidades acústicas.

Assim, esta técnica permite o processamento direto da forma de onda do sinal de fala, possibilitando a realização das alterações de sua *frequência fundamental* e da *duração*. Isso é interessante para o ***SCTF***, já que a proeminência e a entonação das palavras e, por conseguinte, das sentenças serão obtidas através das modificações dos referidos parâmetros.

Em linhas gerais, este modelo consiste na superposição e soma de blocos de sinal (unidades acústicas), deslocados no tempo, de maneira síncrona com o período de pitch. Assim, a grande vantagem que pode ser evidenciada é que, como o processamento é realizado diretamente sobre a forma de onda, uma melhor preservação da qualidade do timbre[6] de voz pode ser obtida. De uma forma mais detalhada, o *TD-PSOLA* conta com três etapas que serão discutidas a seguir.

---

[6] O timbre é a característica de um sinal sonoro que permite distinguir entre um locutor e outro.

3.4.3.1 Etapas da técnica PSOLA

Etapa 1: ***Geração dos Sinais Elementares de Análise***

O sinal original é transformado em uma sequência temporal de sinais janelados de curta duração *(Equação 3.1) - sinais elementares de análise* -, que constituem os blocos básicos utilizados pelo processo de síntese.

$$x_m(n) = h_m(n - t_m) \cdot x(n) \qquad (3.1)$$

Onde: $x_m(n)$ = sinais elementares de análise

t_m = marcas de análise

$h_m(n)$ = janela de Hamming

Na prática, a transformação citada anteriormente é obtida submetendo-se o sinal original a uma sequência de janelamentos de Hamming, onde a frequência de análise é síncrona com o período de pitch do sinal. Entretanto, o janelamento deve ser realizado de modo que haja sobreposição de 50%. A *figura 3.4* ilustra esse processo.

As marcas de análise t_m se sucedem em uma cadência síncrona da *frequência fundamental* das porções sonoras do sinal, sendo obtidas de maneira automática através de um algoritmo de marcação de pitch

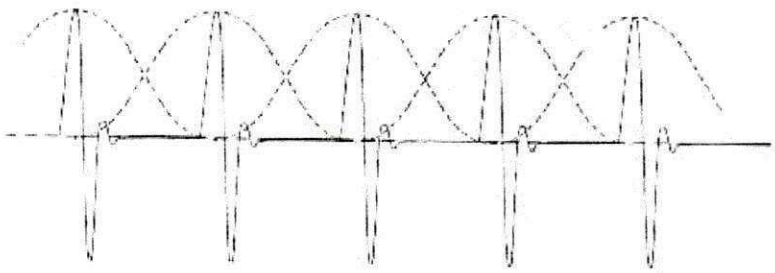

Figura 3.4: Janelamento do Sinal de Análise

Etapa 2: *Geração dos Sinais Elementares de Síntese*

Os *sinais elementares de síntese (Equação 3.2)* são gerados a partir dos *sinais elementares de análise.* Para tanto, marcas de síntese são definidas: cada *sinal elementar de síntese* provém de um *sinal elementar de análise* único, através de uma função de deformação temporal que relaciona as marcas de análise às marcas de síntese. Esta relação é responsável pela modificação da *duração* ou da *frequência fundamental* das unidades.

$$x_q(n) = x_m(n + t_m - t_q) \qquad (3.2)$$

Onde: $x_q(n)$ = sinais elementares de síntese

$x_m(n)$ = sinais elementares de análise

t_m = marcas de análise

t_q = marcas de síntese

Etapa 3: *Geração do Sinal de Síntese*

O *sinal de síntese (Equação 3.3)* é obtido por simples superposição e adição dos *sinais elementares de síntese.*

$$x(n) = \sum x_q(n) \qquad (3.3)$$

Onde: $x(n)$ = sinal de síntese

$x_q(n)$ = sinais elementares de síntese

3.4.3.2 Modificação dos Parâmetros

A *técnica TD-PSOLA* permite a alteração dos parâmetros *frequência fundamental* e *duração.* A seguir serão abordadas as considerações necessárias para a realização dessas mudanças, essenciais para se ter uma fala com características semelhantes à do homem.

- ***Modificação da Duração:*** A modificação da *duração* das unidades acústicas é obtida através da relação entre as marcas de análise e de síntese, que acarretará em eliminação ou adição de períodos do sinal de voz, ou seja, para produzir uma aceleração da fala, faz-se necessário eliminar períodos, enquanto que, para obter-se uma desaceleração, períodos devem ser acrescidos.

Para facilitar o entendimento desse processo, é interessante citar um exemplo. Suponha que a *duração* de uma certa unidade acústica deve ser duplicada. A relação entre as marcas deve, então, ser tal que associe às marcas de síntese apenas metade das marcas de análise. Caso a situação fosse inversa, ou seja, se fosse desejado uma redução na *duração* da unidade acústica, a correspondência entre as marcas deve duplicar os sinais elementares de síntese. As *figuras 3.5 e 3.6* mostram as situações abordadas.

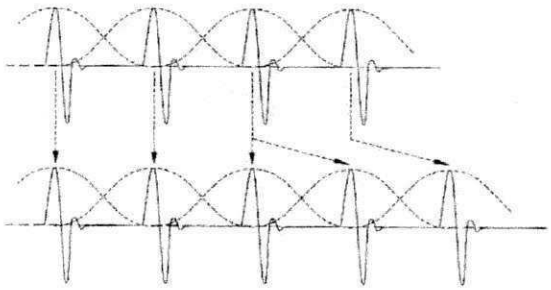

Figura 3.5: Técnica TD-PSOLA - Aumento de Duração

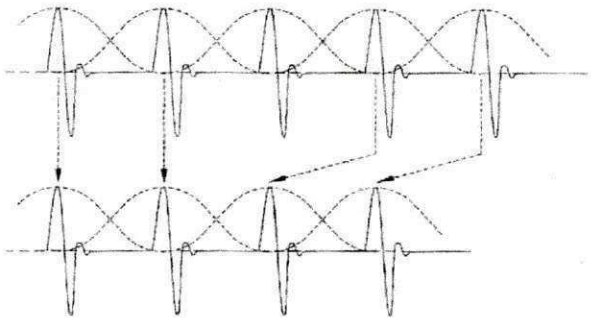

Figura 3.6: Técnica TD-PSOLA - Diminuição de Duração

- ***Modificação da Frequência Fundamental.*** A modificação da *frequência fundamental* das unidades acústicas segue o pensamento desenvolvido para a *duração*, ou seja, é obtida através de relações entre as marcas de análise e de síntese. Portanto, para alterar a *frequência fundamental* por um fator β se faz necessário multiplicar o atraso entre os blocos elementares sucessivos pelo inverso de β. Assim, se for desejado duplicar a *frequência fundamental* de uma unidade acústica, faz-se necessário dividir pela metade o atraso entre os blocos elementares. Caso contrário, se for desejado reduzir a *frequência fundamental* pela metade, deve-se duplicar o atraso entre os sinais elementares. As *figuras 3.7 e 3.8* ilustram o processo em questão.

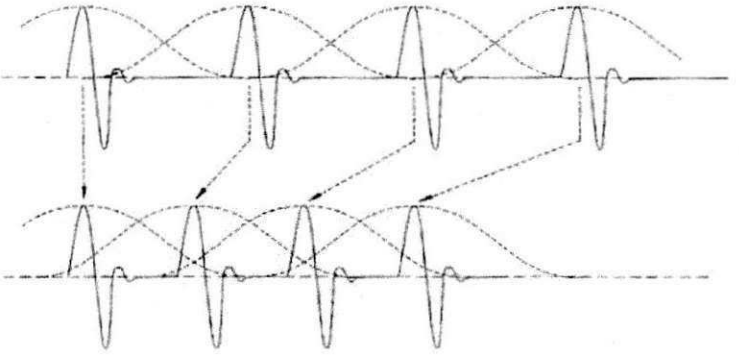

Figura 3.7: Técnica TD-PSOLA - Aumento da Frequência Fundamental

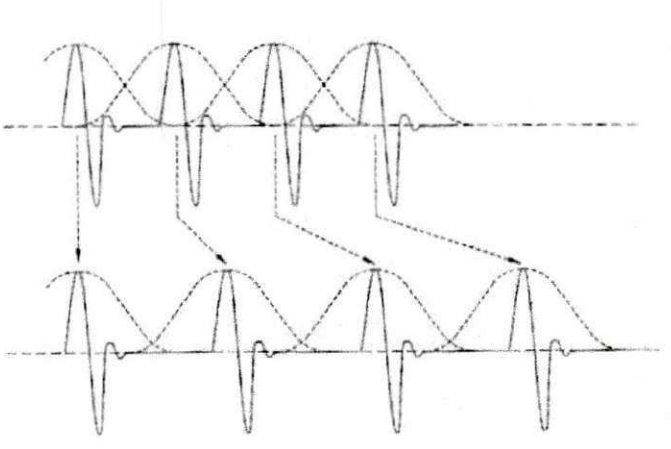

Figura 3.8: Técnica TD-PSOLA - Aumento da Frequência Fundamental

3.4.3.3 Limitação da Técnica TD-PSOLA

Apesar de preservar a forma de onda e ser amplamente utilizada em ***Sistemas Texto-Fala***, a *técnica TD-PSOLA* apresenta algumas inconveniências [13], ou seja:

- As modificações de *duração* só podem ser implementadas de maneira quantizada (..., 1/2, 2/3, 3/4, ..., 4/3, 3/2, 2/1, ...);

- As modificações da *frequência fundamental* introduzem uma alteração na *duração*, causada pela maior ou menor superposição dos segmentos janelados. Logo, esta variação deve ser compensada apropriadamente;

- Durante o aumento de *duração* efetuado em porções não sonora do sinal de fala , a repetição de segmentos introduz uma periodicidade, a qual é responsável por uma aparência "metálica" da fala sintetizada;

- Variações elevadas da *frequência fundamental* causam distorções sensíveis, já que a envoltória resultante da superposição das janelas se afasta consideravelmente de um nível constante.

3.4.3.4 Técnica MBR-PSOLA

A *MBR-PSOLA*, também chamada de *MBROLA*, é a técnica de implementação mais nova dentre as citadas anteriormente. Ela foi desenvolvida objetivando solucionar as inconveniências e limitações originadas a partir do uso da *técnica TD-PSOLA*.

A estrutura geral de um sintetizador *MBR-PSOLA* é apresentada na *figura 3.9*. Como se observa nesta figura, além de uma nova base de dados obtida através da re-síntese harmônica [19], um bloco de interpolação temporal linear é adicionado ao *TD-PSOLA*. Esta

interpolação é realizada, na prática, no início e no final da forma de onda de cada segmento sonoro. Neste caso, o algoritmo de síntese resultante auxilia a interpolação espectral entre os segmentos sonoros, sem o aumento considerável da complexidade do sistema.

Um fator importante no *MBR-PSOLA* é que as marcas de pitch são automáticas. Deste modo, dado o período de pitch constante e fases harmônicas iniciais impostas durante a re-síntese, marcadores de pitch podem ser dados em qualquer posição relativa dentro dos períodos re-sintetizados (partindo do instante de reinicialização), mantendo-o constante nos segmentos sonoros da base de dados. Como o instante de reinicialização é totalmente controlado, marcadores de pitch são impostos implicitamente.

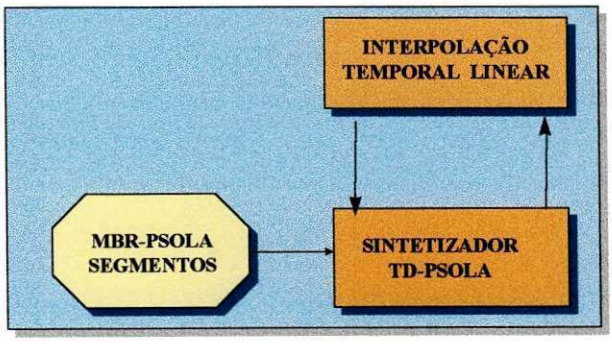

Figura 3.9: Diagrama de Blocos do Sintetizador MBR-PSOLA (Dutoit et. Al, 1993. [9])

CAPÍTULO 4

O SCTF PROPOSTO PARA *L*EITURA *E*FICIENTE, *I*NTELIGÍVEL E *A*UTOMÁTICA : LEIA

*Neste capítulo é apresentada a estrutura geral para o **SCTF**. Assim, em seguida, os vários módulos integrantes são descritos de forma a possibilitar, posteriormente, o desenvolvimento/implementação do sistema.*

4.1 Sistema de Conversão Texto-Fala

Um ***SCTF*** transforma um texto em sua forma ortográfica em fala na forma de onda sonora. Um texto ortográfico é formado por uma combinação das *letras* que constituem o

alfabeto empregado. A fala, por sua vez, é originada a partir de combinações dos possíveis tipos de sons - *fonemas* - que o trato vocal humano é capaz de produzir. Em princípio, a produção da fala por síntese a partir de um texto é um processo simples, formado por dois estágios: o primeiro, responsável pela determinação da sequência de *fonemas* referentes ao texto que se deseja sintetizar, e o segundo, referente à produção dos respectivos *alofones* - realizações físicas dos *fonemas*. Entretanto, esse procedimento não é suficiente para se produzir uma fala natural e inteligível. Para que tais características sejam alcançadas, o texto deve ser submetido a certos tratamentos.

A implementação do processo de conversão texto-fala no ***SCTF*** é realizada em duas fases:

- conversão do texto em alguma forma de representação linguística, a qual engloba a informação dos *fonemas* a serem produzidos, suas durações, as localizações dos limites de frase e o contorno de "*pitch*" a ser usado;
- conversão da representação linguística em uma forma de onda da fala.

O ***SCTF*** existente [17] apresenta quatro etapas. Inicialmente o texto ortográfico é submetido a um pré-processamento, no sentido de adequar siglas, abreviações e números. Em seguida é realizada uma transcrição ortográfico fonética, objetivando associar a *letra* ao *fonema*. Assim, tem-se o *tratamento prosódico*, de forma a permitir uma entoação correta dos vocábulos. Para concluir, a fala sintetizada é originada. A *figura 4.1* ilustra o citado sistema. Entretanto, é conveniente ressaltar que não se tem nenhum conhecimento a cerca de como realmente caracterizar as palavras: que *fonemas* podem ser associados às *letras*? Que procedimentos devem ser definidos para se ter um *tratamento prosódico* capaz de dotar a fala de naturalidade e inteligibilidade?

4.2 Sistema LEIA: Estrutura Geral

Para o ***SCTF*** é fundamental, inicialmente, estabelecer tratamentos direcionados a caracterização do idioma português, objetivando obter a formatação do texto num padrão

compreendido pelo sintetizador. Este último, por sua vez, deve realizar ajustes em certos parâmetros característicos da fala.

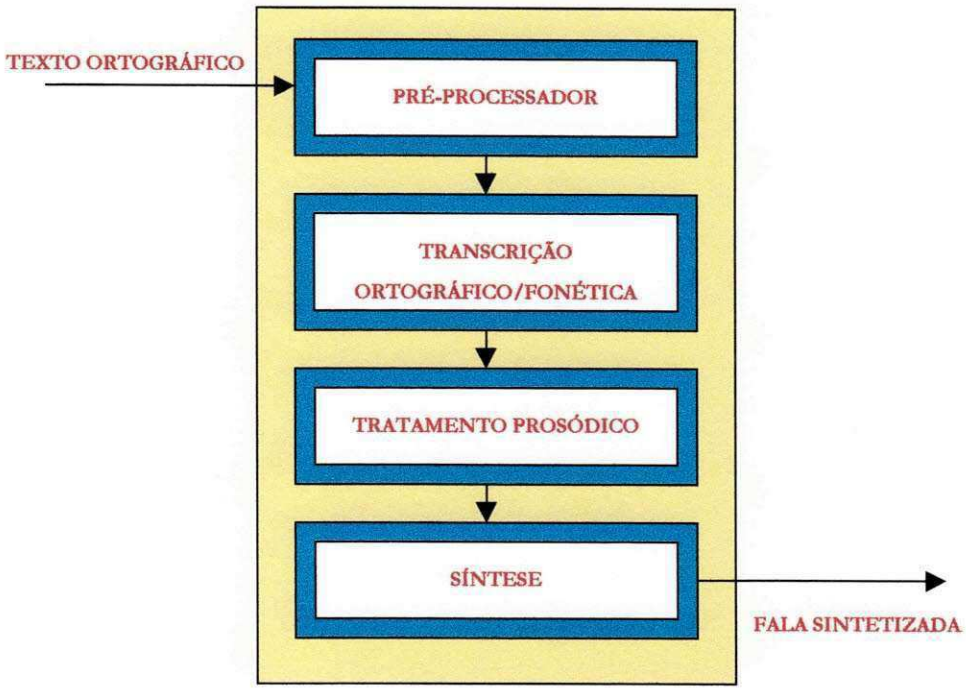

Figura 4.1: Diagrama Básico de um *SCTF*

Considerando as discussões realizadas até o momento [1, 4, 5, 17, 20, 21, 22], propõe-se a utilização da estrutura geral (estrutura hierárquica) [23], descrita na *figura 4.2*, para o ***Sistema de Conversão Texto-Fala para a Língua Portuguesa***, o qual foi denominado de ***LEIA***.

4.2.1 Considerações Gerais

A primeira etapa do ***Sistema LEIA*** consiste na ***Normalização do Texto***, de forma que as muitas abreviações, números e siglas encontrados em um texto possam ser melhor trabalhadas. Por outro lado, alguns caracteres de pontuação podem apresentar mais de um significado. O ponto (.), por exemplo, ora representa uma abreviação ora marca o final de

uma sentença. O hífen (-), por sua vez, pode simbolizar a separação de uma palavra ou a delimitação do texto ou ainda uma enumeração.

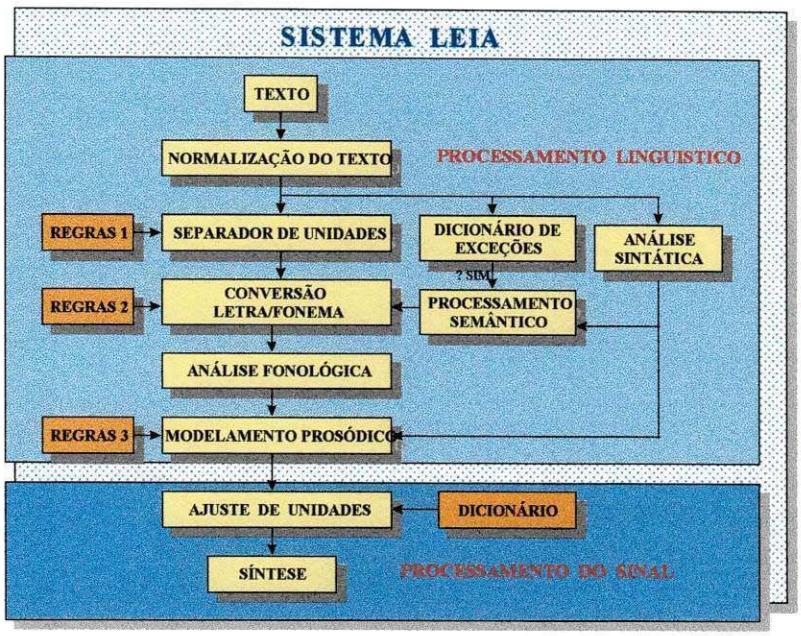

Figura 4.2: Sistema LEIA: Estrutura Geral

A etapa de ***Normalização do Texto***, portanto, consiste em identificar o significado correto dos caracteres de pontuação e, posteriormente, expandir as *abreviações*, as *siglas* e os *números*. Assim, quando da ocorrência da seqüência ***sr.***, o ponto (.) deve ser interpretado como um fator determinante de uma *abreviação* e automaticamente deve-se processar uma tradução que originará a palavra ***senhor***. Por sua vez, ***PB*** deve ser identificada como uma *sigla* e, assim, a forma correta de pronúncia é determinada - no caso "***pebê***". Já o número ***321-0909*** deve ser traduzido para "***três dois um zero nove zero nove***", por apresentar o formato de um número de telefone. Mas, convém acrescentar que certas *siglas* são comumente proferidas abreviadamente. Surge, então, a questão de como proceder para identificar a forma como as *siglas* realmente deverão ser proferidas.

Desta maneira, o texto é reescrito, ficando pronto para ser submetido a um conjunto de etapas: ***Separador de Unidades***, ***Dicionário de Exceções*** e ***Análise Sintática***.

O ***Separador de Unidades*** é essencial para o ***Sistema LEIA***. Esta afirmativa pode ser explicada considerando o método de síntese escolhido, ou seja, *Síntese por Concatenação de Unidades Acústicas*. Portanto, um bom caminho para identificar as unidades que formarão as palavras consiste em determinar as sílabas que as compõem, mesmo sabendo que, em certas situações, se faz necessário dividir uma sílaba ao meio devido ao problema da coarticulação. Esses casos devem ser identificados para tornar a fala mais natural.

O ***Separador de Unidades***, na realidade, não realiza apenas a separação das sílabas, já que em alguns casos duas palavras distintas são pronunciadas como sendo apenas uma, ou seja, não há uma pausa entre elas. Um exemplo bastante evidente consiste nos vocábulos "*seis anos*": o leitor ao pronunciá-los não introduz nenhuma pausa e, além do mais, o /*s*/ final de "*seis*" assume um som de /*z*/ quando se une ao /*a*/ de "*anos*". Assim, o que se pode concluir é que a não consideração deste fato leva a uma pronúncia pouco natural e a uma indesejável semelhança com a voz artificial. Portanto, para identificar esses casos, um conjunto de regras, que será detalhado mais a seguir, é formalizado.

O ***Dicionário de Exceções*** é responsável pela identificação das palavras homônimas heterofônicas[7] - palavras com mesma grafia mas que apresentam pronúncias diferentes -, as quais devem, posteriormente, ser submetidas a um ***Processamento Semântico***, que é responsável por determinar os seus formatos adequados, em termos da separação fonética e tonicidade.

Para entender a definição da ***Análise Sintática*** como uma das partes integrantes do ***Sistema LEIA*** considere o termo de maneira individual. Tem-se que a *análise* significa decompor um todo em suas partes constituintes, enquanto que *sintaxe* trata da relação lógica das palavras na frase. Integrando os conceitos, pode-se dizer que a ***Análise Sintática*** implica na decomposição de uma frase em seus elementos constituintes, a fim de verificar a relação lógica existente entre eles. Logo, é um processo que provê informações necessárias para a correta pronunciação das palavras e mais alto nível de *prosódia*, ou seja, a entoação da frase e a variação na proeminência que os locutores humanos podem usar.

---

[7] Estas palavras também são denominadas de *homônimas homógrafas*. Como exemplo podem ser citados os vocábulos *colher*, *molho*, *acordo* e *sede*.

Então, informações de partes da fala, obtidas a partir de uma ***Análise Sintática***, são usadas para fazer decisões de limites de sentença e de tonicidade. Para tanto, as palavras devem ser classificadas - rotuladas - em dois grupos:

- *Palavras Funções*: preposições, artigos, conjunções, etc.
- *Palavras Conteúdo*: nome, verbo, advérbio, etc.

Analisando a tonicidade em nível de palavra, tem-se que as *palavras conteúdo* são mais proeminentes do que as *palavras funções*. É, porém, imprescindível considerar a tonicidade em nível de sentença: o agrupamento de palavras em uma sentença[8] faz com que as frases soem com mais naturalidade. Logo, esta classificação realizada na ***Análise Sintática***, além de determinar a proeminência das palavras, também determina a entoação da frase, já que as pausas mentais entre as sentenças serão determinadas com base nestas informações.

Por último, a etapa de ***Análise Sintática*** ainda fornece informações para o ***Processamento Semântico***, já que em alguns casos as pronúncias das palavras homógrafas heterofônicas podem ser diferenciadas a partir da sua classificação gramatical: a pronúncia do primeiro ***o*** da palavra ***acordo***, por exemplo, pode apresentar um som aberto ou fechado caso ela represente um verbo ou um substantivo, respectivamente.

A etapa do ***Processamento Semântico*** é ativada caso tenha sido identificada alguma palavra no ***Dicionário de Exceções***, responsável por armazenar *as palavras homógrafas heterofônicas*. Neste caso, estas palavras são associadas a sua correta pronúncia. Convém ainda destacar que há dois tipos de tratamento dentro deste módulo: em certos casos, podem-se distinguir as palavras a partir de sua classificação - é o caso, já citado, da palavra ***acordo***. Entretanto, há situações em que apenas este procedimento não é suficiente: o primeiro ***e*** da palavra ***sede***, quando substantivo, pode assumir o som aberto ou fechado. Essa ocorrência evidencia a necessidade de se considerar o contexto para possibilitar a correta identificação do som das unidades.

Após terem sido submetidas ao ***Separador de Unidades*** e ao ***Processamento Semântico***, as palavras estão prontas para passarem pela etapa de ***Conversão Letra-Fonema.***

---

[8] Em geral, este agrupamento de palavras em uma sentença é denominado de unidades prosódicas.

O texto, na sua forma ortográfica, é uma maneira simples e eficiente de representar a comunicação por escrito, já que ele é originado a partir de um alfabeto e este, por sua vez, apresenta um número bastante limitado de símbolos (*letras*). Entretanto, na comunicação através da fala, não se pode dizer o mesmo, ou seja, as palavras apresentam uma variedade enorme de sons (*fonemas*): não há uma correspondência "um a um" entre as *letras* e os *fonemas*. Um exemplo bem evidente é a *letra* ***x***, que pode assumir os *fonemas* ***x***, ***s***, ***z*** e ***ks***, como nas palavras ***lixo***, ***texto***, ***exame*** e ***fixo***, respectivamente.

Logo, como o objetivo principal do ***Sistema LEIA*** é obter uma fala natural, se faz necessário representar as palavras do texto em uma formato que mais se aproxime da maneira como são faladas e, portanto, as considerações feitas no parágrafo anterior devem ser levadas em conta. A referida representação é uma transcrição das palavras em uma forma gráfica que simboliza os *fonemas*. A palavra ***fixo*** seria convertida em "***fikso***".

Como as palavras estão sendo preparadas para adquirirem um formato adequado para serem buscadas num ***Dicionário de Unidades Acústicas***, convém determinar um alfabeto para representar a *transcrição letra/fonema*. Em geral, esta conversão obedece ao *Alfabeto Fonético Internacional* (*AFI*) - ilustrado na *figura 4.3* -, que utiliza *letras* latinas e gregas para representar os sons das palavras. O *AFI* é um instrumento muito útil para os linguistas, pois suas convenções foram cuidadosamente definidas e aceitas por foneticistas de diversos países, o que permite um rápido reconhecimento dos sinais encontrados em livros. Mas, duas inconveniências surgem imediatamente. Analisando a simbologia do citado alfabeto, pode-se ver claramente que ele não apresenta uma representação tão amigável. Por outro lado, como já citado anteriormente, não há uma correspondência única e direta entre *letra* e *fonema*. Assim, objetivando tornar o processo de transcrição ortográfico/fonético mais simples, foi construído um alfabeto representativo direcionado à Língua Portuguesa - *quadro 4.1* - para simbolizar o seu universo de *fonemas* e, além do mais, determinar as *letras* que assumem mais de um *fonema*. Em paralelo, é essencial desenvolver um conjunto de regras para relacionar a *letra* ao *fonema* corretamente.

Após o processo de conversão, segue-se a ***Análise Fonológica***. Ela é responsável por associar os *fonemas* aos seus *alofones*, permitindo, então, particularizar os ***SCTF*** de acordo com os sotaques, pois analisando a fala de duas pessoas de regiões diferentes, por exemplo nordeste e sul, verifica-se que elas pronunciam uma mesma palavra diferentemente: a *letra* ***e***

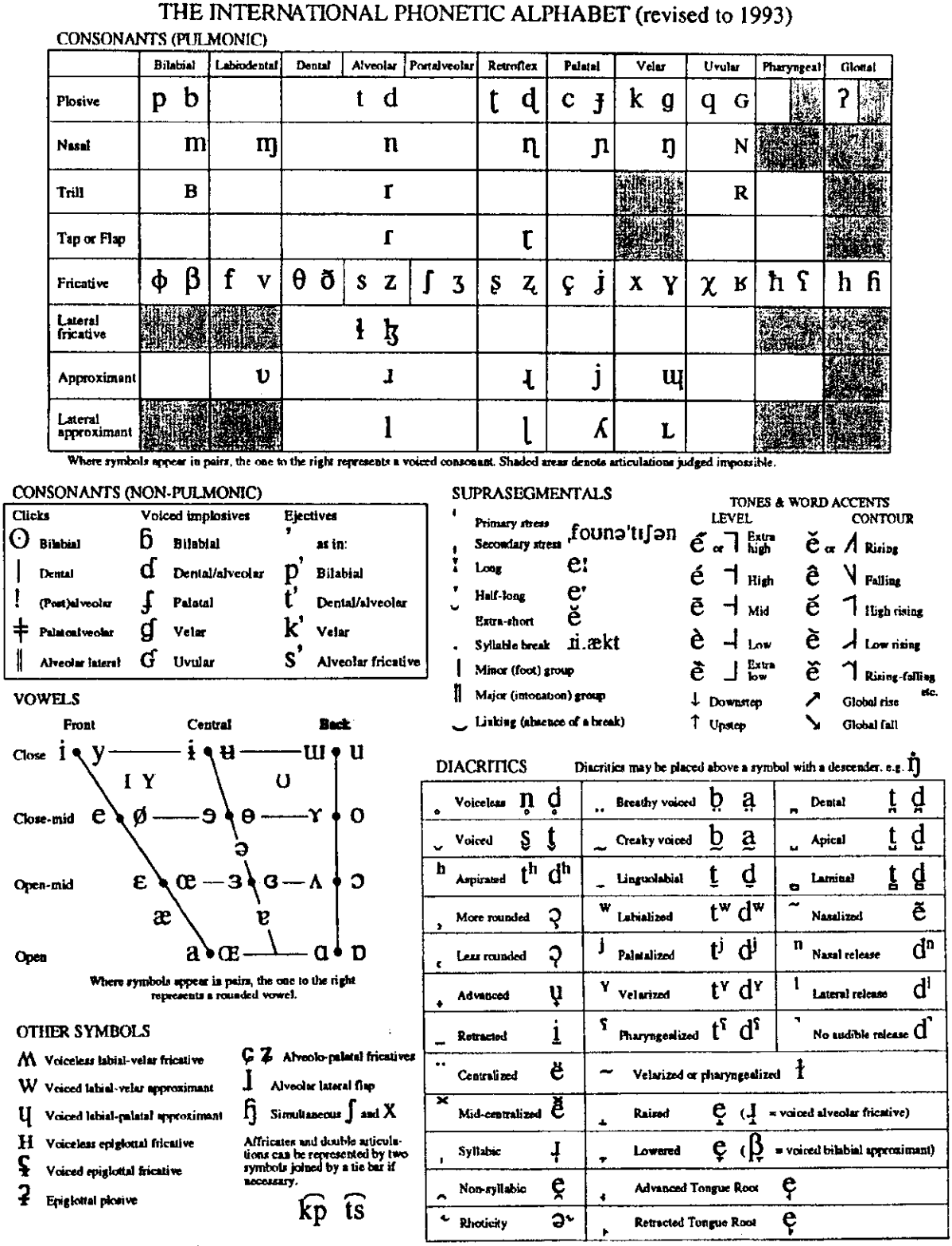

Figura 4.3: Alfabeto Fonético Internacional

A *prosódia* diz respeito à pronúncia regular das palavras, a qual apresenta-se vinculada a três fatores: acento, pausa e entoação. Assim, pode-se dizer que o ***Modelamento Prosódico*** consiste num tratamento realizado nos *fonemas*, para que algumas de suas características sofram alterações quantitativas e a fala produzida se torne o mais semelhante possível à fala humana. Essas alterações englobam modificações quanto à *freqüência fundamental* e à *duração* dos *fonemas*. Logo, a determinação da *prosódia* apropriada para uma pronúncia sintética envolve, pelo menos, a determinação de um tempo adequado para as palavras serem sintetizadas, bem como sua associação a um contorno entonacional adequado.

Na realidade, os processamentos dentro deste módulo são complexos. Para auxiliá-los, tem-se duas fontes de informações: ***Análise Sintática*** - que já foi citada em parágrafos anteriores - e o conjunto de regras, fundamental para a determinação da tonicidade em nível de palavra.

Após terem passado por esta série de etapas, as palavras assumem um formato destacando suas sílabas e tonicidade em nível de palavra e de frase. O próximo passo consiste em originar as formas de ondas propriamente ditas. Para tanto, se faz necessário a concepção de um ***Dicionário de Unidades***. Mas, cada unidade armazenada apresenta valores associados à *freqüência fundamental* e à *duração*, que precisam ser ajustados para atingirem o valor desejado, que são realizados na etapa de ***Ajustes de Unidades***.

Por fim, de posse das unidades que formarão os vocábulos, tem-se a ***Síntese*** propriamente dita, ou seja, a onda sonora é gerada e, assim, a fala é originada.

4.3 Tratamento Prosódico

4.3.1 Considerações Gerais

Antes de começar a tratar o desenvolvimento das etapas integrantes do ***Sistema LEIA***, é interessante enfocar o *tratamento prosódico* de uma forma mais detalhada, por ser um

ponto que ainda merece ser bastante estudado. A importância das características *prosódicas* justifica este direcionamento, pois só assim se conseguirá validar a estrutura proposta.

A *prosódia* pode ser encarada como uma parte das mais desafiadoras do ***SCTF***. É responsável por definir atributos que conduzirão à obtenção de uma fala natural e inteligível. Logo, é interessante comparar os aspectos vinculados à oração relacionados à língua falada e a escrita.

A oração na língua falada conta com numerosos recursos para alcançar seu objetivo de unidade de comunicação. Em seu auxílio, além dos elementos linguísticos de que dispõe o idioma, há uma série de recursos extralinguísticos elocucionais - riso, suspiro, bocejo -, bem como não elocucionais - mímica. Na língua escrita entram em jogo outros fatores: o recurso da entoação desaparece; ela tem que ser deduzida do texto pelo leitor mediante uma técnica bem conhecida que é a leitura. Por outro lado, o leitor encontra-se muito distante no tempo e no espaço, e não é em regra um indivíduo determinado e conhecido pelo escritor.

Com base nas considerações acima citadas, pode-se perceber que a fala não é um processo simples. A entoação e as pausas tornam o desenvolvimento do ***SCTF*** um problema não trivial. Mas, se estes fatores não fossem evidenciados, a fala produzida não apresentaria as características desejadas. Então, certos procedimentos - discutidos a seguir - devem tornar possível a realização do conversor.

4.3.2 Modelamento Prosódico

O ***Modelamento Prosódico*** pode ser representado por dois procedimentos principais: tratamento da pronúncia das palavras e identificação das pausas.

4.3.2.1 A Pronúncia das Palavras e as Pausas

A pronúncia das palavras está ligada a uma série de fatores. Inicialmente, serão considerados os seguintes:

- classificação gramatical;
- tonicidade das palavras.

A *classificação gramatical* é essencial para o tratamento dos vocábulos em nível de sentença, permitindo a identificação daqueles que serão proferidos com maior ênfase. A *tonicidade das palavras*, por sua vez, é fundamental para as considerações a nível de palavra, já que informa qual sílaba é mais proeminente, qual apresenta maior destaque. Então, com base nestas afirmações, um questionamento pode surgir: como identificar a sentença no texto a ser convertido?

A resposta para o problema está diretamente relacionada com a identificação das *pausas*, que podem ser vistas como elementos delimitadores das sentenças.

Assim, relacionadas as *pausas*, duas situações surgem. Inicialmente, tem-se aquelas derivadas diretamente dos sinais de pontuação. Entretanto, há outras ocorrências que não apresentam nenhum símbolo representativo e que são conhecidas como *pausas mentais*.

A explicação para o direcionamento acima proposto pode ser dada citando o seguinte fato: a partir de estudos, chegou-se à conclusão de que a identificação das palavras deveria ser feita com base em análises em nível de oração, pois embora o falante tenha liberdade de escolher os vocábulos quando elabora uma, ele não pode criar a estrutura em que os vocábulos se combinam na comunicação de suas idéias: as estruturas oracionais obedecem a certos modelos formais que podem não ser coincidentes de uma língua para outra e que constituem os padrões estruturais. Entretanto, a análise do vocábulo em si também é importante, pois sabe-se que existe uma sílaba que se sobressai sobre as outras e, além do mais, é interessante para se estabelecer uma visão do possível comportamento dos parâmetros prosódicos - *frequência fundamental* e *duração* -, que permitirá a determinação de uma tendência para as várias unidades, facilitando o processo de ***Ajuste de Unidades*** e contribuindo para uma fala mais natural e inteligível.

A identificação da tonicidade em nível de frase será feita com base em informações geradas a partir da ***Análise Sintática***, isto é, do rotulamento das palavras - *palavras funções* (menos proeminentes) e *palavras conteúdo* (mais proeminentes). Esse fato origina o problema de identificar as várias classes gramaticais, ou seja, artigo, pronome, substantivo, verbo, ...

A determinação da tonicidade em nível de palavra consiste em determinar a sua sílaba tônica. Para possibilitar a realização desta etapa, é essencial a formulação de um conjunto de ***regras*** - que será mostrado no próximo capítulo.

Com relação às *pausas mentais*, elas são localizadas a partir de uma investigação da tendência da Língua Portuguesa em realizar estas interrupções. As informações fornecidas pela ***Análise Sintática*** são fundamentais para identificar, no texto, o local da parada.

Pode-se perceber claramente que, antes de tudo, é importante investigar uma maneira de caracterizar as estruturas oracionais da Língua Portuguesa. De uma forma geral, três fatores podem ser citados:

- associação dos vocábulos de acordo com sua função sintática;
- concordância dos vocábulos de acordo com certos princípios fixados na língua;
- ordem dos vocábulos de acordo com sua função sintática.

Assim, tentando-se chegar a uma generalização dos fatores descritos anteriormente, será considerada a "***Sintaxe de Colocação***".

4.3.3 Sintaxe de Colocação

A ***Sintaxe de Colocação*** trata a maneira de dispor os termos dentro da oração e as orações dentro do período. Logo, a colocação, dentro de um idioma, obedece a tendências variadas, quer de ordem estritamente gramatical, quer de ordem psicológica e estilística. Entretanto, o maior responsável pela ordem favorita numa língua parece ser a entoação oracional.

Até o momento, se falou muito em entoação, mas ainda não foi formalizada nenhuma definição, como também não foi especificada a característica encontrada no português. Assim, pode-se dizer que entoação consiste na maneira como são proferidas as orações dentro de certa cadência melódica. Logo, a parte final de uma oração é sempre marcada por alguns tipos de entoação: declarativa, interrogativa, exclamativa ou pausal. A língua portuguesa, por sua vez, se caracteriza pelo ritmo ascendente, em que se anuncia o termo menos importante e

depois, com acentuação mais forte, a informação nova e de relevância para o ouvinte. Esse fato leva a uma ordem considerada habitual que consiste em enunciar o sujeito, depois o verbo e em seguida seus complementos. Quando este esquema é desobedecido tem-se uma ordem inversa.

Nos próximos capítulos, portanto, serão mostrados procedimentos que podem ser adotados para representar os vocábulos de forma a possibilitar a conversão texto-fala.

CAPÍTULO 5

DESCRIÇÃO DOS MÓDULOS DO SISTEMA *LEIA*

Este capítulo descreve, em detalhes, os módulos do ***Sistema LEIA****, onde os procedimentos sugeridos são apresentados e explicados.*

5.1 Introdução

A concepção do ***Sistema LEIA***, representado pela estrutura proposta na *figura 4.2*, exige o desenvolvimento de um *ambiente de trabalho* no qual os vários módulos devem ser implementados. Assim, os procedimentos e critérios considerados são discutidos a seguir.

5.2 Tela de Apresentação

O acesso ao ***Sistema LEIA*** é realizado a partir da tela principal ilustrada na *figura 5.1*. É verdade que o objetivo consiste na conversão texto-fala de forma automática, o que eliminaria a representação individual de cada etapa. Entretanto, para permitir um acompanhamento mais detalhado do Sistema e, assim, conseguir determinar os pontos que merecem ser otimizados, ele é apresentado através dos vários módulos integrantes.

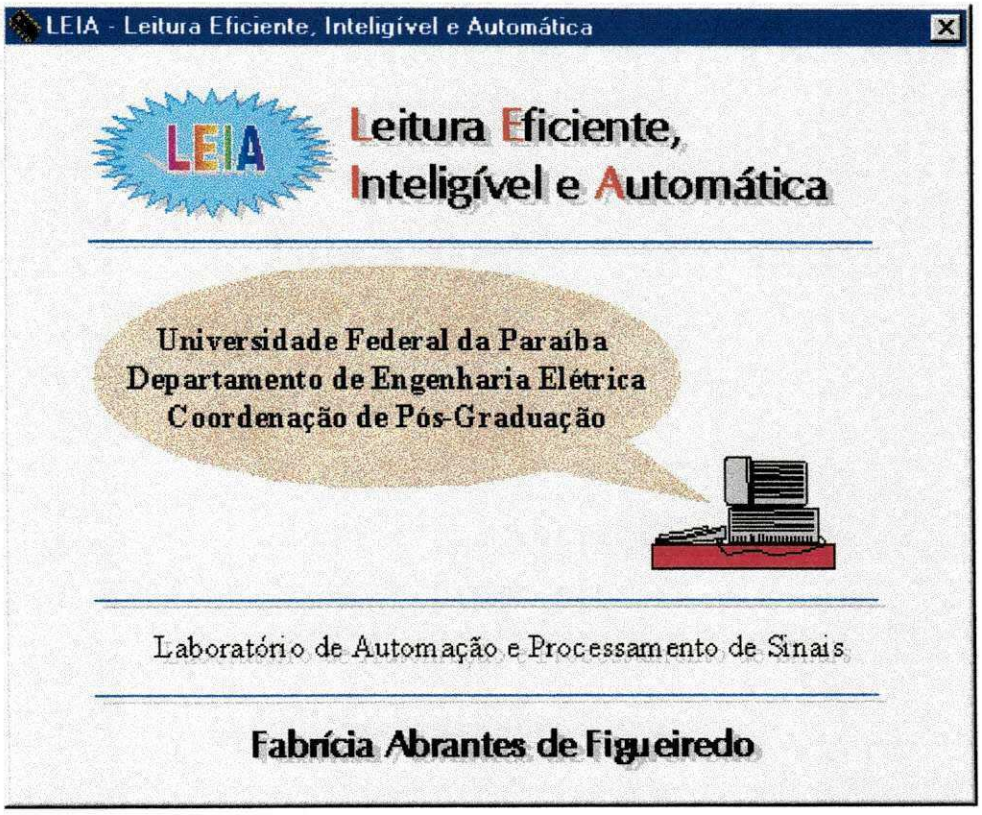

Figura 5.1: Tela de Apresentação do Sistema LEIA

5.3. Normalização do Texto

Esta etapa, em uma primeira vista, não apresenta grandes dificuldades. Entretanto, a

medida que as ocorrências possíveis começam a ser levantadas, a afirmação perde o sentido. A seguir são evidenciadas, portanto, as considerações que devem ser levadas em conta para permitir o desenvolvimento de um pré-processador do texto capaz de transformá-lo em um formato adequado ao processamento do sinal.

O algoritmo para proceder a *normalização do texto* pode ser concebido a partir de três divisões: a primeira, responsável em dar um tratamento adequado às *palavras abreviadas;* a segunda, por sua vez, encarregada de determinar o formato correto dos *algarismos;* e, por fim, uma destinada às *siglas.*

Na realidade, cada divisão - citada no parágrafo anterior - deve apresentar uma espécie de dicionário, que conterá informações essenciais para que o texto atinja um formato adequado. Além desses dicionários, outros pontos merecem ser considerados para que um sistema mais robusto possa ser originado. De início, pode-se destacar que em algumas situações as siglas são pronunciadas conforme a escrita - como é o caso de *CPF* -, enquanto que outras são lidas - como na *sigla INCRA.* Mas, em certas situações, dúvidas podem surgir no que diz respeito à expansão da *sigla.* Quando da ocorrência da seqüência *BB*, o certo é pronunciar "*bebe*" ou "*Banco do Brasil*"?

A princípio, pode-se achar que seria mais conveniente obter a forma por extenso. Entretanto, é importante ressaltar que o objetivo do Sistema em estudo consiste na reprodução fiel do texto em fala. A não transcrição de uma *sigla* em sua forma extensa não implica em perda de naturalidade e inteligibilidade, que são as características desejadas. É verdade que ao promover a expansão, o texto tornar-se-á mais claro, facilitando, por exemplo, a interpretação. Mas esse ponto não é fator determinante do estudo em desenvolvimento.

Assim, para associar a forma adequada assumida pela *sigla*, se pensou em estabelecer uma série de *"máscaras"* representando as possíveis ocorrências. Entretanto, percebeu-se que, em muitas situações, sequências pertencentes ao mesmo padrão podiam tanto ser pronunciadas da forma como se apresentava na escrita quanto obedecer a transcrição das *letras.* Considerando *"USP"* e *"UNP"* pode-se justificar a afirmação anterior, pois a primeira sigla é lida - *"usp"* -, enquanto que a segunda é traduzida para *"uenepe"*. Assim, com a finalidade de tornar o Sistema o mais geral possível, o processo de determinação do formato mais

conveniente contará com um dicionário onde serão armazenadas as *siglas* e sua devida transcrição.

Com relação às *palavras abreviadas*, poderia ser considerada a mesma linha de pensamento estabelecida para as *siglas*. Mas, deixar de realizar a transcrição neste caso vai de encontro a maneira de originação de uma leitura. No seguinte trecho, por exemplo, *"... Campina Grande, PB ..."*, o leitor poderá pronunciar *"pebe"* ou *"Paraíba"*. Entretanto, na ocorrência de *"O Sr. José ..."*, com certeza o *"Sr."* será lido como *"Senhor"* e não como *"esse erre ponto"*. Assim, surge a necessidade de se identificar as abreviações.

As *palavras abreviadas* não vem, necessariamente, finalizadas por um ponto (.) - como é o caso de ha[9]. O ponto (.), por sua vez, também pode caracterizar o final de uma sentença e, assim, pode vir após um vocábulo normal ou mesmo seguir uma *sigla* - *... é o número do meu RG.* Por fim, os *algarismos* podem apresentar três formatos: um primeiro engloba os casos que representam valores monetários, números de casas, código postal, etc. Outro tem uma abrangência mais particular, pois se preocupa apenas com as datas. E o último atua apenas nos casos que simbolizam números de telefones.

5.3.1 Algoritmo1 (Normalização do Texto)

O *algoritmo 1*, de uma forma geral, pode ser representado conforme o *quadro 5.1*. Ele foi estruturado a partir das considerações citadas na seção anterior. Entretanto, para possibilitar a ocorrência de todas as situações especificadas, é necessário estabelecer alguns pontos:

- Os vocábulos do texto podem ser classificados em quatro classes: *formato adequado*, *sigla*, *algarismo* e *abreviação*;
- Se o vocábulo for uma sequência de *letras maiúsculas*, tem-se uma *sigla*;
- Se o vocábulo for uma sequência de *números*, tem-se um *algarismo*;

---

[9] *ha* é a abreviação de *hectare*.

- Se o último caracter for um ponto (.), deve-se verificar os caracteres anteriores:
- se os caracteres anteriores forem *maiúsculos*, tem-se uma *sigla*;
- se os caracteres anteriores fazem parte de um dicionário de *abreviações*;
- senão o vocábulo já se encontra no *formato adequado.*
- Se o vocábulo não se enquadrou em nenhuma consideração acima, então deve-se verificar se ele faz parte do *dicionário de exceções de siglas* (sem o sinal de pontuação);
- Se nenhuma das afirmações anteriormente citadas não foi verificada, então o vocábulo já apresenta um *formato adequado.*

```
repeat até encontrar o fim do texto
        ler palavra
        if (palavra = SIGLA)
                acionar o dicionário de siglas
                else if (palavra = ALGARISMO)
                        determinar o formato adequado
                        else if (palavra = ABREVIAÇÃO)
                                proceder a expansão
```

Quadro 5.1: Representação do Algoritmo 1 (Normalização do Texto)

Para melhor entender o objetivo deste módulo, algumas ocorrências, e suas respectivas expansões, são mostradas no *quadro 5.2.*

5.4 Separador de Unidades

Como já abordado, o ***Sistema LEIA*** adota a *Concatenação de Unidades Acústicas - difones, trifones, polifones* -, as quais são previamente gravadas e armazenadas para originar a fala. A utilização das referidas unidades acústicas resolvem, em parte, o problema da coarticulação.

Mas, por outro lado, surge a questão de identificar que unidades devem ser resgatadas do dicionário para formar as palavras do texto.

Uma possível saída consiste em analisar as palavras contidas no texto que se deseja converter e, assim, determinar os locais onde podem ser inseridas as quebras. Para tanto, as várias possibilidades de posicionamento das vogais e das consoantes devem ser estudadas, para se estabelecer uma tendência (regra) geral. Na realidade, isto é semelhante ao processo de separação de sílabas. Entretanto, um ponto chave deve ser considerado: como o objetivo é identificar as unidades que irão originar a fala, deve-se considerar o modo como os vocábulos realmente são pronunciados. Por exemplo, sejam as palavras "*cooperação*" e "*colaboração*". A regra de separação, quando aplicada a primeira, estabelece uma quebra entre o encontro vocálico "*oo*". Mas, quando esta é pronunciada, a referida separação é eliminada. O encontro vocálico é proferido de uma única vez seguindo a *letra c*. A unidade */coo/* pode se vista como sendo igual a unidade */co/* do primeiro vocábulo. A diferença entre elas diz respeito apenas às suas durações.

Ocorrência	Expansão
A camisa custa R$ 45,00.	a camisa custa quarenta e cinco reais.
Rua Floriano Peixoto, 121	rua floriano peixoto, cento e vinte e um
Dia 25/04/97 ...	dia vinte e cinco de abril de noventa e sete ...
Fone 231-2323	fone dois três um dois três dois três
A casa é grande.	a casa é grande.
O meu CPF é ...	o meu cepeefe é ...

Quadro 5.2: Normalização do Texto

5.4.1 Regras para Estabelecer a Quebra das Palavras

Para proceder a identificação das unidades que formam as palavras do texto,

determinadas regras e procedimentos foram considerados. As normas gramaticais, mais uma vez, figuraram como fatores indispensáveis. Mas, devido a riqueza do idioma português, bem como a derivação de muitas palavras a partir de outras línguas, certas adequações devem ser feitas para possibilitar a separação correta das unidades de forma direcionada à obtenção da fala.

A identificação das quebras baseia-se em dois pontos, ou seja, uma *análise em nível de palavra* - a fim de descobrir uma tendência de ocorrência das *letras* - e outra em *nível de pronúncia.* Para o primeiro, o procedimento adotado pode ser ilustrado a partir do *quadro 5.3*, que exibe algumas sequências possíveis das *letras.*

O segundo ponto - *análise em nível de pronúncia* - pode ser encarado como sendo um refinamento. Considerando a palavra *caminho*, tem-se que sua separação apresenta o seguinte formato: /ca-mi-nho/. Mas, observando a maneira como o vocábulo realmente é pronunciado, a separação assume o padrão /ca-min-ho/.

5.4.2 Algoritmo 2 (Separador de Unidades)

A segmentação de unidades pode ser vista como um processo repetitivo. A partir do estudo prévio descrito acima, "*máscaras*", contendo sequências de *letras* proferidas sem quebras, são determinadas. Cada vocábulo do texto é analisado e quando da concordância entre a sequência da palavra e as "máscaras" representativas, a palavra é separada. O restante, por sua vez, é submetido a uma nova verificação.

O procedimento de verificação *vocábulo/máscara* exige o estabelecimento de prioridades que dependem do número de *letras* sequenciais consideradas. Logo, a disposição das "*máscaras*" segue um formato decrescente, sendo que o maior conjunto apresenta seis *letras.* Foram definidos, desta forma, vinte e cinco padrões de "*máscaras*", os quais apresentam uma série de considerações que determinarão o formato da separação dos vocábulos.

As abordagens realizadas acima são fundamentais para a separação em *nível de palavra.* Entretanto, o Sistema apresenta como entrada um texto, exigindo um direcionamento para a

separação em *nível de sentença.* Convém , portanto, analisar o comportamento do fim e do início das palavras. Há situações em que não se verifica uma pausa entre dois vocábulos, bem como outras em que as *letras* final e inicial se unem e modificam seus sons. Nestes casos, as palavras devem ser consideradas como sendo apenas uma, pois, caso contrário, não serão determinadas as *unidades acústicas* adequadas. O conjunto de regras [1] – designado na estrutura geral (*figura 4.2*) - é apresentado a seguir.

Seqüências	Seqüências Separadas	Exemplo
CCVCCCVCVVCVC	CCVCC / CV / CV / V / CVC	trans / na/ ci / o / nal
CVCV	CV / CV	ca / ma
CVVCVC	CV / V / CVC	co / a / gir
VCCVCVCVV	VC / CV / CV / CVV	al / te / ra / ção
VVCVCVC	VV / CV / CVC	au / di / tor
CVCCVCV	CVC / CV / CV	pan / ca / da
CCVCV	CCV / CV	tre / vo
CCVCCV	CCVC / CV	plan / ta
CCVCCCVCVCCVV	CCVCC / CV / CVC / CVV	trans / pa / rên / ci / a
CCVCCCCVCCV	CCVCC / CCVC / CV	trans / plan / te
CVVVC	CVV / VC	mai / or
CVVVCVV	CVV / V / CVV	mai / o / ria
CVVCCV	CV / VC / CV	fa / ís / ca

Quadro 5.3: Análise - Separação das sílabas[10]

[10] O *C* e o *V* representam uma consoante e uma vogal, respectivamente.

5.4.3 Regras para Modelar o Comportamento da Coarticulação

As regras para modelar o comportamento da coarticulação são as seguintes:

- ***Regra 1***: Quando uma palavra termina em *vogal tônica* e a seguinte é iniciada com *vogal* ou *ditongos tônicos*;
- ***Regra 2***: Quando uma palavra termina em *vogal tônica* e a seguinte é iniciada com *vogal* ou *ditongo átonos*;
- ***Regra 3***: Quando uma palavra termina em *vogal átona* e a seguinte é iniciada por *vogal tônica*;
- ***Regra 4***: Quando uma palavra termina com a *consoante b* e a seguinte é iniciada por outra *consoante*, o *fonema /b/* deve ser proferido sem o aparecimento dos *fonema /i/* e */e/* intermediariamente;
- ***Regra 5***: Quando uma palavra termina com a *consoante b* e a seguinte e iniciada com *vogal* ou *semivogal*, as palavras podem ser unidas diretamente ou com o apoio dos *fonemas /i/* e */u/*;
- ***Regra 6***: Quando uma palavra termina com uma *consoante* e a seguinte é iniciada com a mesma *consoante* os vocábulos são proferidos sem pausa, entretanto os dois *fonemas* permanecem;
- ***Regra 7***: Quando uma palavra termina com a *consoante s* e a seguinte é *vogal* os vocábulos são proferidos sem pausa e o */s/* assume o som de */z/*.

Obs.: As regras 1 e 2 permitem estabelecer uma generalização: se a palavra for terminada por *vogal tônica* e a seguinte for iniciada por *vogal* tem-se a junção das palavras. A representação de todas as regras citadas anteriormente pode ser vista no *quadro 5.4*.

Com a formulação das sete regras acima citadas surge a necessidade de se estabelecer um procedimento para identificar as *sílabas tônicas* e *átonas* das palavras. Para tanto, um outro

conjunto de regras [1] - ***REGRAS 3*** - deve ser considerado, sendo que este, além de fornecer informações para resolver o problema da tonicidade, também é de grande importância para o *modelamento prosódico*, quando da determinação das proeminências em *nível de sentença* e em *nível de palavra*. Em seções posteriores estes algoritmos serão descritos.

```
repeat até encontrar o fim do texto
    ler pal1
    if (pal1 for terminada por vogal)
        if vogal for tônica
            ler pal2
            if (pal2 for iniciada por vogal)
                (paln = pal1 + pal2)
        else if vogal for átona
            ler pal2
            if pal2 for iniciada por vogal tônica
                (paln = pal1 + pal2)
    else if (pal1 for terminada por consoante)
        if (consoante for "b")
            (paln = pal1 + pal2)
        else if (consoante for diferente de "b")
            ler pal2
            if (pal2 for iniciada com a mesma consoante)
                (paln = pal1 + pal2)
        else if (consoante for "s")
            ler pal2
            if (pal2 for iniciada por vogal)
                (paln = pal1 + pal2)
```

Quadro 5.4: Representação das Regras para Modelar o Comportamento da Coarticulação

5.5 Dicionário de Exceções

O ***Dicionário de Exceções*** reúne um conjunto de palavras que apresentam a mesma

grafia, mas são pronunciadas de forma diferente. Assim, como visto, duas divisões podem ser evidenciadas. A primeira enquadra os vocábulos que tem seus sons diferenciados a partir de sua classe gramatical - por exemplo a palavra *acordo*. A segunda, por sua vez, considera as palavras que apresentam a mesma classe gramatical, porém possuem sons diferentes - como é o caso do vocábulo *sede*.

Assim, o texto é submetido a tal dicionário. As *palavras homógrafas heterofônicas* são determinadas, sendo posteriormente classificadas segundo a divisão considerada no parágrafo anterior. Em seguida, a partir da ***Análise Sintática***, a cada palavra é associada a sua classe gramatical. Para os vocábulos pertencentes a primeira divisão a identificação das *unidades acústicas* é obtida sem maiores problemas. Entretanto, os da segunda divisão exigem uma análise contextual. Então, um possível caminho a ser buscado consiste em analisar a vizinhança dos vocábulos, que devem estar devidamente rotulados.

5.6 Análise Sintática

Analisando a estrutura geral do sistema – mostrada na *figura 4.2* -, pode-se concluir que a A***nálise Sintática*** gera informações fundamentais à determinação das *pausas mentais*, das entoações e das proeminências das palavras e sentenças. Entretanto, a importância deste módulo é contrastada com a sua complexidade, já que a língua portuguesa não apresenta uma estrutura única para a formação e classificação das frases e palavras.

Assim, como já abordado em capítulos anteriores, o módulo ***Análise Sintática*** objetiva realizar um rotulamento das palavras e, então, classificá-las em duas categorias - *funções* e *conteúdo*. Logo, em uma primeira análise, a complexidade desta etapa poderia ser combatida, pois definindo-se os artigos, as preposições, as conjunções e os pronomes ter-se-ia identificado as *palavras funções*. As palavras restantes, portanto, seriam as *palavras conteúdo*. Entretanto, a obtenção de um sistema de conversão com resultados naturais e inteligíveis exige informações mais detalhadas: algumas situações exige a determinação da classe gramatical da palavra.

As classes gramaticais dos vocábulos não são determinadas diretamente, pelo menos para a Língua Portuguesa, já que esta não apresenta uma regra geral para a disposição das palavras. É verdade que esta classificação pode ser realizada a partir da *sintaxe de colocação* - detalhada no capítulo anterior -, que tenta generalizar as estruturas oracionais. Mas, mesmo assim, tem-se muitas variações dessas estruturas, impossibilitando o rotulamento detalhado de todas as palavras do texto.

5.6.1 Análise Sintática e a Proeminência das Palavras

Uma das funções da ***Análise Sintática*** consiste em determinar a proeminência das palavras, isto é, que palavras são pronunciadas com maior ênfase. Assim, analisando um locutor, pode-se verificar que ele dá um maior destaque aos verbos, substantivos e adjetivos quando está falando, ao passo que os artigos, conjunções e pronomes não são tão evidentes.

A determinação, portanto, da proeminência das palavras pode ser realizada sem grandes problemas. Na realidade, esta afirmação deve ser melhor analisada. Apesar da Língua Portuguesa permitir a identificação dos artigos, conjunções, preposições e pronomes através de suas definições - pois eles apresentam um número limitado, tornando viável seus armazenamentos -, há ocorrências onde uma palavra, dependendo de sua localização, apresenta uma determinada classe gramatical. A *letra "a"*, por exemplo, pode representar um artigo ou uma preposição. Considerando um exemplo mais complexo, o vocábulo "*entre*" pode ser uma preposição ou um verbo.

Dessa forma, formalizando as definições das *palavras funções* e estabelecendo algumas condições de identificação, os vocábulos podem ser classificados. Aqueles que não fazem parte desta categoria são rotulados como sendo *palavras conteúdo.*

5.6.2 A Análise Sintática e as Pausas Mentais

A identificação e determinação das *pausas mentais* exige um rotulamento mais detalhado das palavras, ou seja, exige a classificação gramatical de cada vocábulo.

Na seção anterior foi sugerida a definição de algumas classes gramaticais. Entretanto, definir os substantivos, os verbos e os adjetivos é um procedimento inviável: seria mais aconselhável optar por um ***SVA*** se a única saída fosse realizar a referida definição.

Um possível e viável caminho a ser seguido consiste em identificar as palavras a partir da composição de uma oração, que, em geral, apresenta os termos sujeito e predicado. Mas, inúmeras variações podem ocorrer: há orações sem sujeito, com sujeito indeterminado, há orações em que o predicado vem antes do sujeito, etc. Por outro lado, não existe uma estrutura definida para o sujeito e para o predicado. Assim, a identificação dos vocábulos não é tarefa fácil. Na realidade, pode-se afirmar que de imediato não é possível classificar todos os vocábulos do texto. Entretanto, é importante estabelecer de início, pelo menos, uma tendência geral. Para tanto, será analisado um conjunto de situações, a partir do qual serão estabelecidas as possíveis ocorrências.

5.6.2.1 Identificação das Classes Gramaticais

Algumas classes gramaticais, como já citado, podem ser identificadas diretamente. Para tanto, se faz necessário apenas construir um arquivo de dados. Muito embora exista a questão de um vocábulo poder ser rotulado em mais de uma classe. Assim, nesse procedimento podem ser enquadrados os artigos, pronomes, preposições e conjunções.

A identificação das outras classes gramaticais apresentam bem mais problemas. É impossível representar todos os verbos, adjetivos e substantivos através da composição de um arquivo de dados. Assim, a saída é estabelecer regras que permitam representar a tendência da Língua Portuguesa.

A análise realizada a partir de um conjunto de frases permitiu identificar ocorrências gerais, tais como: os artigos se antepõem aos substantivos; após um pronome (pessoal) ocorre um verbo. Entretanto, a determinação de todas as possíveis situações é um desafio. Em seções posteriores serão descritos os procedimentos adotados, bem como as limitações existentes.

5.6.3 Determinação das Pausas

A determinação das pausas será realizada com base em dois algoritmos - de *colocação de marcas* e de *eliminação de marcas prosódicas* - desenvolvidos em [6]. No citado estudo, a identificação das quebras é denominada de fronteiras prosódicas, que se apresentam de vários tipos:

- início de frase;
- final de frase;
- sinais de pontuação;
- início de predicado;
- início de oração;
- início de complemento com preposição.

A seguir, portanto, tem-se a descrição dos citados algoritmos.

5.6.3.1 Algoritmo 3 (Colocação de Marcas Prosódicas)

A colocação das *marcas prosódicas* seguem quatro passos:

1. Os sinais de pontuação são simplesmente substituídos por suas *marcas* correspondentes;
2. É atribuída uma *marca de início de oração* imediatamente antes de cada conjunção;
3. É atribuída uma *marca de início de complemento com preposição* antes de cada preposição ou contração. Exceção feita à preposição "*de*" e suas contrações;
4. É atribuída uma *marca de início de predicado* imediatamente antes de cada verbo e antes de cada seqüência: pronome + verbo, advérbio + advérbio, advérbio + verbo e preposição + verbo.

5.6.3.2 Algoritmo 4 (Eliminação de Marcas Prosódicas)

Este algoritmo atribui pesos diferentes para cada tipo de *marca*. É um procedimento de estabelecimento de prioridades para identificar dentre os locais rotulados com as *marcas* aqueles nos quais a pausa realmente irá existir. Assim, tem-se um conjunto de situações determinando as *marcas*, que deve ser seguido ordenadamente.

1. As *marcas* referentes a sinais de pontuação e a início e final de frase recebem peso máximo e jamais são eliminadas;
2. *Marca de início de oração*;
3. *Marca de início de predicado*;
4. *Marca de início de complemento com preposição.*

Assim, o critério para eliminação das *marcas prosódicas* consiste na análise das vizinhanças[11]. Aquelas que apresentarem maiores pesos prevalecerão sobre as demais, ou seja, uma *marca* será eliminada se em sua vizinhança houver outra *marca* de peso superior ao seu.

5.6.4 Considerações em Nível de Palavra

Além de rotular as palavras, ou seja, classificá-las de acordo com sua classe gramatical, é fundamental desenvolver um procedimento para marcar a acentuação em *nível de palavra.* Esta consideração é importante para a concepção do ***Sistema LEIA***, pois quando os vocábulos são proferidos é evidente a presença de uma sílaba mais forte.

Assim, seguindo este pensamento, surgiu a necessidade de se pensar em como originar esse tratamento. Uma possível solução consiste em realizá-lo com base em regras que determinam a *sílaba tônica* para a Língua Portuguesa. A seguir, é descrito um conjunto formado por quatro etapas, responsável pela marcação da tonicidade em *nível de palavra.*

---

[11] Nesse trabalho, vizinhança de uma sílaba consiste nas cinco sílabas à esquerda e nas cinco sílabas à direita.

5.6.4.1 Marcação da Tonicidade em Nível de Palavra

A determinação da *sílaba tônica* das palavras será baseada em várias etapas descritas a seguir [20].

***Etapa 1*: Regra Geral**

A primeira etapa é considerada como regra geral, já que se deriva da própria ortografia, ou seja, as palavras dotadas de sinais que representam acentuação são transcritas para um formato que simboliza o acento. Entretanto, é interessante ressaltar que devem ser consideradas que certas *vogais* apresentam mais de um som, que será determinado a partir do acento. O *quadro5.5* ilustra este caso.

Obs.: As vogais *a*, *e* e *o* podem representar dois sons distintos - aberto e fechado. Então, para representá-las foram definidos dois *fonemas* para cada *letra*, ou seja, o *a*, *e* e *o* para representar o som fechado e *A*, *E* e *O* para representar o som aberto.

***Etapa 2*: Palavras não Acentuadas que Apresentam Maior Ênfase na Última Sílaba**

A segunda etapa - que pode ser exemplificada a partir do *quadro 5.6* - abrange as palavras não acentuadas com maior proeminência na última *sílaba*. Assim, coloca-se um acento simbólico na última *vogal* das palavras terminadas pelas sequências:

- vogal + l;
- vogal + r;
- vogal + z.

Dentro desta etapa ainda se enquadram as palavras terminadas pelas *vogais i* e *u*, exceto as sequências 'vogal + *i*' ou 'vogal + *u*': nestes casos, o acento é colocado na segunda vogal (direita para a esquerda).

Palavras	*Fonemas*/Acentuação
parâmetro	a (fechado) - par'ametro
rádio	a (aberto) - r'Adio
você	e (fechado) - voc'e
médico	e (aberto) - m'Edico
paralelepípedo	i - paralelep'ipedo
tônica	O (fechado) - t'onica
próximo	O (aberto) - pr'Oximo
açúcar	U - aç'ucar
coração	a (fechado) - coraç'ao

Quadro 5.5: Etapa 1 - Exemplos

Palavras	*Fonemas*/Acentuação
dedal	A (aberto) - ded'Au
ureter	E (aberto) - urEt'Er
rapaz	A (aberto) - rAp'Az
colibri	Kolibr'i
caramuru	KarAmur'u
samurai	A (aberto) - sAmur'Ai
museu	E (aberto) - muz'Eu

Quadro 5.6: Etapa 2 – Exemplos

Etapa 3: Palavras não Acentuadas que Apresentam Apenas uma Sílaba

A terceira etapa engloba as palavras monossílabas não acentuadas terminadas em *i* e *u*.

Nesta situação o acento simbólico é colocado nas referidas *vogais*. Mas se as *vogais* vierem antecedidas de outra *vogal* o acento é colocado antes da segunda *vogal* (direita para esquerda). Uma exemplificação dessas ocorrências é mostrada no *quadro 5.7*.

Palavras	*Fonemas*/Acentuação
Pai	a (aberto) - p'Ai
Mau	a (aberto) - m'Au
Tu	t'u
Vi	v'i

Quadro 5.7: Etapa 3 - Exemplos

Etapa 4: Palavras não Acentuadas que Apresentam Maior Ênfase na Penúltima Sílaba

A quarta etapa engloba as palavras que não se enquadraram nas regras anteriores, mais precisamente as palavras paroxítonas. Neste caso, o acento simbólico é colocado na segunda *vogal* (direita para a esquerda), exceto as palavras terminadas em *que* e *gue* que são acentuadas na terceira vogal. O *quadro 5.8* exibe exemplos destas situações.

Palavras	*Fonemas*/Acentuação
Sapato	A (aberto) - sAp'Atu
Cachorro	o (fechado) - kax'oRu
Rapaz	A (aberto) - rAp'Az
Almanaque	A (aberto) - Auman'A qe
Albergue	E (aberto) - Aub'Erge

Quadro 5.8: Etapa 4 – Exemplos

5.6.4.2 Algoritmo 5 (Marcação da Tonicidade a Nível de Palavra)

O *algoritmo 5*, descrito no *quadro 5.9*, tem a finalidade de identificar a tonicidade das palavras. Entretanto, a ele devem ser incorporados certos detalhamentos - que podem ser vistos no *quadro 5.10* - para permitir a identificação da tonicidade em *nível de vogal.* Isso é necessário devido ao procedimento de verificar o fim/início dos vocábulos (***REGRAS 1***).

5.7 Dicionário de Unidades

A concepção do dicionário de *unidades acústicas* pode ser considerada como uma etapa bastante importante do *sistema de síntese.* Logo, é fundamental estabelecer um conjunto de procedimentos essenciais e que implicam em pontos contribuintes à obtenção de uma fala natural e inteligível.

```
repeat até encontrar fim do texto
    ler palavra
    separar as sílabas
if palavra for acentuada
        Então sílaba tônica = sílaba acentuada
    else if palavra for terminada pelas seqüências (vogal + l ou vogal + r ou vogal + z)
        Então sílaba tônica = última sílaba
    else if palavra for  terminada em i ou u
        Então sílaba tônica = última sílaba
    else if palavra for monossílaba
        Então sílaba tônica = única sílaba
    else if sílaba tônica = penúltima sílaba
```

Quadro 5.9: Representação do Algoritmo 5 (Identificação da Tonicidade - Palavra)

O ponto de partida desta etapa consiste em determinar que unidades irão compor o

dicionário. Assim, como já enfatizado, o ***Sistema LEIA*** utiliza os *difones*, *trifones* e *polifones* devido ao problema da coarticulação. Entretanto, outras indagações podem surgir: que fator, além da coarticulação, reforça o uso das citadas *unidades acústicas*? Como elas podem ser obtidas?

Primeiramente, pode-se evidenciar que as unidades podem ser extraídas e armazenadas diretamente a partir do sinal de voz natural. Portanto, utilizando um software específico é procedida a gravação de palavras isoladas contendo as unidades desejadas. Mas, qualquer palavra pode ser usada?

Uma primeira resposta para a questão formulada poderia ser afirmativa. Entretanto, seria uma precipitação fazer tal comentário. Dois importantes pontos merecem ser lembrados [1]: a coarticulação deve ser evitada e deve compreender em seus limites porções espectralmente estáveis. Estas considerações conduzem ao estabelecimento de critérios para a escolha das palavras a serem gravadas.

```
ler a palavra
    separar as sílabas
        if (última sílaba = vogal)
    if (vogal for acentuada)
        vogal = tônica
            else if (palavra = monossílaba) and (última vogal for precedida de
                consoante)
            vogal = tônica
            else if (palavra = monossílaba) and (última vogal não for precedida de
                consoante)
                vogal = átona
                else (vogal = átona)
```

Quadro 5.10: Representação do Algoritmo 3 (Identificação da Tonicidade)

5.7.1 Critérios para a Escolha das Palavras a serem Gravadas

- A unidade desejada deve situar-se na parte central da palavra, já que nessa "região" é atingida uma estabilidade quando o vocábulo é enunciado. Nas extremidades tem-se as maiores variações. Então, pode-se concluir que as palavras apropriadas apresentam, no mínimo, três sílabas;
- As palavras que apresentam as *letras* ***p*** e ***a*** na sua formação devem ser as preferidas, já que facilitam o processo de segmentação e evitam problemas de coarticulação;
- Para as unidades formadas pelas sequências ***VC***[12] e ***CV*** deve-se optar pelas palavras iniciadas com vogal. Por outro lado, as unidades compostas pelas sequências ***CV*** e ***CC*** devem ser extraídas a partir das palavras iniciadas com consoantes.

5.8. Conversão Letra/Fonema: Preparação para o Resgate das Unidades

Para a realização da conversão *letra/fonema* e a identificação da *unidade acústica* que será buscada no dicionário, várias considerações devem ser feitas, ou seja:

- Devem ser definidos as *letras* e os *fonemas*;
- Devem ser analisadas as possibilidades de formação das palavras - vogais e consoantes;
- Devem ser analisadas o fim e o início de cada palavra.

---

[12] *C* e *V* significam consoante e vogal, respectivamente.

5.8.1 Definição de Letras e Fonemas

Como citado, para facilitar a representação das palavras a serem tratadas pelo ***Conversor Texto-Fala para a Língua Portuguesa***, é indicado a realização de algumas modificações fundamentais no *AFI* para simbolizar o seu universo de *fonemas*. O alfabeto proposto foi apresentado no *quadro 4.1*.

5.8.1.1 A Transcrição Letra/Fonema

Como foi dito, não há uma correspondência única entre *fonema* e *letra*. Assim, estudos merecem ser desenvolvidos para que a fala resultante apresente-se semelhante a do homem. O módulo de ***REGRAS 2*** da estrutura proposta (*figura 4.2*), define condições que devem ser consideradas na etapa de separação de unidades. Tais regras são tratadas a seguir.

- *Letra c*

A *letra* ***c*** assume dois *fonemas*, ou seja, /***k***/ e /***s***/. A determinação da forma correta pode ser representada a partir das seguintes considerações:

- Quando o ***c*** anteceder as vogais ***a***, ***o***, ***u*** ou quando anteceder uma *consoante*, assume o *fonema* /***k***/ - como nos vocábulos *casa*, *copo*, *culinária* e *tecla*, respectivamente;
- Quando o ***c*** anteceder as *vogais* ***e*** e ***i***, então tem-se o *fonema* /***s***/ - nessa regra têm-se as palavras *ceia* e *cinto*, por exemplo.

- *Letra g*

A *letra* ***g***, por sua vez, assume dois *fonemas* distintos: /***g***/ e /***j***/. Portanto, a sua transcrição ortográfico-fonética deve obedecer as seguintes regras:

- Se a *letra* ***g*** vier seguida das *letras* ***e*** ou ***i***, assume o *fonema* ***/j/*** - como nas palavras *geladeira* e *girafa*;
- Se a *letra* ***g*** vier seguida da *letra* ***u*** e das *letras* ***e*** ou i, as *letras* ***g*** e ***u*** se unem e dão origem ao *fonema* ***/g/*** - é o que ocorre nos vocábulos *negue* e *lânguido*, respectivamente;
- Para as demais ocorrências a *letra* ***g*** assume o *fonema* ***/g/*** - como exemplo têm-se as palavras *gato* e *gosto*.

- Letra l

A *letra* ***l*** assume dois *fonemas* - ***/l/*** e ***/u/***. Para determinar as ocorrências possíveis tem-se:

- Se a *letra* ***l*** vier seguida de *consoante* assume o *fonema* ***/u/*** - nessa regra têm-se os vocábulos *falta* e *alta*, por exemplo;
- Se a *letra* ***l*** vier seguida de uma *vogal*, então verifica-se o *fonema* ***/l/*** - como é o caso da palavra *lata*;
- Se a *letra* ***l*** coincidir com o final do vocábulo, tem-se o *fonema* ***/u/*** - por exemplo, tem-se a palavra *fatal*.

- Letra q

A *letra* ***q*** assume apenas o *fonema* ***/k/***. Entretanto, quanto vier seguida da *letra* ***u*** e das *letras* ***e*** ou ***i***, as *letras* ***q*** e ***u*** se unem e originam apenas o *fonema* ***/k/***. Como exemplo têm-se as palavras *qual*, *queijo* e *quinta*.

- Letra s

Com relação a *letra* ***s*** dois *fonemas* podem ser originados, ou seja:

- Se a *letra* ***s*** vier entre *vogais*, tem-se o *fonema* ***/z/*** - como na palavra *casa*, por exemplo;

- Caso contrário, a *letra* ***s*** assume o *fonema* ***/s/*** - nessa regra se enquadram os vocábulos *semáforo* e *solo*, por exemplo.

- Letra x

De acordo com a definição a *letra* ***x*** assume quatro *fonemas* distintos: ***/x/***, ***/z/***, ***/s/*** e ***/ks/***. Entretanto, determinar a forma adequada para cada situação em que aparece não é um trabalho direto. Logo, pode-se dizer que é impossível determinar de imediato, uma regra/tendência geral para as várias ocorrências existentes. Assim, a seguir tem-se a descrição de algumas situações. É verdade que o objetivo é conceber um sistema de vocabulário ilimitado. Mas, inicialmente, foi realizado um estudo considerando os vocábulos presentes no Minidicionário Aurélio [24], ou seja, foram observadas cerca de quatrocentas palavras.

Sequência AXA

Em geral, quando o ***x*** vem entre as vogais ***a***, tal *letra* assume o *fonema* ***/x/*** - como é o caso dos vocábulos *taxa*, *laxante* e *engraxate*. Entretanto, se a sequência vier antecedida pela *letra* ***s*** – como na palavra *saxão* -, o ***x*** assume o *fonema* ***/ks/***.

Sequência AXE

Em geral, quando o ***x*** vem entre as *letras* ***a*** e ***e***, assume o *fonema* ***/x/*** - ocorrência verificada no vocábulo *guaxe*, por exemplo.

Sequência AXI

Quando o ***x*** surge entre as vogais ***a*** e ***i***, verifica-se a ocorrência das quatro possíveis formas de *fonemas*. Esse fato, portanto, confirma a importância de se realizar um estudo detalhado para se determinar o *fonema* correto assumido pela unidade em certa situação. Mas, por outro lado, também evidencia a dificuldade de se estabelecer uma regra capaz de definir todas as ocorrências. Assim, a seguir tem-se a descrição dos casos detectados.

- Em geral, o ***x*** assume o *fonema* ***/x/*** - como em *abacaxi*, *maxixe* e *faxina*;

- Se a sequência vier antecedida ou seguida pela *letra* ***l***, então o ***x*** assume o *fonema* ***/ks/*** - ocorrência verificada nas palavras *galáxia* e *maxilar*;
- Se a sequência vier antecedida pela *letra* ***t***, o ***x*** assume o *fonema* ***/ks/*** - fato verificado nos vocábulos *táxi* e *taxidermia*;
- Na palavra ***máxima*** - e suas derivações - o ***x*** assume o *fonema* ***/s/***.

Sequência AXO

Para esta situação, o ***x*** assume o *fonema* ***/x/*** - como exemplo têm-se os vocábulos *coaxo* e *laxo*. Mas, quando a *letra* ***s*** antecede a sequência em questão – como é o caso da palavra *saxofone* -, o ***x*** assume o *fonema* ***/ks/***.

Sequência AXU

Neste caso, o ***x*** assume o *fonema* ***/x/***. Como exemplo tem-se o vocábulo *caxumba*.

Sequência EXA

Para este caso, várias ocorrências podem ser enumeradas, ou seja:

- Se a sequência coincidir com o início do vocábulo – *exame* e *exação*, por exemplo -, a *letra* ***x*** assume o *fonema* ***/z/***;
- Se a sequência vier antecedida pela *letra* ***h*** ou ***s*** ou pelas *letras* ***fl***, o ***x*** assume o *fonema* ***/ks/*** - casos verificados nas palavras *hexaedro*, *sexagenário* e *reflexão*, respectivamente;
- Caso contrário, o ***x*** assume o *fonema* ***/x/*** - como é o caso da palavra *vexame*.

Sequência EXE

Este caso também apresenta várias situações:

- Se a sequência coincidir com o início do vocábulo – como em *exemplo* e *execução* -, o ***x*** assume o *fonema* ***/z/***;

- Se a sequência vier antecedida pelas *letras* **s**, então o **x** assume o *fonema* **/ks/** - como exemplo tem-se o vocábulo *sexênio*;
- Caso contrário, o **x** assume o *fonema* **/x/** - como em *mexerico* e *remexer*.

Sequência EXI

Quando esta sequência ocorrer tem-se:

- Se a sequência coincidir com o início do vocábulo – como em *existir* e *existência* -, então o **x** assume o *fonema* **/z/**;
- Caso a sequência venha antecedida pela *letra* **l** ou pelas *letras* **fl** ou **compl**, o **x** assume o *fonema* **/ks/** - ocorrência verificada nos vocábulos *léxico*, *flexível* e *complexidade*;
- Caso contrário, o **x** assume corresponde ao *fonema* **/x/** - como nas palavras *mexicano* e *mexido*.

Sequência EXO

A ocorrência desta sequência implica nas seguintes situações:

- Se a sequência coincidir com o início do vocábulo – nessa regra têm-se as palavras *exótico* e *exorcismo* -, então o **x** assume o *fonema* **/z/**;
- Se a sequência vier antecedida das *letras* **n**, **s** ou **fl**, então o **x** assume o *fonema* **/ks/**. Como exemplo têm-se os vocábulos *nexo*, *sexo* e *reflexo*, respectivamente;
- Caso contrário, o **x** assume corresponde ao *fonema* **/x/** - como na palavra *pexotada*.

Sequência EXU

Quando o **x** obedecer a citada sequência tem-se:

- Se a sequência coincidir com o início do vocábulo e não for a monossílaba - que assume o *fonema* **/x/** -, então o **x** representa o *fonema* **/z/** - como verificado na palavra *exuberância*;

- Se a sequência vier antecedida pela *letra* **s** - como na palavra *sexual* -, então o **x** assume o *fonema* **/ks/**.

Sequência IXA

Quando o **x** obedecer a citada sequência tem-se:

- Se a sequência vier antecedida da *letra* **f**, então o **x** assume o *fonema* **/ks/** - como é o caso da palavra *fixa*;
- Caso contrário, o **x** assume corresponde ao *fonema* **/x/** - como nos vocábulos *lixa*, *ameixa* e *embaixada*.

Sequência IXE

Para esta sequência o *x* assume o *fonema* /x/ - nessa regra têm-se as palavras *baixela*, *peixe* e *rixento*.

Sequência IXI

Quando o **x** pertencer a citada sequência tem-se:

- Se a sequência vier antecedida da *letras* **f**, então o **x** assume o *fonema* **/ks/** - ocorrência verificada no vocábulo *asfixia*;
- Caso contrário, o **x** assume corresponde ao *fonema* **/x/** - como exemplo têm-se as palavras *baixio* e *caixilho*.

Sequência IXO

Neste caso, o **x** assume o seguintes *fonemas*:

- Se a sequência vier antecedida da *letra* **f** – como ocorre nas palavras *fixo* e *crucifixo* -, então o **x** assume o *fonema* **/ks/**;
- Caso contrário, o **x** assume corresponde ao *fonema* **/x/** - nessa regra enquadram-se as palavras *baixo*, *queixo* e *lixo*.

Sequência IXU

Para esta sequência o *x* assume o *fonema* /*x*/ - como exemplo tem-se a palavra *mixuruca.*

Sequência OXA

Nesta situação, o *x* é representado pelo *fonema* /*x*/ - como é o caso dos vocábulos *almoxarifado*, *coxa* e *oxalá.*

Sequência OXE

Neste caso, o ***x*** assume o seguintes *fonemas*:

- Se a sequência coincidir com o início do vocábulo – como na palavra *oxente* -, então o ***x*** assume o *fonema* ***/x/***;
- Caso contrário, a *letra* ***x*** assume o *fonema* ***/ks/*** - nessa regra pode-se citar a palavra *boxe.*

Sequência OXI

Quando esta sequência é verificada o *x* assume o *fonema* /*ks*/- como nos vocábulos *tóxico* e *oxigênio.* Entretanto, verifica-se a exceção para a palavra *próximo*, onde o ***x*** assume o fonema /***s***/.

Sequência OXO

Diante desta ocorrência – como na palavra *muxoxo* -, o *x* é representado pelo *fonema* /*x*/.

Sequência OXU

Nesta situação, o *x* é representado pelo *fonema* /*x*/- como exemplo tem-se o vocábulo *oxum.*

Sequência UXA

Neste caso, o *x* é representado pelo *fonema* /x/- nessa regra têm-se os vocábulos *luxação* e *trouxa.*

Sequência UXE

Diante desta ocorrência o *x* é representado pelo *fonema* /x/- como exemplo tem-se o vocábulo *luxento.*

Sequência UXI

Quando esta sequência é verificada – como nas palavras *tauxia* e *muxirão* -, o *x* assume o *fonema* /x/.

Sequência UXO

Nesta situação, o ***x*** assume o seguintes *fonemas*:

- Se a sequência vier antecedida das *letras* ***fl*** ou então vier no início do vocábulo, então o ***x*** assume o *fonema* ***/ks/*** - como nos vocábulos *refluxo* e *uxoricida*, respectivamente;
- Caso contrário, o ***x*** assume corresponde ao *fonema* ***/x/*** - nessa regra têm-se as palavras *luxo* e *repuxo.*

Sequência UXU

Quando o *x* obedece a citada sequência – como no vocábulo *luxúria* -, este assume o *fonema* /x/.

Ainda é conveniente e necessário realizar certas considerações relacionadas a *letra x*:

- O *x* assume o *fonema* /***ks***/ quando vier no final da palavra – como em *xerox*;
- O *x* assume o *fonema* /***s***/ quando vier entre uma vogal e uma consoante – nessa regra tem-

se a palavra *explicação*;

- O *x* assume o *fonema* /***x***/ quando vier no início da palavra – como ocorre na palavra *xerox*;
- Para todas as regras citadas acima devem ser consideradas todas as sequências que representam *prefixos*, ou seja, as sequências ***in***, ***des***, dentre outras.

CAPÍTULO 6

A IMPLEMENTAÇÃO DOS MÓDULOS E OS RESULTADOS OBTIDOS

Este capítulo mostra a implementação de alguns módulos do ***Sistema LEIA*** *, bem como os resultados obtidos.*

6.1 Considerações Gerais

A implementação do ***Sistema LEIA*** se deu a partir dos algoritmos descritos no capítulo anterior. Eles foram desenvolvidos utilizando a linguagem de programação Delphi, já

que esta pode ser vista como uma ferramenta bastante flexível, permitindo a concepção de um ambiente mais amigável. Assim, todo o funcionamento do ***Sistema LEIA*** será discutido a seguir.

6.2 Ambiente de Trabalho

A apresentação geral do ***Sistema LEIA*** pode ser visualizada na *figura 5.1*. A tela principal de acesso - ilustrada na *figura 6.1* -, exibe todas as suas partes integrantes. É verdade que nem todos os módulos estão habilitados. Entretanto, o estudo desenvolvido definiu toda a estrutura e particularidades necessárias para o funcionamento completo do *Sistema*. Nas seções seguintes ter-se-á a verificação das etapas implementadas.

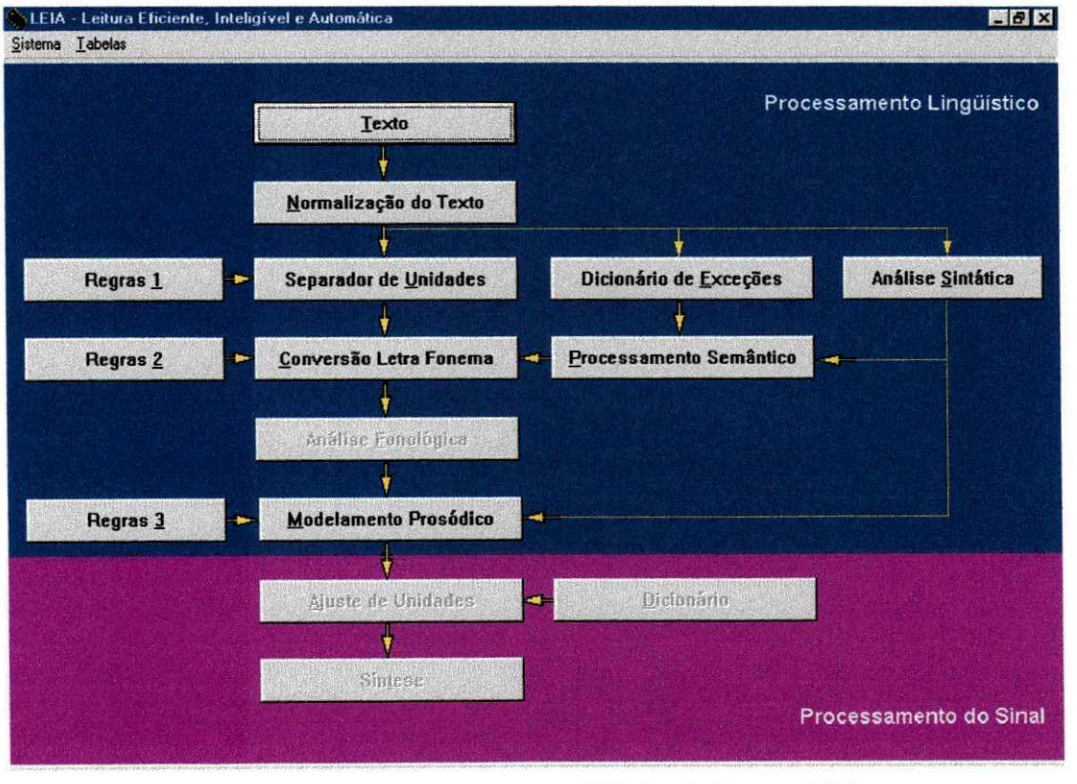

Figura 6.1: Tela de Acesso aos Módulos do Sistema LEIA

6.3 Sistema LEIA: Visualização e Verificação

O grande objetivo do *Sistema* consiste em obter, de maneira automática, a conversão texto-fala. Assim, não é de interesse verificar partes isoladas. O desejado é originar a fala tão logo a introdução do texto tenha sido realizada. Mas, como tem-se a proposição de uma nova estrutura, é interessante realizar um acompanhamento detalhado dos vários módulos. Por outro lado, este posicionamento de análise permitirá o tratamento de certas características particulares e características da Língua Portuguesa.

A introdução do texto ao *Sistema* é realizada acessando o módulo *texto* da estrutura proposta (*figura 6.1*). Assim, uma nova tela é originada - *figura 6.2*. Em seguida, para proceder ao processo de transcrição, os módulos subsequentes devem ser acionados para, assim, poder realizar um acompanhamento passo-a-passo das transformações do texto. Dessa forma tem-se uma verificação mais precisa do desempenho do Sistema projetado para a Língua Portuguesa.

Para verificar o funcionamento do ***Sistema LEIA*** considere o texto a seguir. Foi escolhido um texto que apresentasse uma série de anomalias a serem tratadas. No caso, pode-se ver a existência de *siglas*, *algarismos*, *abreviações* e *palavras compostas*.

" O *Sistema LEIA* está sendo desenvolvido no LAPS. É um projeto que teve início em 15/03/97, através dos alunos de pós-graduação Fabrícia Figueiredo e Leonel Costa. Na parte de orientação figuram os seguintes professores Dr. Benedito Aguiar e Dra. Lírida Barros ".

A *figura 6.3* exibe a tela mostrando o texto considerado como exemplo. Nela pode-se visualizar que, de imediato, tem-se uma junção dos módulos de *normalização*, *separação de unidades* e *conversão letra/fonema*. Estes, por sua vez, já incorporam outros módulos que serão detalhados a seguir.

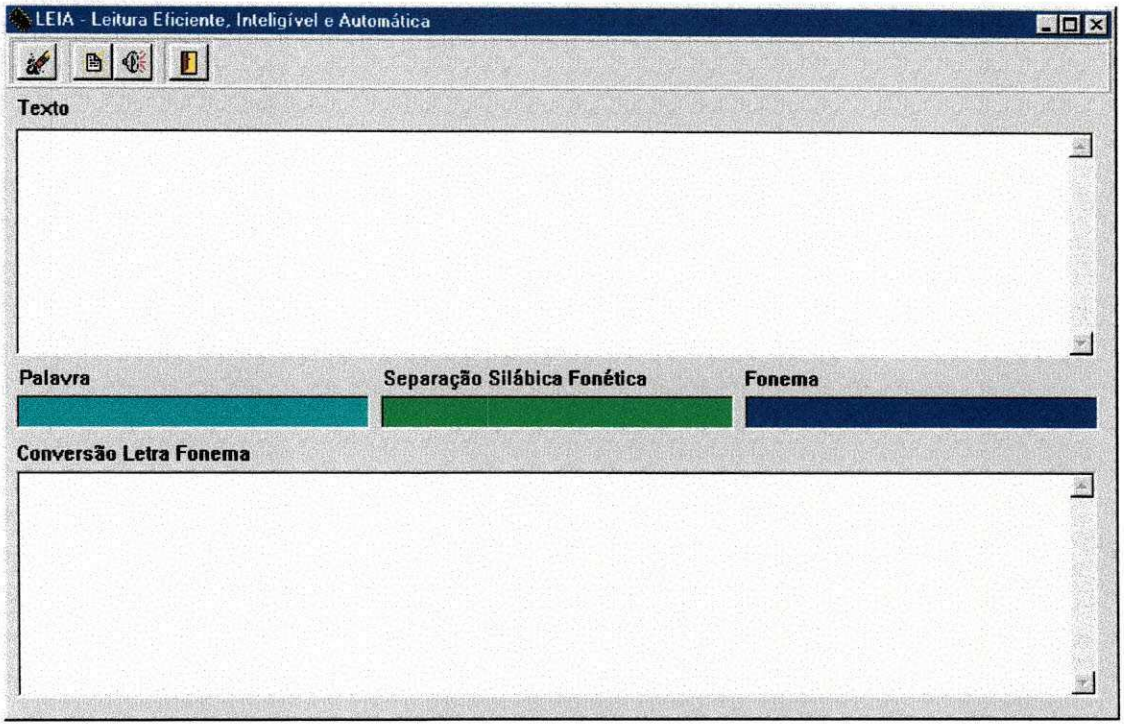

Figura 6.2: Tela para a Introdução do Texto a ser Convertido

6.3.1 Descrição dos Módulos do Sistema LEIA

Para possibilitar a transformação do texto a ser convertido, em um formato adequado a originação da fala, foram criadas várias entradas de dados – que passaram a ser chamadas de *tabelas* – para alimentar os módulos integrantes do *Sistema.* Esse procedimento foi escolhido devido a dois fatores principais. Primeiramente, sabe-se que este estudo marca o início do desenvolvimento de um ***SCTF*** direcionado ao idioma português, considerando um vocabulário ilimitado. Assim, de início pode-se afirmar que será impossível esgotar todas as ocorrências da língua em análise. Por outro lado, tem-se a preocupação de tornar o *Sistema* mais flexível e adaptável às terminologias e nomenclaturas, relacionadas ao idioma português, que possam surgir. Dessa forma, ter-se-á sempre uma fácil atualização de dados.

A *figura 6.4* exibe a *tabela - Índice de Normalização* - que gera dados para a tradução de

siglas e *abreviações*. Logo, tem-se um campo destinado a identificar o termo que será expandido, o qual encontra-se associado a informação *unidade de medida*. Diante de um dado que não represente uma *unidade de medida* apenas o campo *expansão* será utilizado. Um exemplo desta situação pode ser a abreviação *ABR* que será traduzida para *abril*. Caso contrário, os demais campos da tela serão utilizados. Para facilitar o entendimento do procedimento em descrição considere a abreviação *cm*. De acordo com o pensamento adotado, tem-se a composição do arquivo da seguinte forma: centímetro(s) e milímetro(s) para os campos parte inteira (singular e plural) e fracionária (singular e plural), respectivamente.

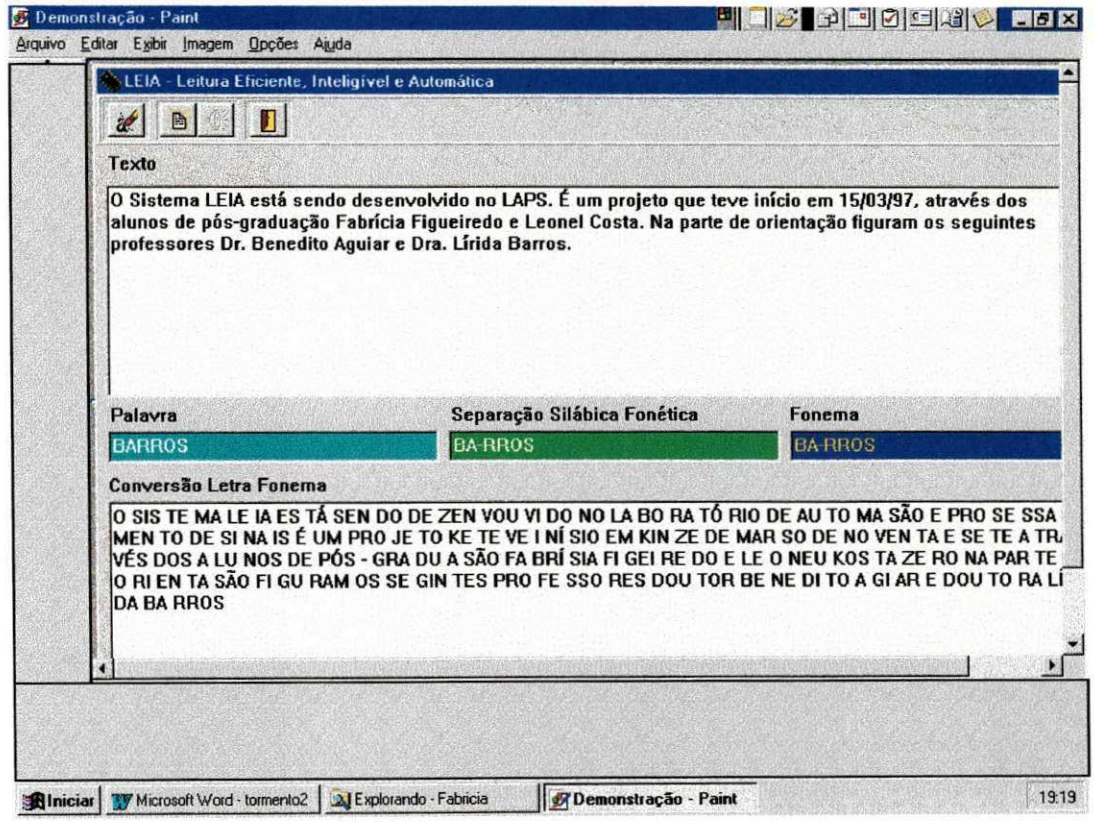

Figura 6.3: Sistema LEIA – Visualização da Formatação de um Texto

Dando continuidade às eventualidades que podem surgir diante do processo de formatação do texto tem-se as palavras homógrafas heterofônicas. A *figura 6.5* ilustra a *tabela – Dicionário de Exceções –* que processa as informações para a identificação correta das *unidades* representativas do vocábulo. Assim, tem-se um campo que apresentará a palavra a ser analisada, ao qual será atribuída uma classificação gramatical, e outros que conterão a

classificação dos vocábulos vizinhos – *anticontexto* e *póscontexto*. Após estas informações e o posterior cruzamento delas, têm-se os *fonemas* e a *separação das unidades*.

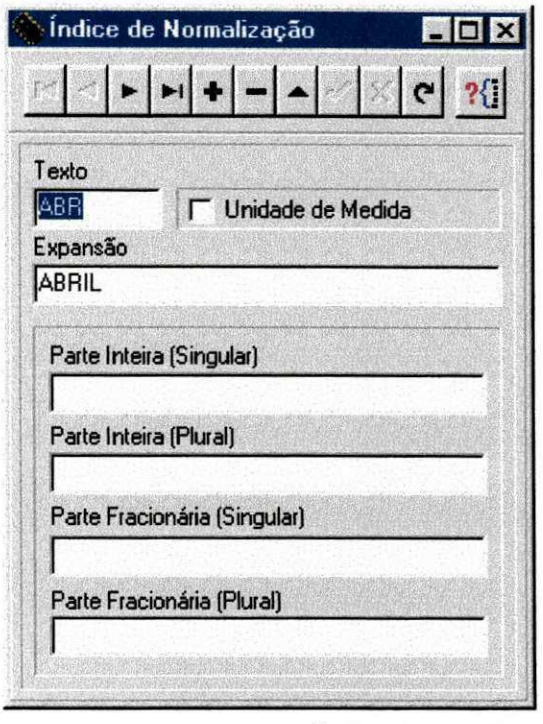

Figura 6.4: Sistema LEIA – Índice de Normalização

Relacionado à associação correta da pronúncia das *palavras homógrafas heterofônicas*, é conveniente ressaltar alguns pontos que ainda precisam ser bastante estudados. Na realidade, considerar apenas as informações nas vizinhanças do vocábulo pode ser insuficiente para se chegar ao pretendido. Este procedimento enquadra os casos mais gerais. Então, foi cogitada a possibilidade de considerar um número maior de palavras antecedentes e precedentes. Entretanto, surgiu um ponto de reflexão: em ocorrências isoladas, ou seja, quando o vocábulo representa a sentença, como proceder? Este ponto fez com que, de início, a *tabela* permanecesse sem alterações. Mas estudos direcionados para a identificação automática de todas as ocorrências estão em andamento, até porque esta particularidade é considerada como um grande desafio do ***Sistema de Síntese Texto-Fala***.

Em seguida tem-se uma das etapas que mais contribui para a geração de informações para todo o *Sistema*, isto é, ***Análise Sintática***. A *figura 6.6* ilustra a tela representativa do citado módulo. Na realidade ela servirá para dar suporte a identificação de certas classes gramaticais, inviáveis de serem definidas na íntegra devido ao seu grande número de elementos integrantes. Comparando os artigos e os verbos, percebe-se claramente que a definição dos primeiros é possível, enquanto que a dos segundos não é aconselhável. Assim, tem-se a formação de um arquivo contendo as classes que apresentam um pequeno número de elementos.

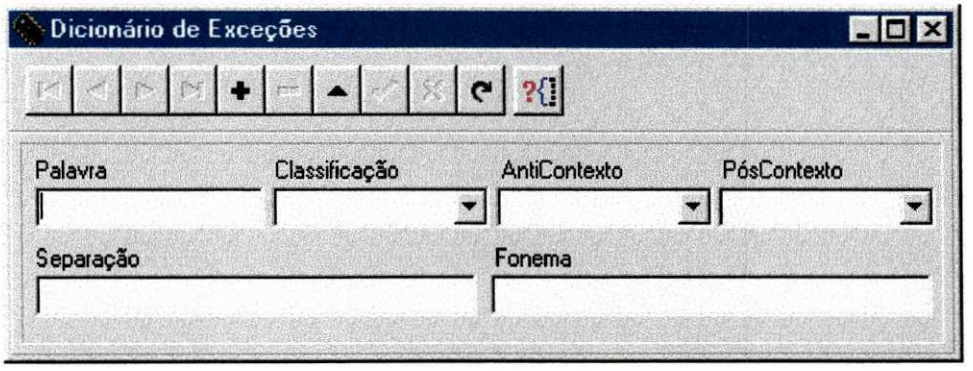

Figura 6.5: Sistema LEIA – Dicionário de Exceções

A concepção dos demais módulos – exceto *tratamento prosódico* – não necessitou da geração de tabelas individuais. As regras formuladas, descritas em Seções anteriores, foram incorporadas nos algoritmos desenvolvidos de modo a proceder a *separação de unidades* e a *conversão letra/fonema.*

6.3.2 A Concepção do Dicionário

Com relação à determinação das *unidades* que devem figurar no dicionário (arquivo de dados) algumas considerações devem ser realizadas. É preciso identificar todas as unidades que devem ser gravadas e armazenadas e associá-las aos valores de *frequência fundamental* e

duração. Entretanto, sabe-se que tais parâmetros sofrerão alterações em virtude do posicionamento das *unidades* na palavra: nas palavras *cabana* e *banana* os valores do *difone* /*ba*/ irão variar. Na primeira situação, ele coincide com a sílaba tônica, o que representa uma maior proeminência. Já na segunda situação, a citada *unidade* não é tão enfatizada. O vocábulo *banana* ainda sugere um outro caso: a *unidade na* apresenta tonicidades distintas. Tem-se primeiramente um som nasalizado e forte seguido por um som aberto.

Figura 6.6: Sistema LEIA – Análise Sintática

Dessa forma, como determinar os valores dos parâmetros que devem ser associados às *unidades*? Realizar uma análise individual para cada ocorrência pode ser considerado um processo demorado e inesgotável, pois estima-se em cerca de 4000 unidades o tamanho do dicionário. Assim, uma medida que pode ser seguida consiste em determinar *unidades padrão* que servirão para projetar os valores de outras *unidades* que apresentem características semelhantes.

A determinação das *unidades* deve obedecer, portanto a *Nomenclatura Gramatical Brasileira* [12] que estabelece uma classificação para as vogais e consoantes, com relação à zona de articulação, intensidade, timbre e papel das cavidades bucal e nasal. Assim, tem-se, por exemplo, que as consoantes *p*, *b* e *m* apresentam semelhanças quanto a zona de articulação, sendo consideradas como *bilabiais* (lábio contra lábio). Por outro lado, quanto ao papel das cordas vocais as consoantes *b* e *m* são classificadas como sendo sonoras, enquanto que o *p* é dito explosivo.

Inicialmente, convém estabelecer as semelhanças existentes entre as letras do alfabeto. Em seguida, deve-se estabelecer os valores para as situações possíveis, ou seja, considerando a

unidade ba verificar e determinar os valores quando esta representa uma sílaba tônica ou uma sílaba átona. Entretanto, para identificar os valores é importante considerar vários vocábulos nos quais a unidade ocorre com a mesma característica. Considerando as palavras *ba*cia, *ba*tata, *ba*nana e mala*ba*rista, tem-se que o *difone ba* em todas as situações representa uma sílaba átona. Mas, pode-se também verificar que há diferenças entre as pronúncias, ou seja, em *batata* tem-se um som aberto, enquanto em *banana* tem-se um som nasalizado. Já comparando as palavras *bacia* e *batata* pode-se perceber uma semelhança. Assim, convém identificar estas ocorrências e analisar a variação da *unidade acústica.* Com base neste estudo, portanto, deve-se fazer uma média de valor.

Seguindo este pensamento, o próximo caso consiste na estimação dos valores para as *unidades* pertencentes a mesma classe. Por exemplo, deve-se analisar até que ponto os valores da *unidade ma* coincidem com os da *unidade ba* – já que *b* e *m* apresentam semelhanças quanto a articulação e vibração das cordas vocais.

Todo este procedimento sugerido foi concebido no sentido de agilizar a formação do arquivo de dados. É esperado que sejam realizadas alterações de valores a medida que o Sistema esteja sendo testado. A concepção de um conversor para um idioma com tantas particularidades, capaz de ser aplicado para texto com vocabulário ilimitado, apresentando uma fala natural e inteligível, requer simulações, que por sua vez implicará em mudanças e assim por diante.

CAPÍTULO 7

CONCLUSÕES E PERSPECTIVAS

Este capítulo aborda as conclusões e perspectivas de trabalho referentes ao Sistema de Conversão Texto-Fala direcionado para a Língua Portuguesa

A realização deste estudo na área de *Processamento Digital de Voz*, enfocando a parte de ***Síntese Texto-Fala***, pode ser vista como um desafio. Até então, neste laboratório (*LAPS*), ainda não tinha sido desenvolvido nenhum trabalho nesta direção. Por outro lado, os estudos existentes ainda deixam a desejar, ou seja, a naturalidade e a inteligibilidade da fala são

características que ainda devem ser bastante exploradas. Para marcar todo o caminho seguido, é importante citar os problemas enfrentados e as soluções encontradas. É verdade que, de certa forma, tudo isso já foi tratado no texto. Entretanto, é importante destacar de forma individual até para estabelecer as possibilidades/necessidades de pesquisas.

Inicialmente, pode-se dizer que a ***Síntese Texto-Fala*** é um campo de pesquisa que ainda deve ser bastante explorado. Analisando o material existente, tem-se um número considerável de publicações. Entretanto, elas, em quase sua totalidade, apresentam-se direcionadas para outros idiomas. Não se pode negar a importância dos artigos voltados para a Língua Portuguesa. Mas, para se conseguir representar as características fundamentais à obtenção de uma fala natural e inteligível, se fazem necessários estudos mais aprofundados.

Seguindo esta linha de pensamento e visando contribuir para a referida área, foi desenvolvida uma estrutura de ***Conversão Texto-Fala*** denominada ***LEIA***[13]. Na realidade, foi criado um ambiente de trabalho para o *Sistema*, incorporando todos os seus módulos integrantes. Essa preocupação em se trabalhar de forma modular é justamente para possibilitar um acompanhamento mais preciso do seu funcionamento, bem como permitir uma exploração mais minuciosa de cada um. Essa afirmações evidenciam a importância deste estudo na área de ***Síntese de Voz***.

O ***Sistema LEIA*** foi constituído considerando os traços característicos básicos encontrados na bibliografia, ou seja, processamento linguístico e processamento do sinal. Entretanto, ao processamento linguístico foi dada uma maior ênfase. Desenvolveu-se uma abordagem mais detalhada baseada em características do português. Esse fato, portanto, impôs certas dificuldades que impossibilitaram a implementação de todos os módulos. À medida que a estrutura ia sendo formada, uma série de questionamentos surgiam e a solução para os problemas levantados era marcada por particularidades que, por sua vez, originavam outras e assim por diante.

As inconveniências começaram logo no primeiro módulo. Uma primeira, consistiu em saber se a *siglas* seriam pronunciadas de forma extensa ou não. Outra dizia respeito ao

---

[13] Sistema de Leitura Eficiente Inteligível e Automática.

estabelecimento das unidades integrantes. Assim, devido ao grande número de *siglas* existentes foi sugerida a concepção de vários arquivos, sendo cada um direcionado para certos assuntos (enfoques). O problema, então, recaiu em como direcionar o texto ao dicionário adequado.

Na etapa responsável pela separação de unidades surgiu o problema de como identificar o local onde as quebras seriam inseridas, pois se fosse levada em conta apenas a separação de sílabas o ***Sistema LEIA*** provavelmente perderia certas características relacionadas à naturalidade e inteligibilidade. Dessa forma, a maneira de como os vocábulos são realmente proferidos foi considerada. Entretanto, um outro fator também deve ser evidenciado: um texto é formado por um conjunto de palavras e quando a leitura é realizada tem-se um encadeamento delas. Isso exige a realização de um estudo envolvendo os vocábulos e suas vizinhanças. Nesse direcionamento, começa-se a identificar a presença dos aspectos ligados a tonicidade.

Relacionada às *pausas mentais* um questionamento pode ser levantado: será que realmente é necessário identificar tais paradas? Essa indagação surgiu quando se começou a investigar os fatores que deveriam ser considerados para determinar os valores dos parâmetros - *frequência fundamental* e *duração* - das unidades. A partir do momento que se analisa os vocábulos em nível de palavra e de sentença, tem-se um estudo dos vocábulos e de suas vizinhanças. Assim, foi visto que, em certas situações, duas palavras eram proferidas sem interrupções. Isto mostra que, indiretamente, há introdução de paradas entre as palavras.

O módulo responsável pela *análise sintática* pode ser considerado bastante complexo. A identificação de certas classes gramaticais é difícil de ser estabelecida. O português é um idioma que não apresenta uma generalização quanto à formação/disposição dos vocábulos. E, como visto, este rotulamento gera informações essenciais para diversas etapas - *modelamento prosódico* e *semântico* -, pois a partir delas consegue-se dotar as palavras de aspectos contendo os valores dos parâmetros *frequência fundamental* e *duração.*

Um outro ponto que merece destaque está relacionado a determinação das "*máscaras*". O estabelecimento das sequências e das condições de ocorrência mereceu a realização de testes e adaptações, já que o objetivo era atingir o maior número possível de vocábulos do idioma - para não dizer todos. Assim, em muitas situações a palavra não era separada de

maneira correta. Isso exigia a incorporação de mais uma condição. Entretanto, às vezes, corrigir a separação de um vocábulo inibia outras regras anteriormente definidas. Após a realização de uma série de análises chegou-se a conclusão de que todo o processo de verificação deveria ser feito através de comparações decrescentes das "*máscaras*".

Todas essas afirmações realizadas até o momento só reafirmam o quanto é difícil representar todas as características e particularidades do idioma português. A consideração de um procedimento para tratar algo sempre gera uma série de caminhos a serem investigados. E não adianta contar com um processamento considerável do sinal se não tiver uma representação adequada para as unidades acústicas que originarão a fala.

A determinação das unidades a serem buscadas no dicionário foi um ponto um tanto conflitante. Como citado, o critério utilizado para proceder a identificação delas baseou-se em dois fatores: separação das sílabas e modo como os vocábulos são pronunciados. Logo, levando em consideração a realização física dos *fonemas*, pode-se perceber que por mais genérico que seja o Sistema, ele sempre apresentará alguns traços particulares. Mas isso pode ser encarado como um aspecto positivo, no sentido de torná-lo mais flexível. A partir de ajustes realizados na etapa denominada de *análise fonológica*, a fala originada poderá ter consonância com costumes de ordem regional, por exemplo.

Assim, reunindo todos esses pontos, citados anteriormente, foi conseguida a implementação de um ***Sistema de Conversão Texto-Fala*** contendo as partes básicas para transformar o texto em um formato possível de ser tratado pela etapa responsável pelo processamento do sinal. A estrutura obtida realiza a *normalização do texto* - a partir da qual tem-se a expansão das *siglas*, *algarismos* e *abreviações* -, bem como determina os *fones* que serão buscados no dicionário de unidades acústicas. Entretanto, a esse processo de identificação fonética, foram incorporados tratamentos objetivando a caracterização do idioma português: várias regras de tonicidade e ocorrência de sons foram associadas ao módulo.

Pode-se questionar o fato de que nem todas as etapas do Sistemas estão funcionando. Essa afirmação é verdadeira, entretanto, o presente trabalho abordou a estrutura de ***Conversão Texto-Fala*** de forma geral e, dessa forma, destacou os pontos básicos que todos os módulos devem conter. Toda a concepção da estrutura foi feita objetivando produzir um

Sistema que apresentasse uma interface amigável, fácil de ser utilizada e também que permitisse a incorporação de outros módulos para torná-la mais robusta.

Assim, a idéia básica do trabalho era obter, de fato, um ***Conversor Texto-Fala*** completo para a língua portuguesa. Porém, com o decorrer da pesquisa verificou-se que o desenvolvimento do processamento lingüístico e o de sinais são distintos e demandam um tempo bem maior, que o considerado inicialmente, sobretudo levando-se em consideração as peculiaridades da língua portuguesa. Desta forma sugere-se, para trabalhos futuros, o desenvolvimento de um estudo mais aprofundado envolvendo o *processamento prosódico*, o desenvolvimento da etapa responsável pelo processamento de sinais (sintetizador), juntamente com o dicionário de unidades acústicas – considerando pontos destacados neste trabalho – e a interface entre os diversos estágios, de modo a resultar em uma ***Síntese Texto-Fala*** de alta qualidade.

Referências

[1] Egashira, F., "*Síntese de Voz a partir de Texto para a Língua Portuguesa*", Universidade Estadual de Campinas - Faculdade de Engenharia Elétrica - Dissertação de mestrado, julho de 1992.

[2] Flanagan J. L., "*Computers that Talk and Listen: Man-machine Communication by Voice*", Proceedings of the IEEE, Vol. 64, N. 4, pp 405-415, 1976.

[3] Allen J., "*A Perspective on Man-Machine Communication by Speech*", Proceedings of the IEEE, Vol. 73, N. 11, pp 1541-1550, 1985.

[4] Hirschberg J. B., Riederer S.A., Rowley J. E., Sydral A. K., "*Voice Response Systems:Technologies and Applications*", AT&T Techinal Journal, pp 42-51, 1990.

[5] Sproat R. W., Olive J. P., "*Text-to-Speech Synthesis*", AT&T Techinal Journal, pp 35-44, 1995.

[6] Silva C. H. da, Violaro F., "*Modelamento Prosódico para Conversão Texto-Fala do Português falado no Brasil*", XIII Simpósio Brasileiro de Telecomunicações - Águas de Lindóia, pp 77-82, setembro de 1995.

[7] Petric I., Bergh H. van den, "*Text Features Affecting Listening Performance in One Way Speech Communication*", Analysis and Synthesis of Speech - Strategic Research Towards / High-Quality Text-to-Speech Generation, Editors: Vicent J. van Heuven and Louis C. W. Pols, Mounton de Gruyter, Berlin - New York, 1993, pp 13-26.

[8] Prado P. P. L. do, "*Sintetizador Articulatório de Voz: Mapeamento Acústico/Articulatório*", XI Simpósio Brasileiro de Telecomunicações - Natal, pp 708-712, setembro de 1993.

[9] Dutoit T., "*High Quality Text-to-Speech Synthesis of the French Language*", Faculté Polytechnique de Mons - Ph. D. Dissertation, october 1993.

[10] Dicionário Brasileiro da Língua Portuguesa, Enciclopédia Mirador Internacional, Volume 1 (A – INDEBI), Terceira Edição, São Paulo, 1979.

[11] Nicola J. de, Infante U., "*Gramática Contemporânea da Língua Portuguesa*", Editora Scipione, São Paulo, 1989.

[12] Bechara E., "*Moderna Gramática Portuguesa*", Companhia Editora Nacional, 25ª. Edição, São Paulo, 1980.

[13] Violaro F., "*Processamento Digital de Sinais de Fala*", Minicurso, XV Simpósio Brasileiro de Telecomunicações, Recife, 8-11 de Setembro de 1997.

[14] Aguiar Neto, B. G., "*Processamento Digital de Sinais de Voz e Imagem*", DEE/CCT/UFPB, Agosto de 1996. [14] Dutoit T., Leich H., "*MBR-PSOLA: Text-to-Speech Synthesis Based on na MBE Re-Synthesis of the Segments Database*", Faculté Polytechnique de Mons - TCTS-Multitel.

[15] D. Junior, J. R., Proakis J. G. and Hansen J. H. L., "*Discrete-Time Processing of Speech Signals*", Macmillan Publishing Company, 1993.

[16] Rabiner L. R., Schafer R. W., "*Digital Processing of Speech Signal*", Prentice Hall, USA, 1978.

[17] Silva C. H., "*Modelamento Prosódico para Conversão Texto-Fala do Português falado no Brasil*", Universidade Estadual de Campinas - Faculdade de Engenharia Elétrica - Dissertação de mestrado, dezembro de 1995.

[18] Dutoit, T. and Leich, H., "*A Comparison of Four Candidate Algorithms in the context of High Quality Text-to-Speech Synthesis*", ICASSP'94, Adelaide, Austrália.

[19] d'Alessandro C., Liénard J., "*Spoken Output Technologies*", LIMSI-CNRS, Orsay, France.

[20] Egashira F., Violaro F., "*Conversor Texto-Fala para a Língua Portuguesa*", XIII Simpósio Brasileiro de Telecomunicações - Águas de Lindóia, pp 71-76, setembro de 1995.

[21] Egashira F., Violaro F., "*Conversor Texto-Fala*", I Fórum Nacional de Ciência e Tecnologia em Saúde, pp 148-151, 1992.

[22] Egashira F., Violaro F., "*Síntese de Voz a partir de Texto para a Língua Portuguesa*", IX Simpósio Brasileiro de Telecomunicações - São Paulo, Setembro de 91.

[23] Figueiredo F. A., Barros L. A., Aguiar Neto B. G., "*Uma Nova Abordagem para o Sistema de Conversão Texto-Fala para a Língua Portuguesa*", XV Simpósio Brasileiro de Telecomunicações - Recife, Setembro de 97.

[24] Ferreira, A. B. de H., "*Minidicionário Aurélio*" (Minidicionário da Língua Portuguesa), 1ª Edição, 5ª Impressão, Editora Nova Fronteira, Rio de Janeiro, 1997.