Universidade Federal do Rio de Janeiro

Escola Politécnica

Departamento de Eletrônica e de Computação

Sistema de Conversão Texto-Fala com Busca Otimizada de Unidades Acústicas em Banco de Voz

Autora:

Kersey Wirleide Anacleto Xavier da Silva

Orientador:

Prof. Sergio Lima Netto, Ph.D.

Co-orientador:

Vagner Luis Latsch, D.Sc.

Examinador:

Prof. José Gabriel Rodriguez Carneiro Gomes, Ph.D.

Examinador:

Prof. Amaro Azevedo de Lima, Ph.D.

DEL

Dezembro de 2011

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
Escola Politécnica - Departamento de Eletrônica e de Computação
Centro de Tecnologia, bloco H, sala H-217, Cidade Universitária
Rio de Janeiro - RJ CEP 21949-900

"Recife

Não a Veneza americana

Não a Mauritsstad dos armadores das Índias Ocidentais

Não o Recife dos Mascates

Nem mesmo o Recife que aprendi a amar depois"

(trecho do poema "Evocação do Recife" do poeta pernambucano Manuel Bandeira)

AGRADECIMENTO

Inicialmente, agradeço a Deus, por sempre estar ao meu lado em todos os momentos, atendendo às minhas orações e me dando coragem para superar os momentos difíceis.

Aos meus pais, pela educação que me concederam.

A David, pelo apoio e incentivo incondicionais nas horas difícieis.

Aos meus orientadores: professor Sergio Lima Netto, por ter me aceito como aluna de iniciação científica e pela confiança que vem depositando em mim desde então, o que resultou neste trabalho; e Vagner Luis Latsch, pela confiança em me permitir conceder contribuições ao seu projeto e pela paciência em responder todos os meus e-mails, quase que diários, de dúvidas.

Por fim, a todos que de alguma maneira contribuíram para a conclusão deste projeto.

Kersey Wirleide Anacleto Xavier da Silva

RESUMO

Um conversor texto-fala é um sistema que produz sinais de fala sintetizados a partir da entrada de um arquivo de texto escrito em um língua pré-definida. A fala sintetizada já vem sendo utilizada em diversas aplicações, como nos navegadores por GPS, nos sistemas de acessibilidade para deficientes visuais e no auxílio do ensino de idiomas. Este trabalho apresenta o estudo e implementação de métodos de seleção de unidades de síntese de um sistema conversor texto-fala (sistema SASPRO) desenvolvido por Vagner Latsch em sua tese de doutorado.

O trabalho se inicia com uma descrição das principais etapas dos conversores texto-fala, que basicamente, podem ser dividas em duas: processamento do texto e síntese da fala. Em seguida, são vistas as principais características do sistema SASPRO. Também é visto como uma busca elaborada de unidades pode melhorar a naturalidade do sinal de fala gerado. São mostradas, então, as diferentes implementações da seleção de unidades acústicas, indicando a melhoria obtida em cada versão. Por fim, os resultados obtidos são apresentados e analisados de forma subjetiva.

Palavras-Chave: síntese de voz, conversor texto-fala, seleção de unidades, parâmetros prosódicos.

ABSTRACT

A text-to-speech (TTS) system generates a synthesized speech signal starting from a text file input by the user. TTS applications include, for instance, GPS navigators, human-machine interface for blind people, and second-language teaching systems. This project presents the analysis and implementation of several search methods of speech units in the SASPRO system TTS, developed by Vagner Latsch in his D.Sc. thesis.

This work starts by describing the main stages in a TTS system, which are basically two: text processing and speech synthesis. We then describe the main capabilities of the SASPRO system. Following, we consider how an elaborate unit search can improve upon the natural perception of the generated speech. Distinct search procedures are described, indicating the improvement achieved by each method. In the end, practical results are presented along with the corresponding subjective evaluation for each method.

Key-words: speech synthesis, text-to-speech, unit selection, prosodic parameters.

SIGLAS

G2P - Grapheme To Phoneme/Phone

HMMs - hidden Markov models

LINSE - Laboratório de Circuitos e Processamento de Sinais

LP - Linear Prediction

LSF - Line spectral frequencies

MBROLA - Multi Band Resynthesis OverLap Add

NCE - Núcleo de Computação Eletrônica

POS - Parts of Speech

PSOLA - Pitch Synchronous Overlap and Add

SASPRO - Sistema de Análise e Síntese da Prosódia

SERPRO - Serviço Federal de Processamento de Dados

STL - Standard Template Library

TD-PSOLA - Time Domain - Pitch Synchronous Overlap and Add

TTS - Text-to-Speech

UFRJ - Universidade Federal do Rio de Janeiro

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Um conversor texto-fala ou TTS (do inglês *text-to-speech*) é um sistema que se propõe a produzir sinais de fala sintetizados a partir da entrada de um arquivo de texto qualquer escrito em uma língua pré-definida. Para esse tipo de sistema funcionar de maneira "ideal", ele deve ser capaz de realizar, de maneira automática, um processo similar ao que ocorre durante a leitura oral humana. Trata-se, entretanto, de uma tarefa bastante complexa, pois a leitura não se limita apenas à conversão de cada palavra na sua representação fonológica, mas envolve toda a competência do leitor em demonstrar determinadas posturas acerca do que está escrito, como por exemplo, raiva, questionamento, alegria etc.

O desenvolvimento de um sistema TTS, assim, pode ser visto como um assunto multidisciplinar, que exige participação de várias áreas de conhecimento, como linguística e processamento de sinais, o que torna este assunto desafiador e ao mesmo tempo motivador.

1.1 Motivação

A evolução dos computadores trouxe a massificação do uso dos mesmos fazendo com que interação usuário-máquina se tornasse cada vez mais importante. Como a voz se trata de um meio natural de comunicação, pode-se considerá-la como um meio de interface mais direto e efetivo. De fato, já se percebe uma tendência dos sistemas de utilizarem a fala como forma de interação homem-máquina, e esta

é a importância de obtermos sistemas que produzam sinal de fala com alta inteligibilidade e naturalidade.

Os navegadores por GPS, os sistemas de atendimento eletrônico, os sistemas de acessibilidade para deficientes visuais, e o auxílio no estudo de idiomas são exemplos de aplicações dos sistemas que utilizam a fala como uma maneira de interagir com o usuário. Apesar das várias aplicações existentes para os sistemas TTS, esta é ainda uma tecnologia em desenvolvimento, principalmente para o português brasileiro.

Assim, o objeto de estudo deste trabalho é o conversor TTS para português brasileiro do sistema SASPRO [3] que, após uma etapa de processamento de texto, sintetiza o sinal de fala através da concatenação de segmentos temporais previamente gravados e armazenados em um banco de unidades. Este sistema utiliza ainda o algoritmo TD-PSOLA (*time domain - pitch synchronous overlap-and-add*) para manipular as variáveis prosódicas do sinal concatenado.

A síntese por meio da simples concatenação de unidades, tal como é feita no TTS do sistema SASPRO, leva, em geral, a sinais com descontinuidades e a contornos prosódicos diferentes [1] do desejado. Isto ocorre, pois a quantidade limitada de cada unidade no banco não é suficiente para garantir todas variações prosódicas e fonéticas necessárias no processo de síntese.

Em contrapartida, a síntese por seleção de unidades, baseada em um grande *corpus*, permite que as unidades com contexto fonéticos e prosódicos desejados sejam selecionadas em tempo de execução, o que minimiza a manipulação prosódica, aumentando qualidade e naturalidade dos sinais de fala sintetizados.

Desta forma, se for considerado um conjunto de frases com várias unidades de síntese em potencial, em que múltiplas versões de uma mesma unidade podem ser encontradas, há a possibilidade interessante de encontrar uma versão desta unidade que melhor atenda às características desejadas para a síntese. Este é, portanto, o agente motivador deste trabalho: a partir de um conjunto de frases, prover ao sistema SASPRO uma maneira de selecionar as unidades de síntese para obter sinais sintetizados mais próximos da fala humana.

1.2 Objetivos

Tendo em vista que a idéia principal não é propriamente se obter um sistema TTS comercial completo, mas um desenvolvimento de aprendizagem acadêmica, o objetivo geral deste projeto é, então, realizar um estudo acerca da seleção das unidades de síntese. Como resultados práticos deste estudo, têm-se os seguintes objetivos específicos:

- Obter um arquivo único de definição de unidades;
- Desenvolver um método automático de geração deste arquivo único;
- Implementar um método completo de síntese a partir do arquivo de definições e de um banco de frases disponível;
- Estudar a implementação do processo de busca, usando diferentes critérios de seleção das unidades, incluindo aspectos de tonicidade e de prosódia (*pitch*, duração e intensidade) da unidade;
- Avaliar a qualidade e naturalidade do sinal resultante das diferentes implementações dos itens anteriores.

1.3 Organização deste Projeto

Para atingirmos os objetivos acima listados, este projeto é organizado da seguinte forma:

- No Capítulo 2, a seguir, são descritas as etapas básicas que compõem um sistema conversor texto-fala e também são apresentados alguns exemplos de sistemas TTS para português brasileiro;
- No Capítulo 3, são apresentadas as principais características do sistema SASPRO, desenvolvido por Vagner Latsch no âmbito de seu programa de doutorado, e as características importantes do método de seleção de unidade, bem como uma descrição dos possíveis tipos de unidades utilizadas;

- Em seguida, no Capítulo 4 são apresentadas as principais contribuições deste trabalho ao SASPRO, com ênfase na implementação de diferentes métodos de busca de unidades em um amplo *corpus* de frases;

- Os resultados obtidos com os diferentes métodos implementados são apresentados no Capítulo 5, em que procuramos evidenciar as melhorias alcançadas na nova versão do sistema SASPRO.

Capítulo 2

Sistemas Conversores Texto-Fala

2.1 Introdução

Os conversores texto-fala, basicamente, são sistemas que produzem voz sintetizada a partir de um texto escrito em uma língua pré-definida. Para esse tipo de sistema funcionar de maneira ideal, ele deve ser capaz de realizar, de maneira automática, um processo semelhante à leitura oral, produzindo sinais de fala com alta inteligibilidade e naturalidade. Na prática, este processo representa uma tarefa bastante complexa.

Neste capítulo serão discutidas as etapas básicas de um sistema conversor texto-fala, onde serão apresentadas a etapa de processamento de texto e suas subdivisões, e as técnicas de síntese da fala existentes. Por fim, é feita uma breve descrição de sistemas TTS acadêmicos desenvolvidos para o português brasileiro.

2.2 Etapas da Síntese de Voz

Um sistema que converte texto em fala pode ser modelado a partir de dois módulos básicos, como representado na Figura 2.1: o primeiro corresponde ao processamento do texto, que converte o texto inicial, dado como entrada para o sistema, numa representação intermediária; o segundo seria o módulo de síntese, que converte a representação intermediária em sinal de fala [1] propriamente dito.

Na literatura sobre o assunto, foram encontradas diversas denominações di-

Figura 2.1: Diagrama de blocos básico de um sistema TTS.

ferentes para estes dois blocos, mas a funcionalidade descrita para cada um deles é praticamente a mesma nas diversas referências, como detalhado nas seções a seguir.

2.3 Processamento do Texto

O módulo de processamento de texto corresponde à primeira etapa a ser realizada no processo de conversão texto-fala, mapeando o texto de entrada numa representação intermediária e representando, de forma adequada, todo o conhecimento de natureza fonética e prosódica extraído deste texto [2]. Este módulo, como pode ser visto na Figura 2.2, pode ainda ser subdividido em três partes: análise do texto, transcrição fonética e análise prosódica, que serão explicados a seguir.

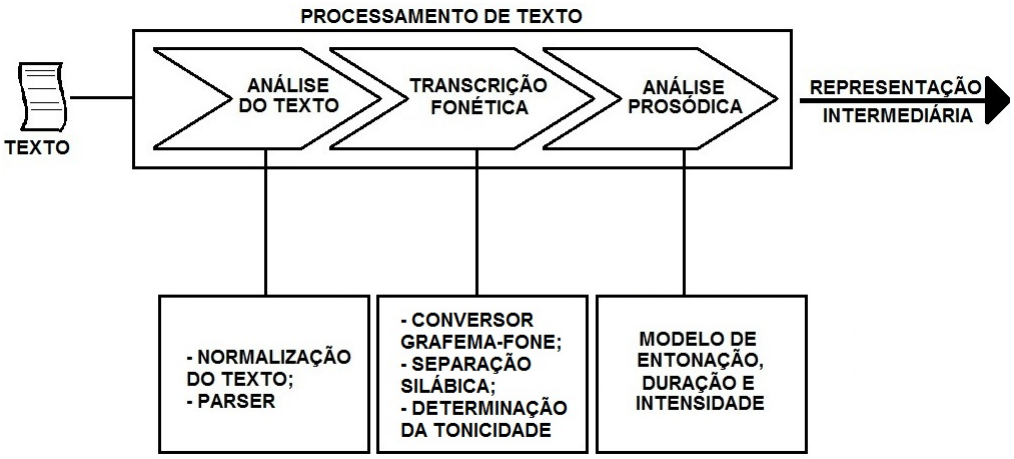


Figura 2.2: Diagrama detalhado do módulo de processamento do texto de um sistema TTS.

2.3.1 Análise do Texto

O submódulo de análise de texto pode ainda ser decomposto nas etapas de normalização e de segmentação (*parsing*) do texto, detalhadas a seguir:

- A normalização de um texto consiste em substituir certos elementos por seus equivalentes "por extenso". Os elementos que normalmente passam por esse processo são as abreviaturas (por exemplo, 'dr.' é substituído por 'doutor'), as siglas ('UFRJ' torna-se 'Universidade Federal do Rio de Janeiro'), números ('22' é mapeado em 'vinte e dois'), datas ('11/11/11' seria substituído por 'onze de novembro de dois mil e onze'), valores monetários ('R$ 10,00' é convertido em 'dez reais') e outros símbolos especiais que não pertencem ao alfabeto, tais como pontuação e barras.

  Embora a normalização, em uma primeira análise, pareça ser uma tarefa trivial, de mera substituição, em alguns casos esse processo se torna extremamente complexo. De fato, algumas operações dependem do contexto com o qual se trabalha (livro, jornal, texto científico, texto literário), e em muitos casos não é possível determinar a transformação adequada dentre as diferentes opções possíveis. As datas, por exemplo, podem aparecer em diversos formatos e mesmo uma sigla pode ter diferentes significados, que dependem do contexto ao qual o texto se refere. Outra fonte de dificuldade é a pontuação do texto, onde o ponto final ('.') normalmente funciona como indicador de final de sentença, mas que também pode ser utilizado nas abreviaturas e siglas. Analogamente, a vírgula também pode ser utilizada na composição de números decimais, e nesse caso deve ser transcrita de forma explícita.

  Uma alternativa utilizada para resolver estes problemas de normalização é o uso de regras que apontam como as diferentes formas de textos devem ser convertidas [5]. Outra maneira seria o uso de marcações em que os numerais e abreviaturas, por exemplo, possuem etiquetas que indicam como deverão ser convertidos [2].

- O papel principal da segmentação é realizar a análise morfo-sintática do texto, fornecendo elementos importantes para as etapas subsequentes de processa-

mento prosódico e de transcrição fonética [2]. A pronúncia da palavra "molho", por exemplo, depende da sua classe gramatical. Assim, nas frases "Eu não me molho" e "O molho estava muito bom", a análise morfológica ou POS (*parts of speech*) determina a pronúncia correta em cada caso.

Na prática, como normalmente os componentes prosódicos estão relacionados à estrutura sintática da sentença, a classificação sintática pode ser utilizada ainda para definir a localização provável das fronteiras prosódicas.

Para algumas palavras homógrafas, a ambiguidade não pode ser resolvida com a análise morfo-sintática. Por exemplo, na sentença " A sede da torcida é grande", somente a partir da análise do contexto é possível determinar a pronúncia correta da palavra "sede", pois em ambos os casos a classe gramatical da palavra é a mesma. Para resolver esses problemas, são agregados desambiguadores de homógrafos ao processo de análise de texto.

2.3.2 Transcrição Fonética

O submódulo de transcrição fonética, em geral, é responsável por obter a sequência de fones que caracteriza o texto de entrada, bem como sua separação silábica e tonicidade das sílabas. O processo que consiste em transformar a sequência de letras (grafemas) de cada uma das palavras do texto em uma cadeia de símbolos que representem seus respectivos sons (fonemas) é realizado pelo conversor grafema-fonema, também conhecido como grafema-fone (G2P - *grapheme to phoneme/phone*).

A transcrição fonética é realizada a partir da utilização de um conjunto de regras pré-determinadas que, em geral, levam em consideração o contexto dos grafemas adjacentes para realizar a transcrição fonética. As palavras cuja transcrição foge às regras básicas são incluídas em um dicionário de exceções cujo tamanho depende do aperfeiçoamento destas regras.

Um exemplo típico de dificuldade de transcrição por regras, na língua portuguesa, são as vogais "e" e "o", pois podem ser convertidas para vogais abertas ou fechadas [3], necessitando, portanto, de regras que se alternem.

A determinação da sílaba tônica e divisão silábica, que são realizadas nesta etapa, normalmente, também são efetuadas a partir de um conjunto de regras cor-

respondentes. Essas informações desempenham papel importante no estudo dos modelos prosódicos, conforme será explicado mais adiante.

2.3.3 Análise Prosódica

De acordo com Taylor [1], a prosódia pode ser vista como a parte da comunicação falada que expressa emoção, atitudes e intenções do locutor. Assim sendo, a prosódia apropriada é essencial para produzir sinais de fala sintetizados com características semelhantes às da fala natural. A prosódia pode ser apresentada em dois níveis: segmental e supra-segmental [1]. O nível segmental ocupa-se com a observação da variação dos parâmetros prosódicos (duração, freqüência fundamental $F0$ e amplitude) ao nível dos segmentos. Esse nível visa principalmente a interação do segmento com seus segmentos vizinhos e a interferência que esses vizinhos realizam sobre o segmento em questão. Já o nível supra-segmental tem objetivo de estruturar a sentença ao nível de sílabas, palavras, frases, etc. Deve-se ressaltar que apesar da prosódia neste caso está sendo observada no nível supra-segmental os parâmetros prosódicos (duração, frequência fundamental F0 e amplitude) estão diretamente associados aos segmentos da sentença.

Em termos quantitativos, a prosódia se refere a como algo é dito, através da variações de *pitch* (entonação), duração e intensidade, discutidas a seguir.

- Taylor [1] define a entonação como o uso sistemático de variações do *pitch* para a comunicação. O *pitch* está intimamente associado a frequência fundamental $F0$, que, por sua vez, está relacionada à periodicidade do sinal nos trechos sonoros. Alguns autores, como citado em [2], afirmam que a frequência fundamental é a característica mais importante no modelamento prosódico.

  O objetivo dos sistemas TTS em determinar a entonação é encontrar o modelo apropriado para a evolução temporal do contorno de *pitch*, ao longo de cada segmento de texto a ser sintetizado. Há diversas estratégias que podem ser utilizadas para modelagem e codificação do contorno de *pitch*.

  No modelo TILT, desenvolvido por Taylor [1], a entonação é representada por uma forma linear de eventos, que podem representar o valor de $F0$, com o

modelo em si tratando apenas de duração de subida e descida da curva de $F0$.

O modelo de Fujisaki tem origem fisiológica e procura representar a parte do aparelho fonador humano responsável pelas variações de $F0$. A descrição do contorno de $F0$ é feita através de duas componentes: as acentuais e as frasais. As componentes frasais são modeladas como funções impulsos e as componentes acentuais como funções pulsos. Todas essas componentes funcionam como entradas de filtros lineares de segunda ordem criticamente amortecidos, cujas saídas são somadas para gerar a curva de contorno de $F0$, na escala logarítmica [1].

- A duração é o parâmetro que mede a distância temporal do início ao fim de um segmento fonético. Ela é considerada o segundo parâmetro prosódico mais importante [1], variando de acordo com a ênfase ou o ritmo do discurso desejados.

  Assim como na entonação, várias estratégias podem ser utilizadas para implementar um modelo de duração.

  O modelo de Klatt, por exemplo, é um modelo duracional baseado em regras [2]. Neste modelo, assume-se que todos os segmentos possuem uma duração inerente e uma duração mínima. Na prática, os segmentos podem ser alongados ou encurtados (nunca abaixo da duração mínima) de acordo com o ambiente prosódico em que se encontram.

  No modelo de Campbell, a sílaba é utilizada como unidade fundamental de duração [1]. Desta forma, a duração da sílaba é calculada a partir dos fatores contextuais no nível da sílaba utilizando uma rede neural que procura implementar o princípio de elasticidade, que afirma que todos segmentos que compõem a sílaba são expandidos ou contraídos do mesmo fator.

  No sistema SASPRO [3], porém, foi proposta uma adaptação do modelo de CAMPBELL, em que a duração da sílaba é definida diferentemente para a sua vogal-núcleo e para as demais componentes da sílaba.

- A intensidade é um parâmetro prosódico que é proporcional à energia do sinal de fala. Por considerarem a intensidade uma variável de menor relevância,

a maioria dos trabalhos de modelagem prosódica não dão atenção a esse parâmentro. No sistema SASPRO [3], a intensidade silábica foi representada pela intensidade da vogal-núcleo.

2.4 Síntese da Fala

Existem diferentes técnicas que podem ser utilizadas para efetuar a geração do sinal de fala sintética, mas o objetivo final de cada uma dessas técnicas é sempre o mesmo: gerar um sinal acústico que corresponda à sequência de fonemas e ao perfil prosódico determinados pelo módulo de processamento de texto. O mecanismo de síntese também deve procurar evitar descontinuidades ao longo do sinal gerado, de forma que a fala produzida ao final do processo seja inteligível e tão natural quanto possível.

De acordo com Taylor [1], as técnicas de síntese da fala podem ser classificadas em três gerações: a primeira baseada na modelagem do trato vocal, como os sintetizadores por formantes ou por predição linear; a segunda, que se fundamenta na síntese concatenativa em que a fala sintética é produzida através da concatenação de segmentos; e, por fim, a terceira, que é baseada em um grande corpus de voz, como os sintetizadores por seleção de unidades e os sintetizadores por modelos estocásticos baseados em HMMs (*hidden Markov models*). Estes três tipos de sintetizadores são apresentados, com um maior detalhamento, nas subseções a seguir.

2.4.1 Síntese por Formantes

O modelo de síntese por formantes ou por regras é baseado no modelo fonte-filtro da teoria acústica de produção de fala. Para geração da voz sintética neste modelo é preciso modelar a fonte de excitação e os filtros que simulam o trato vocal através de suas funções de transferência.

O modelo de síntese por regras precisa da determinação dos parâmetros relacionados à fonte de excitação, tais como período de pitch, amplitude, presença ou ausência de ruído de aspiração (sonoridade) etc. Além destes, são considerados ainda os parâmetros relacionados ao trato vocal, como frequência, amplitude e largura de

banda dos formantes, presença de pólos e zeros nasais etc.

A Figura 2.3 mostra o diagrama de blocos simplificado de um sintetizador por formantes, onde apenas o *pitch* e os formantes aparecem como parâmetros. Na prática, tais sistemas têm cerca de quarenta parâmetros [2].

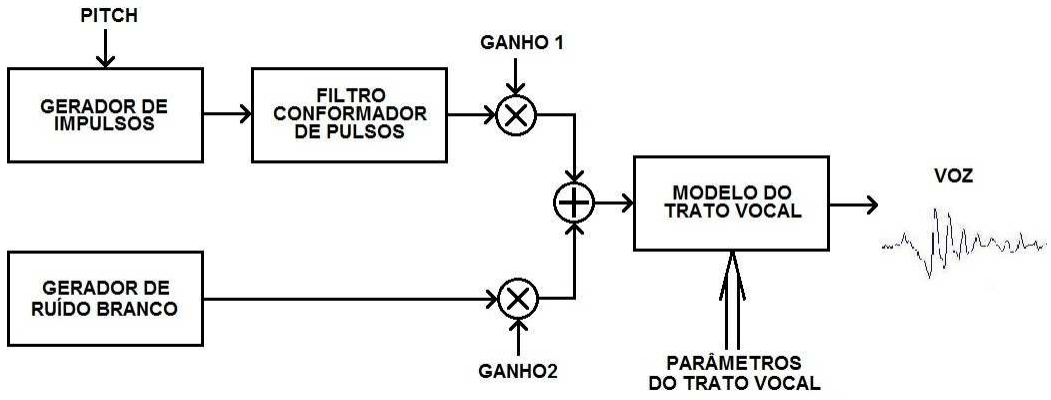


Figura 2.3: Diagrama de blocos simplificado do sintetizador por formantes.

A primeira versão do sintetizador de formantes de Klatt apresentada em 1980 funcionou como alavanca inicial dos sistemas de síntese por regras [5]. A técnica de síntese por formantes é bastante flexível, pois permite a geração de diferentes qualidades de voz a partir do ajuste de parâmetros do modelo da fonte [2].

O principal impedimento prático à utilização da síntese por regras em um sistema de conversão texto-fala deve-se à dificuldade na determinação dos parâmetros de controle do sintetizador que em geral precisa de uma fase de treinamento do modelo complexa e exaustiva. Para os sistemas de síntese a partir de texto, esses parâmetros devem ser determinados de maneira automática.

Quanto à qualidade dos sinais sintetizados a partir desta técnica, os sinais obtidos são intelegíveis mas com baixa naturalidade [1]. Umas das razões da falta de naturalidade ocorre pela dificuldade de se capturar as nuances da fala humana em um determinado conjunto finito de regras e parâmetros que compõem o sintetizador.

2.4.2 Síntese por Concatenação

A idéia da síntese concatenativa é gerar o sinal de fala a partir da justaposição de segmentos pré-gravados de voz. Tais segmentos são selecionados a partir de

um inventário de unidades previamente construído. Ao contrário da síntese por formantes, a síntese concatenativa não requer a definição de regras.

Como cada segmento é gravado de uma fala natural, espera-se que a saída do sistema também seja natural. Na prática, porém, o que se percebe é que a concatenação pura e simples não é capaz de gerar um sinal de fala com qualidade satisfatória. De fato, algumas vezes o sinal de fala gerado por meio dessa técnica não chega nem mesmo a ser inteligível e muito menos natural. A principal razão disso é o fenômeno da coarticulação [2] presente em nossa fala natural, em que um som influencia a geração do som subsequente e mesmo o anterior. Assim, a simples justaposição de dois segmentos de fala pode gerar descontinuidades espectral e prosódica. Descontinuidades espectrais ocorrem quando os formantes na região de transição não coincidem, e a descontinuidade prosódica ocorre quando o *pitch* é distinto em duas unidades concatenadas. Sendo assim, para minimizar, ou mesmo evitar, estes efeitos indesejados, o processo de escolha das unidades segmentais a serem concatenadas é decisivo no desenvolvimento de um sintetizador por concatenação.

Outro problema dos sistemas de segunda geração é que eles não generalizam bem contextos que não foram incluídos no processo de formação dos segmentos, em geral porque a variabilidade prosódica real é muito grande [2]. Existem, portanto, técnicas que permitem modificar a prosódia de um sinal concatenado para atingir a prosódia desejada. Essas técnicas, porém, se não devidamente ajustadas, podem provocar novos efeitos indesejados, degradando a qualidade da fala sintética resultante [1].

As técnicas de manipulação prosódica mais utilizadas são as da família PSOLA (*pitch synchronous overlap and add*) [1]. Neste tipo de técnica, inicialmente, o sinal original é separado em janelas sincronizadas com o *pitch*. Esta segmentação é feita de modo que ocorra sobreposição das janelas consecutivas. Os parâmetros prosódicos são modificados de acordo com o perfil prosódico (de *pitch* e duração) desejado, e, em seguida, uma nova sequência de sinais é gerada a partir da adição da sequência de segmentos originais, como ilustrado na Figura 2.4.

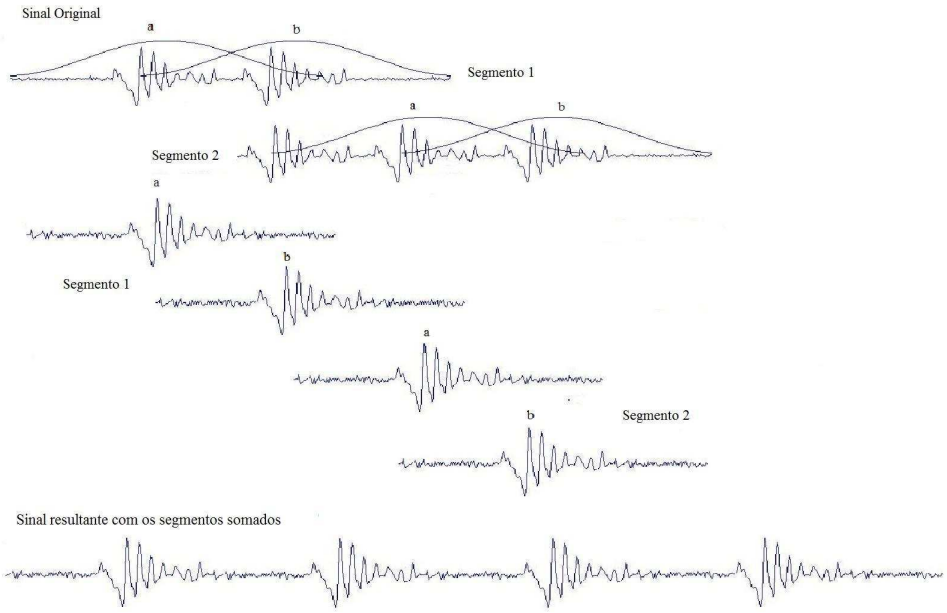


Figura 2.4: Procedimento básico do PSOLA.

2.4.3 Síntese Baseada em um Grande *Corpus*

Os sistemas de terceira geração são baseados em um grande *corpus* de fala, e dentre esses estão os sintetizadores por modelos estocásticos baseados em HMMs e os sintetizadores por seleção de unidades.

Ainda de acordo com Taylor [1], a síntese por HMM começou com o clássico artigo de Tokuda [14], em que são explicados os princípios básicos de como os modelos ocultos de Markov podem ser utilizados para sintetizar sinais de fala. Dentre as vantagens da síntese por HMM tem-se o fato de os modelos treinados poderem produzir sinais de alta qualidade, serem compactos, e de fácil modificação para a transformação de voz e outros fins. Entre as desvantagens ele cita o fato de a qualidade do sinal final ser muito dependente do modelo paramétrico utilizado, o que depende de uma ótima segmentação e etiquetagem do sinal de voz, o que são tarefas muito especializadas.

A técnica de síntese da fala por seleção automática de unidades ou *unit selection*, tema deste trabalho, pode ser vista como uma extensão da técnica por concatenação e tem sido a mais utilizada entre os sistemas comerciais [1] mais recentes. Essa técnica se utiliza de bancos com grande quantidade de unidades, o que aumenta a possibilidade de, no processo de seleção, se encontrar unidades próximas das

"ideais", inclusive nos aspectos prosódicos. Desta forma, possíveis descontinuidades espectrais entre as unidades adjacentes serão minimizadas, o que garante a naturalidade do sinal gerado, uma vez que o contorno prosódico adequado foi empregado. Entretanto, a grande variabilidade dos contextos prosódico e fonético torna-se um fator limitante desta técnica [11], como será visto, mais adiante, na Seção 3.3.

2.5 Outros Sistemas TTS Existentes para o Português Brasileiro

Nesta seção serão apresentados alguns exemplos de desenvolvimento de sistema TTS para o português brasileiro.

2.5.1 ORADOR

O ORADOR [6] é um sistema de conversão texto-fala que vem sendo desenvolvido pelo grupo do Linse (Laboratório de Circuitos e Processamento de Sinais) do Departamento de Engenharia Elétrica da Universidade Federal de Santa Catarina, desde 1998. A abordagem utilizada nesses sistemas é a síntese concatenativa, onde o ponto de articulação dos segmentos é levado em conta para realizar a concatenação.

O banco de unidades que são utilizadas na concatenação foi criado a partir de uma prévia gravação de um *corpus* de texto por um locutor profissional. Em versões anteriores, o repertório de unidade era composto principalmente por trifones e alguns polifones. Atualmente, o sistema se baseia em unidades de tamanho variável, podendo essas unidades constituírem-se de apenas um fonema ou até de unidades maiores como sílabas, palavras, sintagmas e frases curtas.

De acordo com a página oficial do projeto ORADOR [6], onde alguns exemplos de síntese podem ser encontrados, o financiamento da pesquisa tem sido feito pela Dígitro Tecnologia e pelo Ministério de Ciência e Tecnologia (CNPq).

2.5.2 Aiuruetê

O Aiuruetê [7] é um sistema TTS para português brasileiro. O nome do sistema, "Aiuruetê", vem do nome tupi-guarani do papagaio. Este sistema foi desenvolvido em conjunto por pesquisadores das áreas de linguística e de engenharia elétrica da Unicamp, sendo iniciado como um estudo de descrição fonético-acústica da língua, no âmbito do Laboratório de Fonética e Psicolinguística (Lafape) do Instituto de Estudos da Linguagem (IEL). Posteriormente, o projeto passou também a contar com a participação do Laboratório de Processamento Digital de Fala (LPDF) da Faculdade de Engenharia Elétrica (FEEC) da Unicamp.

No Aiuruetê, as regras de transcrição fonética, bem como as regras para uma correta pronúncia no processo de síntese, são realizadas pelo conversor ortográfico-fônico denominado *Ortofon*, desenvolvido pelo Lafape. A técnica utilizada no processo de síntese é a concatenação de polifones, trechos sonoros com dois ou mais fonemas, que estão armazenados em um dicionário sonoro com aproximadamente 2.500 diferentes fragmentos de sons extraídos de gravações. Os módulos e interfaces foram escritos em linguagem C++, e a interface entre o aplicativo e o usuário em Delphi. O Aiuruetê foi desenvolvido para o sistema operacional Windows e financiado pela FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo).

2.5.3 LianeTTS

O LianeTTS [8] é sistema conversor texto-fala para leitura de tela de computador. Foi desenvolvido pela SERPRO em parceria com o Núcleo de Computação Eletrônica (NCE) da UFRJ, numa evolução da plataforma DosVox. Esse sistema gera voz em português brasileiro, com sotaque carioca.

O funcionamento do aplicativo LianeTTS ocorre pela análise do texto que em seguida o traduz em um texto compilado no formato de difones (.pho). Para o procedimento de síntese da voz, o sistema LianeTTS utiliza o MBROLA, uma variante do método PSOLA. O *software* opera nos sistemas operacionais Linux e Windows e está disponível no portal do SERPRO (www.serpro.gov.br).

2.6 Conclusão

Neste capítulo discutimos as etapas básicas de um sistema conversor texto-fala: o processamento de texto, passando pela normalização do texto, segmentação (*parsing*), transcrição fonética e análise prosódica; e as técnicas de síntese, que foram divididas em três grupos: síntese por formantes, síntese concatenativa e síntese baseada em um grande *corpus*. Em seguida, foi feita uma breve apresentação de dois sistemas conversores texto-fala acadêmicos e desenvolvidos para o português brasileiro.

No capítulo seguinte, apresentamos o sistema SASPRO, desenvolvido pelo pesquisador Vagner Latsch no contexto de seu doutorado no Programa de Engenharia Elétrica na COPPE-UFRJ [3].

Capítulo 3

O Sistema SASPRO

3.1 Introdução

O SASPRO é o sistema de análise e síntese da prosódia desenvolvido por Vagner Latsch no âmbito de seu doutorado na COPPE-UFRJ [3]. Vale ressaltar que, nesse contexto, o objetivo do autor não era construir um sistema comercial, mas uma ferramenta de apoio à pesquisa e desenvolvimento de sistemas de conversão texto-fala.

O sistema foi implementado em linguagem C++, na plataforma Microsoft Visual Studio 2010 sob o paradigma de orientação por objetos, o que permite maior reaproveitamento dos códigos além de facilitar a manutenção do sistema.

O SASPRO inclui um conversor TTS protótipo e diversas funcionalidades de processamento de texto e de modelagem prosódica. As ferramentas de *software* pertencentes ao sistema são de fácil manipulação e permitem ao usuário visualizar, por exemplo, todas as etapas de análise e síntese de prosódia.

Na Seção 3.2, serão apresentadas, resumidamente, as principais características das etapas de conversão texto-fala do sistema SASPRO. Além disso, na Seção 3.3 será melhor descrito o método de síntese por seleção automática das unidades, abordando os tipos de unidades de síntese mais utilizadas. E, por fim, será vista a generalização do processo de busca ideal realizado por Hunt e Black, como apresentado em [9].

3.2 Características do Protótipo TTS

Nesta seção, pretende-se dar um idéia geral das etapas de conversão de um texto para fala do sistema SASPRO original, sintetizando as informações contidas em [3].

3.2.1 Análise do Texto

No sistema TTS do SASPRO, a análise do texto de entrada é iniciada com uma análise morfológica realizada a partir de um classificador automático, que atribui, a cada uma das palavras do texto, etiquetas com as suas respectivas categorias morfológicas. Neste caso, o classificador utilizado foi o etiquetador TBL (*Transformation Based Learning*) descrito em [15].

A etapa inicial de normalização do texto foi dispensada considerando, portanto, que o texto de entrada contém apenas palavras da língua portuguesa e os respectivos símbolos de pontuação.

A etapa seguinte é a de transcrição fonética, que é iniciada pela conversão grafema-fonema, resultado da aplicação de um conjunto de regras que operam no domínio da palavra. Para a transcrição fonética das palavras que são exceções às regras utiliza-se um dicionário de exceções. Na representação fonética, o sistema SASPRO utiliza uma sequência de fones identificados por uma estrutura composta por atributos próprios e pré-determinados. Estes atributos são um código, um símbolo e a classificação quanto à sonoridade e a informação sobre sonoro/surdo. Nesse processo considerou-se o dialeto típico de um carioca.

Em seguida, a etapa de separação silábica foi realizada atendendo a um conjunto de regras baseadas na composição da sílaba e levando em consideração o fato de que a separação silábica é mais natural quando realizada em termos do agrupamento de fones ao invés do agrupamento de grafemas/caracteres. Desta forma, a separação silábica foi realizada somente após a conversão grafema-fonema. Para formação da sílaba, o autor adotou um modelo composto por um *núcleo* e componentes adjacentes à sua esquerda (*onset*) e à sua direita (*coda*). O núcleo é sempre ocupado por uma vogal, enquanto que o *onset* e a coda da sílaba podem ser vazios.

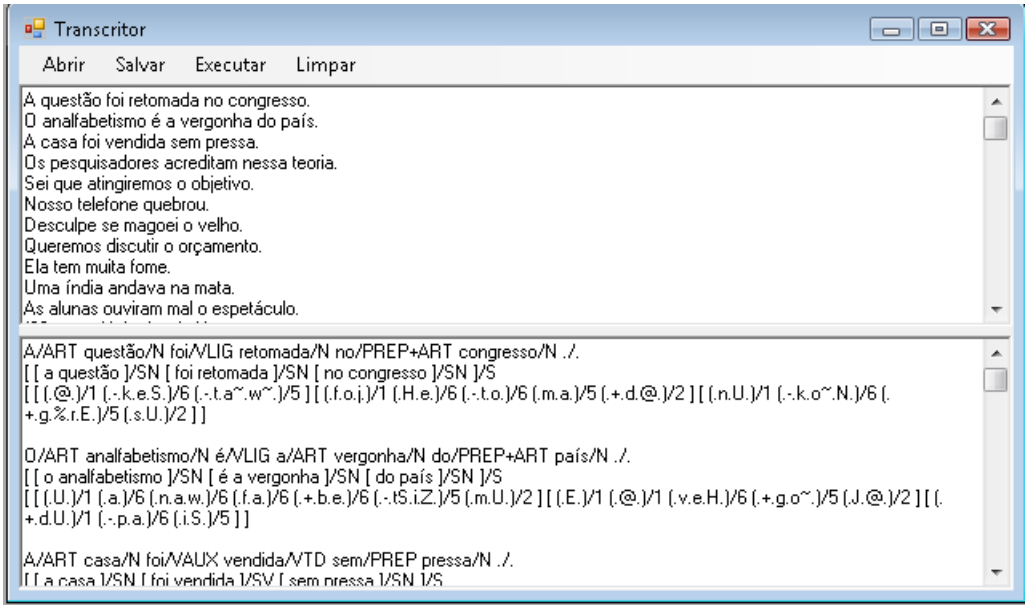


Figura 3.1: Tela de execução do processamento do texto do sitema SASPRO indicando: o texto de entrada, a classificação gramatical de cada palavra, a transcrição fonética e a separação silábica correspondente.

Após a separação silábica é aplicado o procedimento de regras para determinação da sílaba tônica. A partir deste procedimento têm-se a indicação da sílaba tônica e a caracterização do posicionamento da sílaba em relação à sílaba tônica: pretônica, tônica, postônica medial, postônica final, mossílabo átono e monossílabo tônico.

Ainda na etapa de processamento, foram implementadas regras pós-silábicas adicionais que fazem referência à tonicidade ou à sílaba, tal como a redução das vogais. Por fim, para concluir o módulo de processamento de texto, o autor aplica as regras de junção de palavras denominadas de regras pós-lexicais.

Embora a etapa do processamento do texto seja utilizada no sistema protótipo TTS, o sistema SASPRO inclui um aplicativo individual em que as etapas de processamento de texto podem ser observadas, como ilustrado na Figura 3.1. Nesta figura, têm-se o texto de entrada, na parte superior da tela, o resultado do processamento deste texto, na parte inferior. O texto de entrada pode ser escrito ou importado de um arquivo texto com caracteres no formato UTF-8. O resultado do processamento é exibido em três linhas de texto. A primeira linha mostra o

resultado da classificação morfossintática, onde para cada palavra é exibida sua respectiva classificação morfológica. A segunda linha mostra a separação das palavras acompanhada da pontuação correspondente. Na última linha pode ser vista a representação fonética, onde os símbolos que identificam os fonemas são agrupados por sílabas, seguidas da indicação de tonicidade.

3.2.2 Processamento Prosódico

Com o objetivo de observar a evolução dos parâmetros prosódicos, o sistema SASPRO inclui ainda o processo de modelagem prosódica que é realizada em duas ações, a de análise e a de síntese.

Assim, na etapa de análise, os parâmetros dos fones são encontrados e um procedimento de normalização dos parâmetros prosódicos (duração e intensidade) desses fones é estabelecido. Em seguida, os parâmetros da sílaba são obtidos a partir dos parâmetros dos fones que a compõem.

Na etapa de síntese, os parâmetros dos fones individuais são obtidos a partir dos parâmetros da sílaba, definidos por algum padrão prosódico desejado.

No modelo das durações do SASPRO, diferentes padrões prosódicos (atitudes) foram modelados, determinando-se um modelo de alteração da duração da sílaba para cada padrão prosódico, numa adaptação do modelo de Campbell. Desta forma, a duração da sílaba passou a ser representada por dois parâmetros ao invés de um: um relativo ao *onset* e o outro relativo ao núcleo da sílaba.

Para a modelagem da intensidade do sistema SASPRO, a intensidade de cada sílaba é representada pela intensidade da vogal-núcleo. Desta forma, a intensidade dos demais segmentos da sílaba foi obtida pela correlação da intensidade de cada segmento com a intensidade da vogal-núcleo.

Com o intuito de tornar o contorno de *pitch* menos dependente do falante, a modelagem de entonação realiza uma normalização inicial pelo *pitch* característico do falante, no domínio logarítmico. Além disso, a curva de contorno de *pitch* é aproximada por quatro pontos, tomados nos limites da vogal-núcleo da sílaba. Desta forma, a velocidade das variações no contorno de *pitch* acompanha a velocidade de emissão da vogal. Assim, quando a sílaba é alongada, o intervalo entre amostras do

contorno também é alongado, diminuindo a inclinação da curva e vice-versa.

De posse destes modelos, o sistema SASPRO representa as variáveis prosódicas de duração, intensidade e *pitch* codificadas, no nível da sílaba, por um vetor de parâmetros. A evolução deste vetor ao longo das sílabas de uma sentença permite que o comportamento das variáveis prosódicas seja observado mais facilmente.

A prosódia de uma determinada sentença pode ser manipulada utilizando o SASPRO como ferramenta de síntese, como representado na Figura 3.2. Neste sentido, o usuário pode facilmente alterar (diminuir ou aumentar) qualquer uma das três características prosódicas com um simples arrastar do *mouse*, podendo ouvir imediatamente o resultado dessa operação.

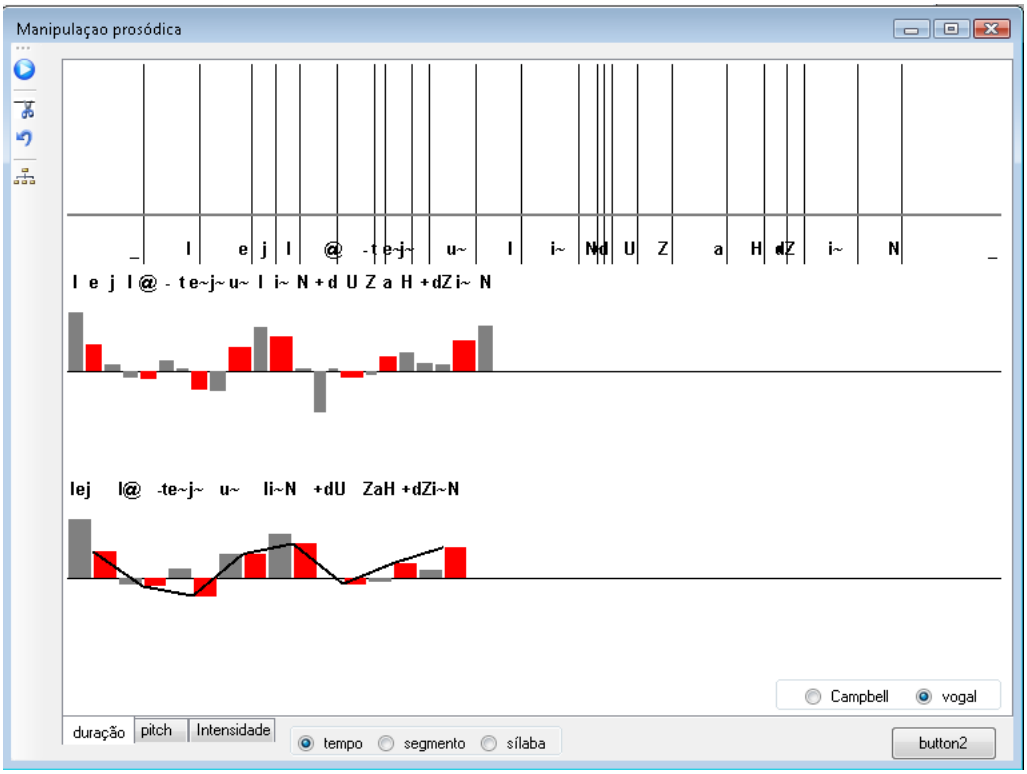


Figura 3.2: Tela de ferramentas de manipulação de prosódia do sistema SASPRO.

3.2.3 Síntese da Fala

A abordagem inicial do sistema SASPRO é a síntese concatenativa, em que o sinal de fala é gerado através da justaposição de segmentos temporais previamente gravados e armazenados em um banco de unidades reduzido. O algoritmo

TD-PSOLA é utilizado, então, para manipular as variáveis prosódicas do sinal concatenado.

A Figura 3.3 mostra a tela de execução do conversor TTS protótipo do SASPRO, onde pode ser vizualizada, como exemplo, a conversão de texto para fala da sentença "Ela tem muita fome". A sentença foi escrita na entrada de texto e em seguida foram executadas as etapas de processamento de texto e síntese do sinal. O primeiro gráfico mostra o sinal gerado da concatenação das unidades e as respectivas marcas de *pitch*. O gráfico inferior mostra o espectograma do sinal resultante com as respectivas etiquetas de fronteira.

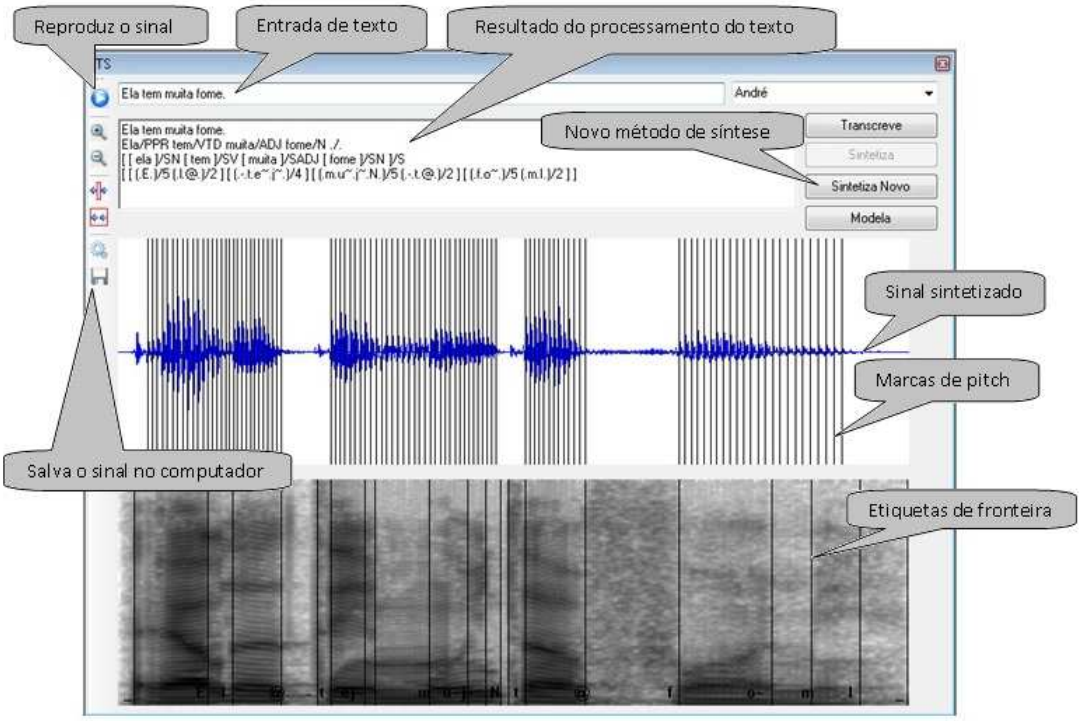


Figura 3.3: Tela do sistema protótipo TTS do SASPRO com exemplo da conversão texto-fala da sentença "Ela tem muita fome".

Ainda na Figura 3.3, na parte superior direita, foi destacado o botão que executa a síntese de fala com o novo método implementado no âmbito do presente projeto de graduação.

3.2.4 Banco de Unidades

O sistema SASPRO foi utilizado para a construção do banco de unidades que são utilizadas na concatenação do sinal de fala. As unidades eram obtidas a partir de logotomas, isto é, palavras sem sentido e que contêm a sequência de fonemas que compõem a unidade desejada. Esses logotomas são escolhidos de forma que o contexto fonético em torno dos fones da unidade a ser extraída seja o mais neutro possível, a fim de minimizar a coarticulação destes com os segmentos vizinhos que não fazem parte da unidade. O recorte das unidades era realizado manualmente pelo usuário, que seleciona o trecho da unidade e solicita que a mesma seja adicionada a um arquivo de definição. O banco em si é descrito por um arquivo único em que cada linha contém o nome-código de cada unidade, o arquivo .wav de onde ela foi retirada, assim como os instantes de tempo que delimitam a unidade no sinal.

3.3 Síntese por Seleção Automática de Unidades

O método de seleção automática de unidades foi proposto inicialmente para que houvesse o mínimo possível de manipulação prosódica [9], o que costuma ser a principal causa da perda da naturalidade dos sinais gerados pelos sistemas TTS de segunda geração. A seguir, serão feitas algumas considerações sobre o tipo de unidades a serem armazenadas e o processo de busca da unidade adequada.

3.3.1 Tamanho das Unidades de Síntese

O processo de definição das unidades deve representar um compromisso com a naturalidade e inteligibilidade. Assim, a escolha do tamanho das unidades a serem utilizadas no processo de síntese é uma das decisões importantes que precisam ser consideradas ao se projetar um sistema conversor texto-fala. Com esta motivação, foi realizado um estudo bibliográfico sobre o assunto. A partir deste estudo, percebe-se que são várias as possibilidades de tamanho e quantidade de unidades que podem ser utilizadas, como é descrito a seguir:

- Fones: Os fones ou fonemas são a unidade básica de uma língua. Com a introdução dos algoritmos de seleção automática de unidades, os fones foram

considerados como possíveis unidades de síntese [9]. Contudo, o que se observa na prática é que os sistemas baseados neste tipo de unidades apresentam um comportamento não muito estável, oscilando entre falas sintetizadas com uma alto grau de naturalidade e falas sintetizadas com distorções desagradáveis [12]. Uma das razões para este problema é que todos os pontos de concatenação passam a ser realizados nas fronteiras dos fones, o que dificulta a representação precisa do efeito de coarticulação [5].

- Metade dos Fones: Como o próprio nome sugere, estas são unidades que têm a metade do tamanho dos fones. Desta forma, estas unidades se estendem desde as fronteiras entre fones até o ponto médio ou se estendem a partir deste ponto médio até o final do fone [1]. Esta unidade oferece mais flexibilidade do que um fone e tem se mostrado útil em sistemas que usam múltiplas instâncias de unidades [2]. Porém, quando utilizadas de forma isolada apresentam a mesma, se não mais, dificuldade com o processo de coarticulação que os fones.

- Difones: O difone é uma unidade formada por uma dupla de fones. Ele se inicia na metade do primeiro fone e termina na metade do fone seguinte [1]. A grande vantagem do difone é que a transição entre os fones está inteiramente contida no interior da unidade, capturando grande parte das transições fonéticas e justificando o amplo uso destas unidades em sistemas TTS até mesmo comerciais de segunda geração. Por outro lado, os difones incluem apenas parte dos vários efeitos coarticulatórios da língua falada, o que justifica o uso, mesmo que parcial, de unidades maiores.

- Trifones: A partir dessa desvantagem dos difones, surgiu a proposição da unidade trifone, que inclui um fone inteiro e suas transições à esquerda e à direita [4]. Neste caso, a dificuldade passa a ser o número excessivo de unidades, o que justifica o uso de trifones como um complemento, para casos de sons especiais, de bancos baseados em unidades menores.

- Sílabas: A coarticulação que ocorre entre fonemas pertencentes a sílabas diferentes normalmente é bem menor do que aquela que ocorre entre segmentos intra-silábicos. Desta forma, as sílabas podem ser consideradas unidades na-

turais [2]. Entretanto, o número de sílabas existentes na língua é muito grande (no inglês, por exemplo, há cerca de dez mil sílabas diferentes [2]), o que dificulta a montagem de um banco puro e simples baseado neste tipo de unidade.

- Demissílabas: As demissílabas são baseadas no mesmo princípio fonológico das sílabas. Esta unidades são formadas a partir da divisão das sílabas em duas partes parcialmente sobrepostas, com o pico silábico (núcleo) pentencendo a ambas as partes. Por exemplo, a sílaba "bar" na palavra "barco" possui uma demissílaba inicial "ba" e uma demissílaba final "ar". Um problema desta divisão é o de que nem sempre é possível desprezar a interação que ocorre entre os segmentos pertencentes a sílabas diferentes [5].

- Palavras: As unidades de um sistema TTS podem até mesmo ser palavras ou mesmo frases. A dificuldade neste caso, naturalmente, é o elevado número de unidades necessárias para um sistema de síntese irrestrita, o que faz com que este tipo de banco só seja utilizado em sistemas de síntese especializada [2].

Os sistemas podem ser homogênos, isto é, utilizando um único tipo de unidade para concatenação, ou heterôgeneos, que fazem uso de dois ou mais tipos de unidades. Em geral, a tendência atual é para sistemas mistos [1], com as unidades maiores sendo usadas para aumentar a naturalidade do sinal gerado e as unidades menores com o papel de limitar o tamanho do banco de unidades.

3.3.2 Busca da Unidade de Síntese

Como já foi dito, com a seleção automática das unidades de síntese, a fala sintetizada será gerada a partir de um grande *corpus* de fala no qual se busca a sequência de unidades de síntese que melhor satifaz as seguintes propriedades:

- A sequência deve apresentar um contorno fonético e prosódico o mais próximo possível do desejado;

- As unidades de síntese selecionadas devem minimizar possíveis descontinuidades espectrais ao longo das junções entre elas.

Assim, em outras palavras, o principal objetivo do processo de seleção automática de unidades é evitar, o máximo possível, qualquer tipo de modificação prosódica ou mesmo espectral das unidades de síntese selecionadas.

A Figura 3.4 ilustra o processo básico de seleção automática proposto por Hunt e Black [9], que se baseia no uso de duas funções-custo principais:

- Custo fonético-prosódico, $C^t(u_i, t_i)$, que estima a diferença (distância) entre o contexto fonético-prosódico unidade de síntese, u_i, presente na base de dados, e o contexto fonético-prosódico desejado, t_i;
- Custo de concatenção, $C^c(u_{i-1}, u_i)$, que estima o grau de descontinuidades entre duas unidades de síntese consecutivas u_{i-1} e u_i.

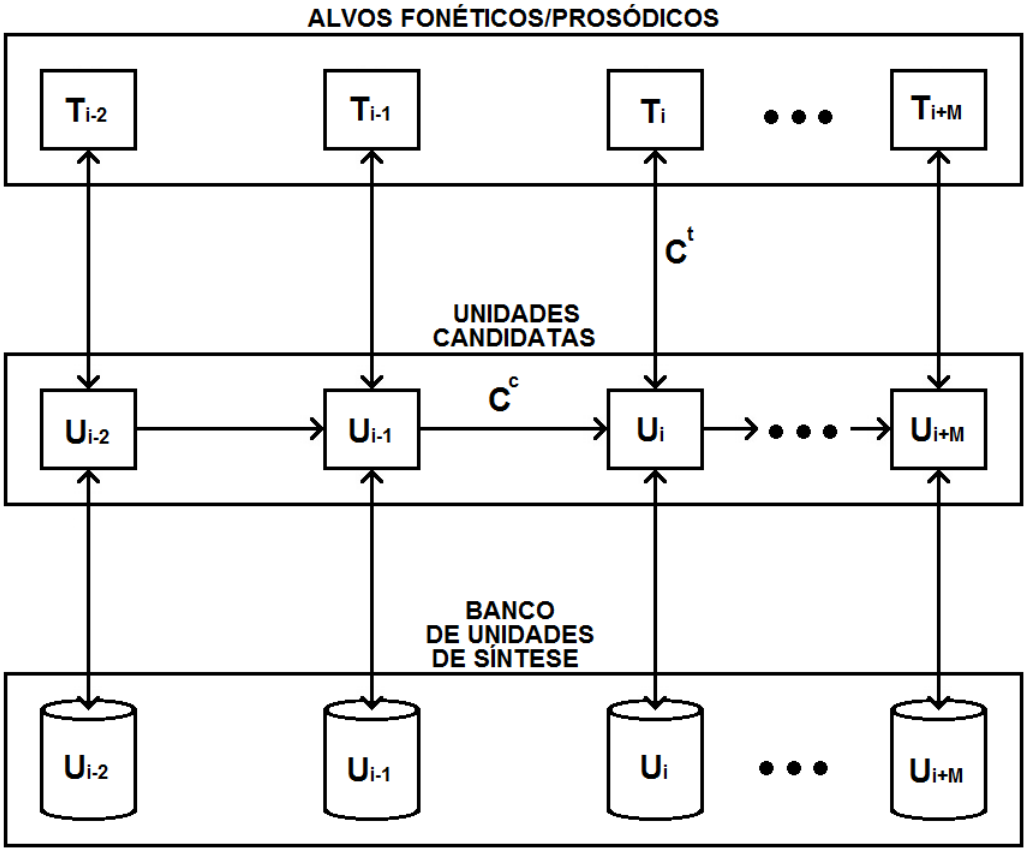


Figura 3.4: Processo de seleção automática de unidades de síntese.

Assim, uma primeira etapa de busca avalia a função de custo fonético-prosódico de todas as unidades de síntese candidatas, enquanto uma segunda etapa estima o

custo de concatenação associado a cada transição entre duas unidades de síntese candidatas. Por último, um algoritmo é utilizado para encontrar a sequência ótima de unidades de síntese que minimiza a soma acumulada dos dois custos acima definidos.

As funções de custo fonético-prosódico incluem, em geral, rótulos fonéticos acerca da duração e *pitch* das unidades. Na prática, a estimação dessas funções é uma tarefa muito complexa, principalmente quando se consideram aspectos perceptuais. Dentre as métricas utilizadas tem-se a distância Euclidiana [1], utilizada neste trabalho.

O custo de concatenação é calculado na fronteira das unidades de síntese, quantificando o grau de descontinuidade no ponto de junção entre as unidades, incluindo o aspecto de descasamento das características espectrais. Vários estudos têm sido realizados em busca de representações acústicas e métricas que possam estimar o custo de concatenação que melhor se aproxime das percepções humanas [1], como por exemplo: coeficientes cepstrais, coeficientes cepstrais a partir da predição linear (LP, *linear prediction*), coeficientes LSF (*line spectrum frequencies*) etc.

Desta forma, a **função-custo total** de uma sentença a ser sintetizada será dada por [9]

$$C(\overline{U}, T) = \sum_{i=1}^{N} C^t(t_i, u_i) + \sum_{i=2}^{N} C^c(u_{i-1}, u_i), \tag{3.1}$$

onde $T = [t_1, t_2, ...t_N]$ é a sequência de alvos prosódicos-fonéticos desejados e $\overline{U} = [u_1, u_2, ...u_N]$ é uma sequência de unidades de síntese candidatas à sequência ótima. As funções-custo são, ainda, divididas em subcustos e seus valores estimados como uma soma ponderada de seus respectivos subcustos. Os pesos atribuídos aos subcustos podem ser calculados, por exemplo, por um método de regressão linear [9].

O objetivo principal da função de busca será encontrar o vetor $\overline{U}$ que minimiza a função-custo total. Este procedimento de busca pode ser solucionado com um algoritmo de programação dinâmica. Contudo, para grandes bases de unidades de síntese, esta busca pode demandar um alto custo computacional, que pode ser reduzido com técnicas de "clusterização" das unidades de síntese [9]. Essas técnicas são bastante utilizadas em sistemas de reconhecimento de voz e permitem controlar a relação entre o tamanho da base de dados, a variabilidade acústica e a complexidade computacional do processo de busca da unidade ótima.

3.4 Conclusão

Neste capítulo, apresentamos o sistema SASPRO desenvolvido por Vagner Latsch ao longo de seu doutorado, bem como suas principais funcionalidades no contexto do processo de conversão texto-fala.

Também foi apresentado o método de síntese para estes conversores utilizando o método de seleção automático das unidades, acompanhado de um breve estudo acerca dos principais tipos de unidades utilizadas no processo de síntese de voz, ressaltando as vantagens e desvantagens de cada tipo.

Por fim, foi descrito o procedimento de busca das unidades idealizado por Hunt e Black [9], que generaliza este processo para um grande corpus de unidades.

O capítulo seguinte descreve o método de busca incorporado ao sistema SASPRO no âmbito deste projeto de graduação, baseado apenas no conceito de custo fonético-prosódico.

Capítulo 4

Contribuições ao Sistema

4.1 Introdução

A partir das considerações feitas no capítulo anterior sobre o método de seleção automática de unidades, no presente capítulo serão apresentadas as contribuições ao SASPRO, principalmente o acréscimo de novas informações úteis ao procedimento de busca. Desta forma, almeja-se um sinal de fala sintetizado mais natural e inteligível, o mais próximo possível da fala humana.

Neste ponto, vale ressaltar que foram necessários uma grande familiarização com o sistema SASPRO e suas ferramentas, bem como um extenso estudo para um melhor entendimento dos métodos como foram implementadas tais ferramentas, usando como fonte principal a própria tese de doutorado de Vagner Latsch [3].

A Seção 4.2 descreve o modo adotado no contexto do presente projeto de graduação para se obter um banco único e o mais completo possível.

Na Seção 4.3, é proposto um novo método de síntese a partir da busca em um arquivo de definições de unidades, enquanto que na Seção 4.4 é descrito o método utilizado na montagem automática deste arquivo.

Por fim, na Seção 4.5 são propostos e comparados diferentes métodos de busca de unidades de síntese por seleção de parâmetros fonéticos-prosódicos.

4.2 Banco Único de Unidades

Conforme já foi citado na Seção 3.2.4, o banco de unidades do sistema SASPRO original era formado a partir de unidades inseridas em logotomas, selecionadas manualmente e incluídas em um arquivo de definição. Uma vez definidas as unidades, a construção do banco era feita a partir deste arquivo de definições, que continha, em cada linha, um código de identificação da unidade, o nome do arquivo onde a unidade se encontra e os instantes de tempo que delimitam a unidade no sinal. Estas informações de definição das unidades eram armazenadas em uma estrutura de dados chamada TUnitDef e o arquivo de definições era denominado de "unitsdef.txt".

Quando um arquivo de texto era aberto na lista de frases, a sua primeira linha continha o diretório de armazenamento dos arquivos e nas linhas seguintes as sentenças, iniciadas pelo nome atribuído aos arquivos. Então, tal como estava implementado, o arquivo de definição possuía apenas o nome do arquivo da unidade, sem incluir o diretório de tal arquivo, o que não possibilitava a leitura de unidades em diferentes diretórios, como representado na Figura 4.1. Este fato dificultava a inserção de novas unidades que poderiam ser utilizadas na ocasião da síntese.

Como o método de seleção automática das unidades se baseia na geração de voz a partir de um grande *corpus*, uma primeira contribuição ao sistema SASPRO foi a adoção de um arquivo de definição único, que pudesse ler todas unidades, independente do diretório/arquivo em que elas estivessem inseridas ou do arquivo que estivesse aberto na lista de frases. Assim, ao arquivo de definição foi acrescentado não apenas o nome do arquivo com a unidade desejada, mas também o diretório completo de localização desta unidade, possibilitando a leitura e escrita de qualquer unidade pelo sistema sempre que necessário, como indicado na Figura 4.2.

4.3 Novo Método de Síntese

A síntese original do sistema SASPRO seguia a técnica concatenativa, em que, de maneira geral, as unidades eram armazenadas em um banco de unidades. Na ocasião da síntese, estas unidades eram trazidas do banco e concatenadas, pro-

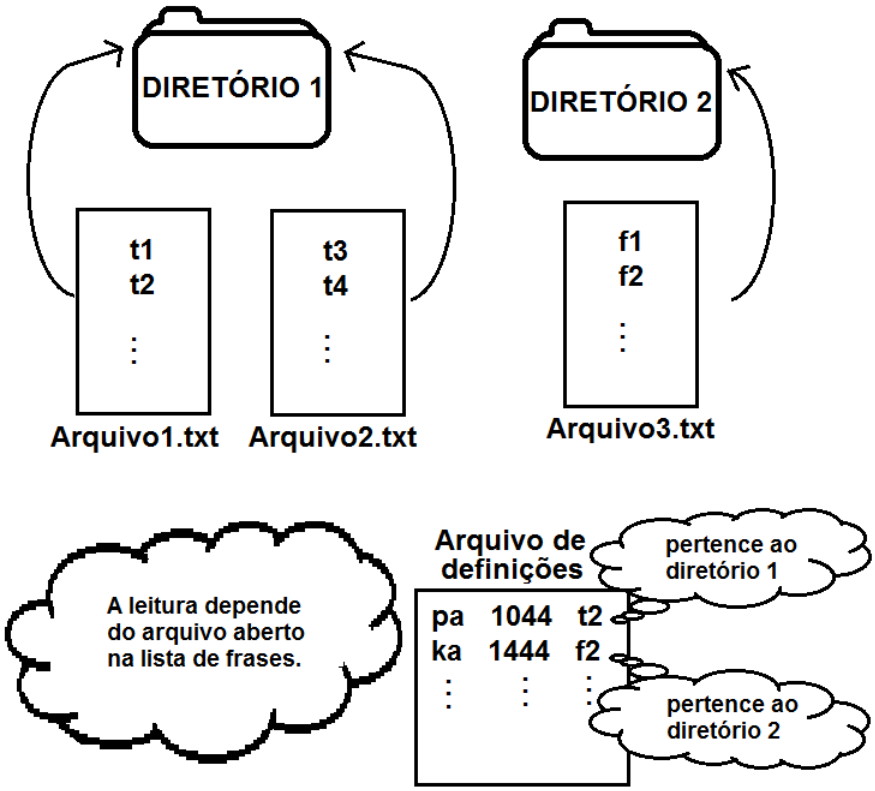


Figura 4.1: Processo antigo de leitura das unidades definidas no arquivo de descrição do banco.

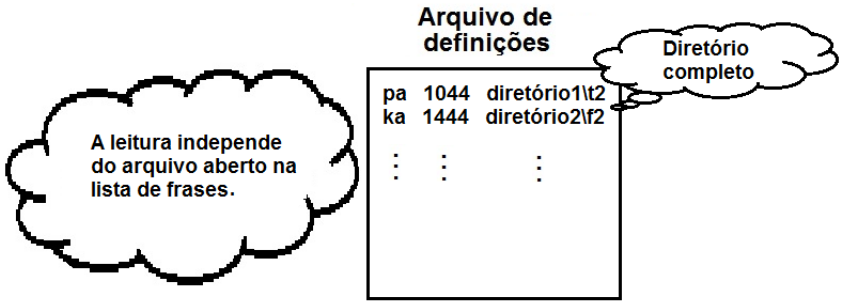


Figura 4.2: Processo atual de leitura das unidades definidas no arquivo de definição.

duzindo, desta maneira, o sinal de voz sintético.

Neste trabalho, é proposta uma nova metodologia, em que as unidades são obtidas diretamente de frases gravadas naturalmente, frases estas que podem, assim, ser vistas como o próprio banco de unidades. Neste novo método, no momento da síntese, a unidade é extraída ou recortada diretamente da frase a que pertence para compor a concatenação do sinal sintetizado. A grande dificuldade desta nova

proposta é que o banco se torna muito maior, pois ele passa a ser composto por todas as frases disponíveis. Com isto, espera-se que o tempo para realizar a síntese aumente bastante, já que a quantidade de unidades a serem consideradas no processo de busca também aumenta consideravelmente. Por outro lado, com esta nova proposta, passa-se a ter diversas versões de uma mesma unidade, com diferentes contextos prosódicos, o que aumenta a possibilidade de encontrar uma versão da unidade que melhor se ajuste às características desejadas para a síntese, provavelmente melhorando a qualidade da fala sintetizada.

Para simplificar o processo de busca, todas as unidades presentes no *corpus* de frases são inicialmente mapeadas num arquivo de definições de unidades, e a busca é realizada ao longo do conteúdo deste arquivo. Neste arquivo, conforme já foi explicado na Seção 4.2, são inseridas as seguintes informações necessárias para encontrar as unidades: um código de identificação da unidade, o diretório completo de onde a unidade se encontra e os instantes de tempo que delimitam a unidade no sinal gravado.

Desta forma, a busca é feita em cima das informações contidas neste arquivo de definições de unidades e, no processo de síntese, a unidade selecionada é extraída diretamente da frase a que pertence, segundo o mesmo arquivo de definições, como representado na Figura 4.3.

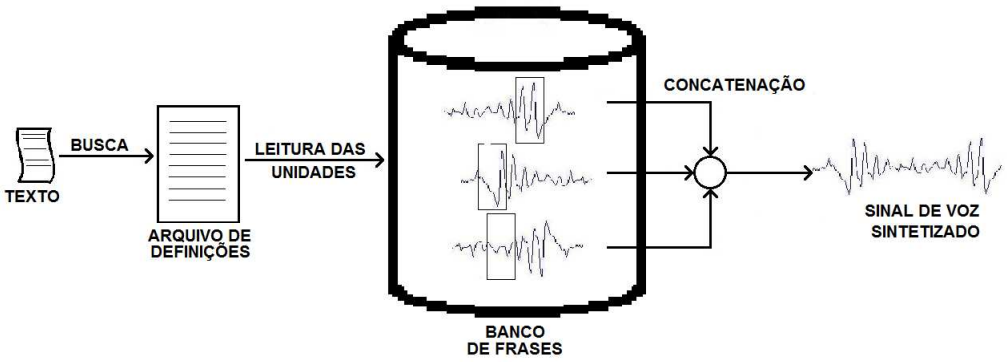


Figura 4.3: Novo procedimento de síntese adotado no sistema SASPRO.

Para realizar este procedimento, tal como era feito anteriormente, também é necessário que as frases gravadas contidas no banco estejam devidamente anotadas, com as marcas de *pitch* inseridas e a etiquetagem fonética, delimitando as fronteiras

de cada unidade, concluída. Vale ressaltar que todos os procedimentos de gravação, recorte, visualização da amplitude dos sinais e do espectro, assim como a edição das marcas de pitch e das etiquetas do conjunto de frases, podem ser realizados com o auxílio das ferramentas de edição pertencentes ao SASPRO, como descrito em [3]. Ao término destes procedimentos, para cada sentença etiquetada, são gerados quatro arquivos: um arquivo (.wav) estéreo, contendo o sinal de fala propriamente dito e o do eletroglotógrafo (EGG) ou do microfone de contato; um arquivo (.mrk) contendo os instantes de tempo das marcas de *pitch* extraídas a partir do sinal EGG; um arquivo (.lmk) contendo os instantes de tempo das etiquetas de fronteira dos segmentos fonéticos contidos no sinal de voz; e um arquivo (.phn) contendo o resultado do processamento de texto. Estes arquivos contêm toda a informação necessária para a posterior análise das sentenças. Em termos de programação, os dados contidos nestes arquivos correspondem aos dados encapsulados pelas classes CWave, CMarks, CLabelMarks e CSentence, respectivamente, e pertencentes a implementação inicial do sistema SASPRO [3].

Neste trabalho, para esta nova metodologia de síntese, foi utilizado um *corpus* contendo 200 sentenças, cujos processos de gravação, inserção das marcas de *pitch* e etiquetagem das fronteiras fonéticas foram realizados utilizando-se as ferramentas do sistema SASPRO conforme descrito em [3].

4.3.1 Implementação

Na nova versão do sistema SASPRO, as unidades ainda são encapsuladas pela classe CUnit, com a diferença de que no novo método de síntese, não é mais necessário que sejam inseridas, removidas ou buscadas no banco em si, já que elas, as unidades, são formadas diretamente a partir da leitura das frases, conforme descrito a seguir.

A classe CUnitData é responsável pela mapeamento das unidades no arquivo de definições, indicando a exata posição das mesmas no banco de frases. Desta forma, a essa classe, foi acrescentado um método para, no momento da síntese, realizar a busca pela unidade desejada no próprio arquivo de definições. Uma vez de posse da localização da unidade desejada, isto é, do arquivo de frase a que pertence

e os instantes de tempo que delimitam a unidade no sinal, foi acrescentado outro método para realizar a leitura das unidades diretamente das frases, encapsulando-as na classe CUnit.

Por fim, na etapa final da síntese, são retornadas as amostras do sinal concatenado, as marcas de pitch e as marcas de fronteiras fonéticas, tal como já era feito. Assim, a síntese pode ser realizada buscando as unidades no arquivo de definições e lendo-as diretamente da frase a que pertencem.

4.4 Automação do Arquivo de Definições

Na proposta original do SASPRO, o arquivo de definições era gerado manualmente, após o recorte das unidades uma a uma. Na nova proposta, porém, pelo grande número de unidades (alguns milhares), é interessante que o arquivo de definições das unidades seja definido automaticamente. A seguir, será descrito o procedimento adotado para a definição automática deste arquivo, de maneira a atender os seguintes objetivos:

- Obter de maneira dinâmica todas as possíveis combinações de unidades de síntese contidas em cada sentença;
- Eliminar a necessidade de indicar os pontos de recorte das unidades no arquivo de definições;
- Facilitar o acréscimo de novos atributos da unidade, relevantes ao procedimento de busca.

Diante do exposto, neste estágio se tornava importante escolher o tipo (tamanho) das unidades síntese. Levando em consideração o levantamento bibliográfico descrito na Seção 3.3, e seguindo a tendência observada nos principais sistemas de síntese atuais, optou-se por utilizar um misto das unidades dos tipos difones e trifones.

Desta forma, por exemplo, para a palavra “pato”, cuja a transcrição fonética é dada por “_pa-tU_”, têm-se todas as possíveis formações de difones e trifones, que devem estar indicadas no arquivo de definições, mostradas na Tabela 4.4, onde

os símbolos "_" e "-" indicam, respectivamente, intervalo de silêncio e início de consoante plosiva não vozeada.

Tabela 4.1: Combinações possíveis de difones e trifones para a palavra "pato".

Difones	**Trifones**
_p	_pa
pa	pa-
a-	a-t
-t	-tU
tU	tU_

Outro procedimento adotado na nova versão do sistema SASPRO foi a inclusão, no arquivo de definições, das etiquetas de fronteiras que delimitam os fones das extremidades das unidades. Desta forma, foi possível eliminar a necessidade de indicar o ponto de recorte das unidades, ou o ponto de concatenação, ponto este que passou a ser determinado no momento da execução do procedimento de síntese. Assim, o arquivo de definições torna-se somente um arquivo que mapeia as unidades indicando as que estão disponíveis.

4.4.1 Implementação

Para realizar o procedimento de criação do arquivo de definições automaticamente foi adicionado um método à classe CUnitdata. Esse novo método é responsável por realizar um *loop* interno a uma dada sentença e encontrar todas as possíveis combinações de unidades aí contidas, bem como a indicação de suas respectivas fronteiras fonéticas. Este *loop* é baseado nos fones, que são encapsulados pela classe já implementada CPhone, que representa cada fone através de um código e símbolo de identificação.

Com isto, em cada sentença do *corpus* de frases, o método implementado lê o sinal correspondente e a respectiva transcrição fonética. Em seguida, são realizadas as atribuições da classe CLabelMarks, referentes às etiquetas de fronteira de cada fone. A leitura das sentenças foi realizada a partir de métodos já existentes na classe

CSentence, assim como as atribuições na classe CLbmarks, em que o método faz uso das informações contidas no arquivo onde estão os respectivos instantes de tempo das etiquetas de fronteira de cada fone.

O próximo passo é efetuar todas as combinações possíveis de dois ou três fones, tal como exemplificado na Tabela 4.4. As extremidadades das unidades serão as fronteiras mais externas do fones contidos nela. Estes pontos apenas indicam a posição da unidade, não o ponto de concatenação das mesmas. Uma vez realizadas todas as possíveis combinações, o arquivo de definições é salvo automaticamente.

Tomando novamente a palavra "pato" como exemplo, o procedimento adotado é ilustrado na Figura 4.4, em que se pode ver a transcrição fonética da palavra acompanhada de valores fictícios dos instantes de tempo das fronteiras de cada fone. Em destaque nesta figura, têm-se um exemplo do difone "pa" e outro do trifone "tU_", com as setas indicando as extremidades das unidades e as linhas tracejadas o ponto de recorte das unidades.

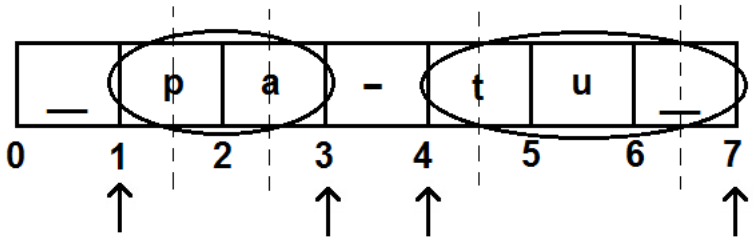


Figura 4.4: Indicação das fornteiras das unidades da palavra "pato".

Na nova versão do sistema SASPRO, o ponto de recorte das unidades passou a ser encontrado em tempo de execução, de modo a minimizar a função custo de concatenção citada na Seção 3.3.2. Este processo de busca percorre os sinais dos fones que irão ser concatenados e compara as características espectrais das unidades selecionadas para escolher o ponto de maior similaridade entre eles. Aqui, por simplificação, este ponto de recorte foi definido como sendo o ponto médio entre as duas marcas de fronteiras que definem os fones mais extremos da unidade. O método da classe CUnitData, implementado neste trabalho para realizar a leitura das unidades diretamente a partir das frases, passou a determinar o ponto de concatenação conforme descrito.

O arquivo de definições foi implementado como um *container* associativo

multimap petencente à biblioteca STL da linguagem C++. Genericamente, um *container* é uma espécie de caixa onde objetos do mesmo tipo são armazenados e organizados [13]. Um *container* associativo utiliza um conceito de chave (no caso, o código de cada unidade) atribuída a cada elemento armazenado. Dentre outras coisas, um *container* possui funções-membro para inserção, busca e remoção dos elementos, funções estas utilizadas na formação e na manipulação do arquivo de definições das unidades.

Dentre as vantagens da utilização da estrutura de *multimap* neste trabalho, está o fato de este tipo de *container* permitir o armazenamento ordenado dos elementos com chaves repetidas. Isso é de fato necessário, pois há a possibilidade (e até mesmo a necessidade) de termos mais de uma versão da mesma unidade. Além disto, é possível explorar o fato de que neste tipo de estrutura a função-membro de busca realiza este procedimento com uma complexidade logarítmica ($\log(n)$), que é mais eficiente que a busca linear, por exemplo.

A ferramenta que permite a indicação das unidades no arquivo de definições, que estão contidas em uma sentença, foi adicionada ao SASPRO, de modo que este procedimento pode ser realizado a partir de uma sentença individual ou a partir de todo o conjunto de frases que estiverem presentes na lista de frases. Neste sentido, a Figura 4.5 mostra os submenus adicionados ao sistema SASPRO para se gerar o arquivo de definições automático.

4.4.2 Resultados Parciais

Na versão original do SASPRO, o banco de unidades era recortado manualmente a partir de logotomas, possuindo um total de cerca de 500 unidades compondo um banco de aproximadamente 500 MB.

Com a nova metodologia automática de criação do arquivo de descrições de unidades, a partir do *corpus* de 200 frases, gerou-se um arquivo de definições de 852 KB contendo um total de cerca de 11 mil unidades, sendo que 3,5 mil com conteúdo fonético distinto. Dessas 3,5 mil unidades fonéticas, aproximadamente 2 mil continham apenas uma única realização, devido ao limitado tamanho do *corpus* inicial. As 30 unidades com maior número de realizações são mostradas na Figura 4.6.

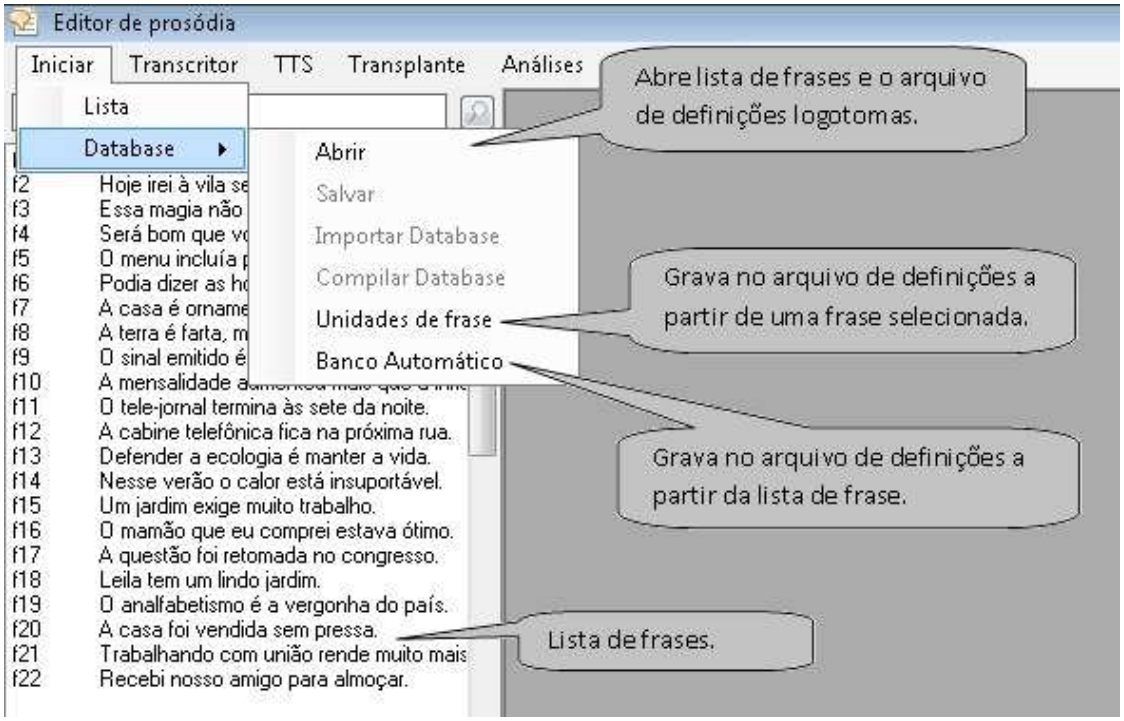


Figura 4.5: Tela do sistema SASPRO no modo de geração automática do arquivo de definições.

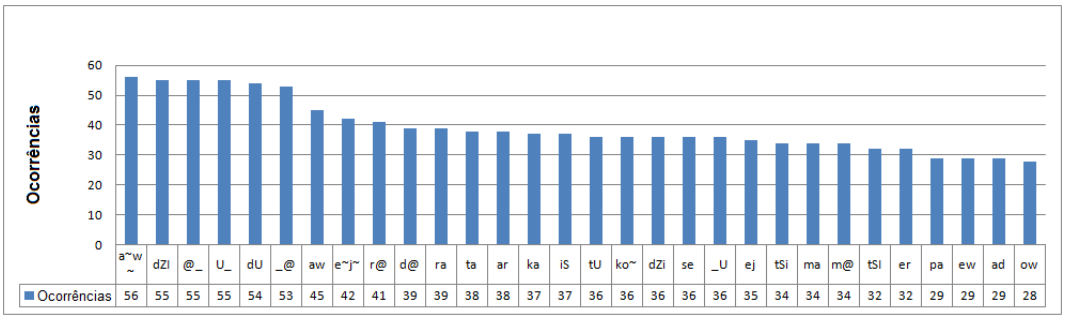


Figura 4.6: Unidades de maior ocorrência no novo arquivo de definições do sistema SASPRO.

O arquivo de definições obtido foi utilizado no procedimento de síntese proposto na Seção 4.3.

Como neste arquivo têm-se várias versões da mesma unidade, num primeiro experimento, a síntese foi realizada de maneira a concatenar a primeira unidade foneticamente compatível encontrada.

Num segundo experimento, outras características das unidades, como a tonicidade sílaba, por exemplo, foram consideradas para diferenciar as unidades no

processo de busca para a síntese.

Os resultados dos sinais obtidos nesses dois experimentos serão apresentados no Capítulo 5, numa comparação com as outras etapas implementadas.

4.5 Busca por Seleção de Unidades

O método de seleção foi proposto inicialmente para que houvesse o mínimo possível de manipulação prosódica, pois é neste processo que ocorre a maior perda da naturalidade na geração de voz sintetizada. No caso, a manipulação prosódica se refere a alterações dos parâmetros de *pitch*, duração e intensidade. Assim, para minimizar as alterações destes parâmetros, o melhor é que as unidades utilizadas na concatenação já tenham estes parâmetros os mais próximos do desejado.

Neste sentido, após a transcrição do texto, pode-se implementar uma etapa de estimação prosódica para determinar os valores prosódicos desejados para uma dada unidade fonética. Esta modelagem da prosódia utiliza os parâmetros do texto, tais como a posição da sílaba da sentença, tonicidade da unidade, posição da sílaba na palavra, pontuação etc.

Uma idéia alternativa, porém, considera a busca das unidades diretamente a partir dessas informações do texto, sem passar pela etapa intermediária de modelagem de prosódia.

Com isto, podemos vislumbrar duas outras versões para a busca de unidades no novo sistema SASPRO: uma busca que considera os parâmetros prosódicos de *pitch*, duração e intensidade, gerados por um módulo adicional de modelagem prosódica, e outra que considera apenas as informações das unidades relativas ao texto de entrada.

Neste trabalho, procuramos implementar estes dois modos de busca, como descrito nas subseções a seguir. Inicialmente, o segundo modo foi implementado considerando apenas a característica de tonicidade da unidade em questão. Mesmo para esta versão simplificada, percebeu-se a necessidade de uma grande variabilidade de realizações das unidades indicadas no banco, o que não era o caso de nosso sistema que possui apenas um *corpus* de 200 frases. Desta forma, um sistema

híbrido foi proposto, considerando também a busca pelos parâmetros prosódicos diretamente, caso a busca textual se mostrasse insatisfatória. Para emular o módulo de modelagem prosódica, cujo desenvolvimento foge ao escopo do presente projeto, utilizou-se uma locução natural da frase de entrada da qual os parâmetros prosódicos desejados eram copiados.

4.5.1 Busca por Parâmetro (tonicidade) do Texto

Neste processo, consideramos a busca de unidades, levando em consideração a tonicidade da unidade na respectiva palavra. Assim, o passo inicial foi acrescentar o parâmetro tonicidade ao arquivo de definições, o que foi feito através da estrutura de dados TUnitdef, que armazena as informações de definições das unidades.

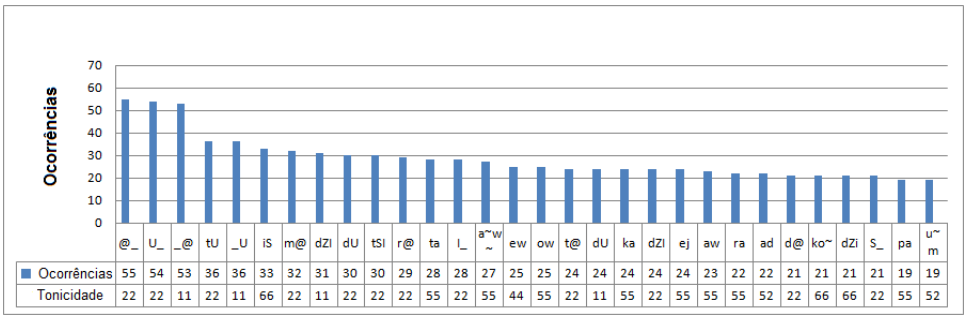
Assim, após as atribuições da classe CLabelMarks, é possível realizar o *loop* interno às sentenças através dos fones. O valor da tonicidade é atribuído no processamento do texto, sendo conferido à cada sílaba, que por sua vez está encapsulada pela classe CSilaba. Portanto, para ter acesso à informação de tonicidade de cada fone, foi necessário consultar a classe CSilaba pela classe CPhone com o método *parent*.

Um problema encontrado neste novo processo foi a atribuição da tonicidade às unidades que contêm fones pertencentes a sílabas diferentes. Na palavra "_pa-to_, por exemplo, a sílaba "_pa" possui um valor de tonicidade e a sílaba "-to_" outro. Então, a questão era como atribuir a tonicidade à unidade "a-t", por exemplo. A solução adotada foi armazenar a tonicidade a cada fone pertencente à unidade indicada, de modo que a tonicidade de cada unidade passa a ser vista como um vetor de duas componentes, para o caso dos difones, ou três, para os trifones.

No processo de conversão texto-fala, sempre que há uma entrada de texto, é realizada a transcrição fonética no nível dos fones, e é atribuída a tonicidade a cada sílaba e por consequência a cada fone. Com isto, tornou-se possível considerar a tonicidade de cada unidade no processo de busca para a síntese. Baseado nos conceitos descritos em [9], como citado na Seção 3.3.2, foi necessário estabelecer uma medida de distância entre unidades para o processo de busca. Neste caso, a distância utilizada foi a euclidiana, calculada a partir do vetor tonicidade da unidade

alvo, obtido na transcrição fonética, e do vetor tonicidade da unidade disponível no banco.

Conforme já foi dito, a partir do *corpus* de 200 frases, cerca de 11 mil unidades foram mapeadas no arquivo de definições. Com o acréscimo do parâmetro tonicidade, o tamanho do arquivo passou a ser de 890 KB. A quantidade de unidades distintas, com respeito à definição e à tonicidade passou a 5,4 mil, e a quantidade de unidades com uma única realização passou a cerca de 4 mil. O gráfico apresentado na Figura 4.7 mostra as 30 unidades de maior ocorrência com respeito no arquivo de definições e tonicidade.



	@_	U_	_@	tU	_U	iS	m@	dZI	dU	tSI	r@	ta	I_	a~w~	ew	ow	t@	dU	ka	dZI	ej	aw	ra	ad	d@	ko~	dZi	S_	pa	u~m
Ocorrências	55	54	53	36	36	33	32	31	30	30	29	28	28	27	25	25	24	24	24	24	24	23	22	22	21	21	21	21	19	19
Tonicidade	22	22	11	22	11	66	22	11	22	22	22	55	22	55	44	55	22	11	55	22	55	55	55	52	22	66	66	22	55	52

Figura 4.7: Unidades de maior ocorrência no arquivo de definições considerando também o aspecto de tonicidade.

Os resultados deste procedimento de síntese com busca dinâmica e seleção de tonicidade também serão apresentados no Capítulo 5, numa comparação com as demais versões implementadas para o método de busca.

4.5.2 Busca por Parâmetros de Prosódia

Quando incorporamos o aspecto de tonicidade de uma unidade ao processo de busca, devido à pouca variabilidade do banco de unidades utilizado, verificou-se ser muito comum a dificuldade de se encontrar uma unidade adequada. Assim, para compensar este efeito, nos casos críticos, usou-se a busca de unidades diretamente a partir dos parâmetro prosódicos, considerando-se a existência de um módulo adicional de modelagem prosódica.

Neste método alternativo de busca, assim como no caso da tonicidade, o passo inicial foi acrescentar os novos parâmetros (no caso, os prosódicos) ao arquivo de

definições. Para ficar consistente com a busca por tonicidade, optou-se por utilizar os parâmetros prosódicos no nível da sílaba. Com isto, após a etapa de síntese, em que os parâmetros prosódicos da sílaba são repassados aos fones, usando os modelos descritos na Seção 3.2. Neste caso, após a leitura da sentença, são relizadas as atribuições das classes CLablMarks, CWave e CaMarks, que a priori, estão envolvidas na etapa de processamento prosódico, com métodos já implementados. O vetor de parâmetros que codifica as variáveis prosódicas de duração, intensidade e *pitch*, no nível da sílaba, também é encapsulado pela classe CSilaba. E, tal como anteriormente, para se ter acesso aos parâmetros prosódicos de cada fone, é necessário consultar a classe CSilaba a partir da classe CPhone com o método *parent*. Desta forma, à estrutura TUnitdef foi acrescentado um vetor para armazenar os parâmetros prosódicos. No caso do *pitch*, que é representado por quatro pontos, optou-se por armazenar o seu valor médio, de modo que ao invés de sete parâmetros prosódicos por unidade são armazenados apenas quatro.

Para realizar a busca das unidades de síntese utilizando os parâmetros prosódicos, é necessário que os parâmetros desejados sejam conhecidos a priori. Em sistemas práticos, estes parâmetros podem ser estimados, por exemplo, a partir de uma rede neural previamente treinada. Entretanto, como o foco deste trabalho é o estudo do processo de seleção das unidades, optou-se neste caso por copiar os parâmetros prosódicos de uma frase gravada naturalmente, postergando a etapa de modelagem prosódica para um trabalho complementar.

Para a etapa de síntese, com o transplante de prosódia a partir de uma frase natural, foi criado um novo método para obter os parâmetros prosódicos de cada fone do sinal gravado. Estes dados são armazenados em um vetor da classe CUnitdata e utilizados no método desta classe que é responsável pela busca das unidades no arquivo de definições. Desta forma, tem-se um vetor com os parâmetros prosódicos desejados para cada unidade definida na etapa de busca.

Neste etapa, também foi necessário definir uma métrica, baseada nos parâmetros prosódicos, para a distância entre as unidades-alvo e as unidade-candidatas. A medida utilizada foi a distância euclidiana, e a medida da distância total $D(u_i, t_i)$ seria a soma das distâncias da tonicidade, $D^t(u_i, t_i)$, e dos parâmetros prosódicos,

$D^p(u_i, t_i)$, isto é

$$D(u_i, t_i) = D^t(u_i, t_i) + D^p(u_i, t_i). \tag{4.1}$$

Deste modo, a unidade de síntese escolhida pelo processo de busca é aquela que, dentre as unidades-candidatas do banco, minimiza esta distância total.

O arquivo de definições automático obtido, que inclui a tonicidade e os parâmetros prosódicos de *pitch* médio, duração e intensidade de cada uma das 11 mil unidades definidas, atingiu o tamanho aproximado de 1,3 MB, e o seu tempo de gravação para o nosso *corpus* de 200 sinais foi de 58 s.

O método de busca mista, considerando os aspectos de tonicidade e prosódicos, foram incorporados ao sistema SASPRO, como indicado na Figura 4.8, e podem ser acionados através dos botões "concat new", referente à seleção apenas por tonicidade, e "prosódia", referente à seleção pelos parâmetros prosódicos e tonicidade. Os resultados deste procedimento de síntese com busca dinâmica e seleção de tonicidade e parâmetros prosódicos serão apresentados no Capítulo 5.

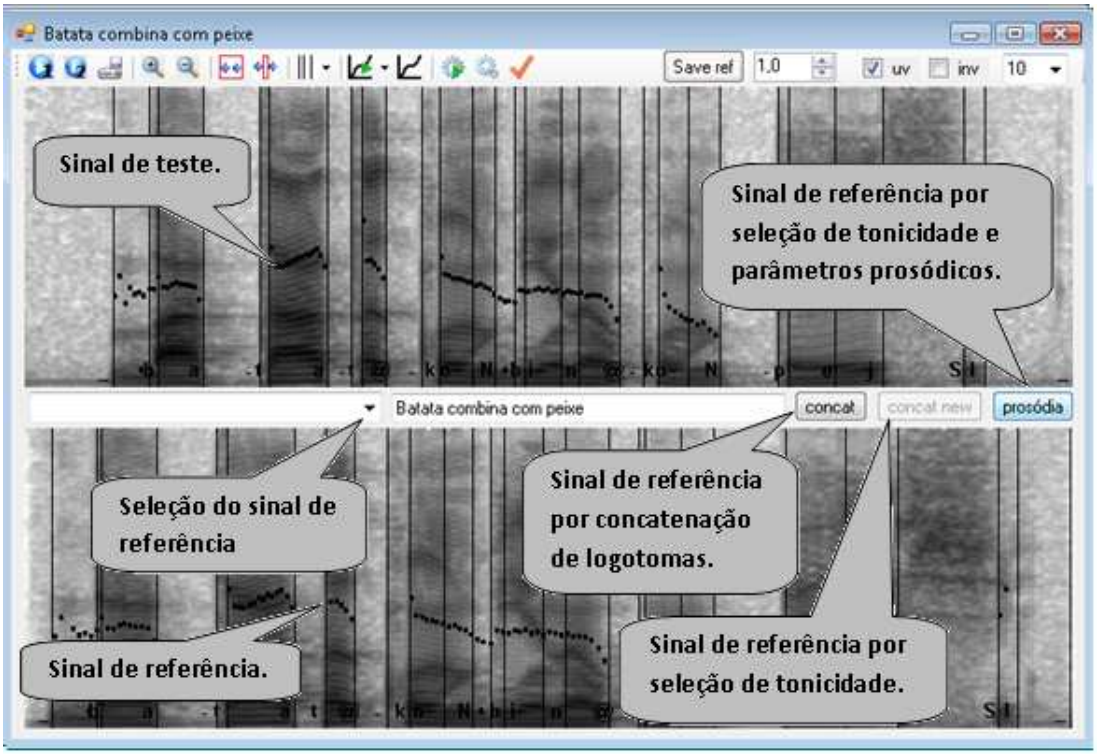


Figura 4.8: Tela atualizada de transplante de prosódia do sistema SASPRO incluindo as novas funcionalidades desenvolvidas no âmbito deste projeto de graduação.

4.6 Conclusão

Este capítulo apresentou as principais contribuições deste trabalho ao sistema SASPRO. Foram descritas as ações adotadas e a maneira como foram implementadas, bem como alguns resultados quantitativos parciais acerca do banco de unidades resultante em cada caso.

Assim, foi descrito o novo modo de leitura das unidades no sistema SASPRO, bem como o novo método de síntese do sistema, que passou a utilizar um banco contínuo de frases em vez de um banco de unidades. Além disso, foi descrito o procedimento para realizar a busca das unidades de síntese no arquivo de definições e como este arquivo pode ser obtido automaticamente.

Por fim, foram propostos diferentes métodos de busca das unidades por seleção de parâmetros fonético-prosódicos.

O capítulo seguinte apresenta os resultados obtidos, mostrando um comparativo das etapas intermediárias até a etapa final.

Capítulo 5

Resultados Obtidos

5.1 Introdução

Neste capítulo, procuramos avaliar os resultados obtidos pela concatenação de unidades que foram buscadas utilizando o método de seleção. O que se espera é que, com a seleção das unidades, as frases sintetizadas pareçam mais naturais. O conceito de naturalidade, neste contexto, refere-se à fala sintetizada ser a mais próxima possível da fala humana. Assim, realizar uma avaliação precisa deste conceito é bastante complexo, tendo em vista a subjetividade do processo envolvido.

Aqui é apresentada uma avaliação comparativa dos sinais sintetizados com as novas técnicas de síntese e, em seguida, incluímos os resultados de um teste subjetivo de qualidade.

Nas análises que se seguem, o sistema de concatenação original será denominado de Sistema 1, enquanto que o modo de busca dinâmica no arquivo de definições, mas sem seleção das unidades, será denominado de Sistema 2. Além disto, o sistema com busca dinâmica e seleção da tonicidade foi denominado Sistema 3, e o sistema com busca dinâmica e seleção de tonicidade e parâmetros prosódicos de Sistema 4.

5.2 Avaliação Comparativa

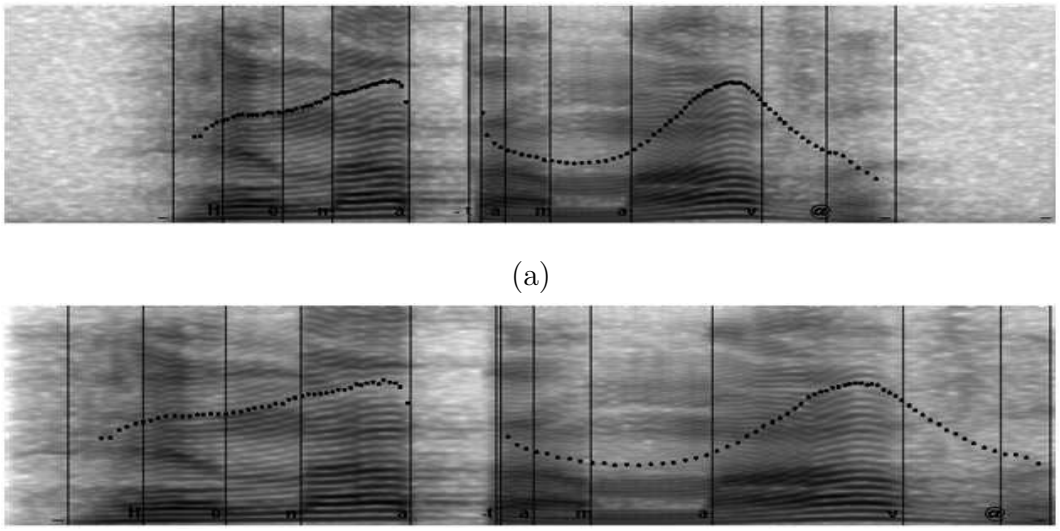
Tendo em vista que no Sistema 1 a síntese foi realizada a partir de logotomas, a duração, energia e pitch são mais constantes e o sinal gerado por vezes

parece artificial. Já no Sistema 2 implementado, o que foi observado é que os sinais resultantes apresentam um comportamento não muito estável, oscilando entre falas sintetizadas com um alto grau de naturalidade ou com algumas distorções proeminentes. Isso, possivelmente, ocorre porque nesse sistema a síntese é realizada com unidades que foram buscadas sem qualquer tipo de seleção. Ou seja, a primeira unidade compatível encontrada é utilizada. Assim, como as primeiras unidades do banco pertencem às primeiras frases. Se uma frase de entrada for semelhante a uma das primeiras frases do banco, o sinal gerado apresenta maior naturalidade. Por outro lado, se as frases de entrada forem muito diferentes das frases iniciais, o resultado parece distorcido.

No Sistema 3, foi acrescentado um método de seleção das unidades. Neste caso, percebeu-se, que para serem observados resultados mais significativos, a variabilidade do banco de frases utilizado precisa ser muito grande, o que não é o caso do banco disponível. Entretanto, percebeu-se que, sempre que possível, o sistema selecionou a unidade com tonicidade mais próxima da desejada e algumas melhoras na naturalidade dos sinais obtidos foram claramente observadas. Por exemplo, a palavra "cocaína" quando sintetizada com os Sistemas 1 e 2, o som se parece com "cocáina", com marcação de tonicidade na sílaba "ca", enquanto que no sistema Sistema 3 a palavra foi sintetizada com a tonicidade correta. Outro exemplo é a frase "Meninos hábeis", em que os Sistemas 1 e Sistema 2 geraram uma pronúncia como se a sílaba tônica da palavra "hábeis" fosse a última, enquanto o Sistema 3 gerou a tonicidade correta.

O Sistema 4 faz a seleção das unidades utilizando o parâmetro de tonicidade e os parâmetros prosódicos de duração, intensidade e *pitch*, onde esses parâmetros foram copiados de uma frase gravada naturalmente. Neste caso, para uma unidade ser escolhida ela deve atender também aos requisitos prosódicos. Com isto, constatou-se informalmente que os sinais obtidos com esse sistema foram melhores, sendo possível perceber as nuances das frases contidas no banco, que demonstravam alguma postura da fala, tais como questionamento, exclamação, raiva, etc. Além disso, conforme esperado pelo método de seleção das unidades, foi possível perceber que as frases sintetizadas a partir desse sistema necessitaram que as alterações prosódicas

pelo método de PSOLA fossem minimizadas. Isso pode ser obervado na comparação do espectograma da Figura 5.1a, relativo à frase "Renata Amava" com postura de questionamento gravada naturalmente, e do espectograma da Figura 5.1b, em que a mesma frase foi sintetizada a partir do Sistema 4, sem qualquer alteração subsequente pelo PSOLA. Nesta figura, observa-se claramente a similaridade dos dois espectogramas.



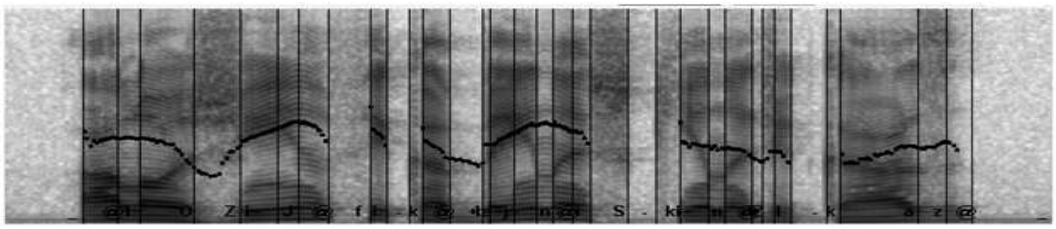
(a)

(b)

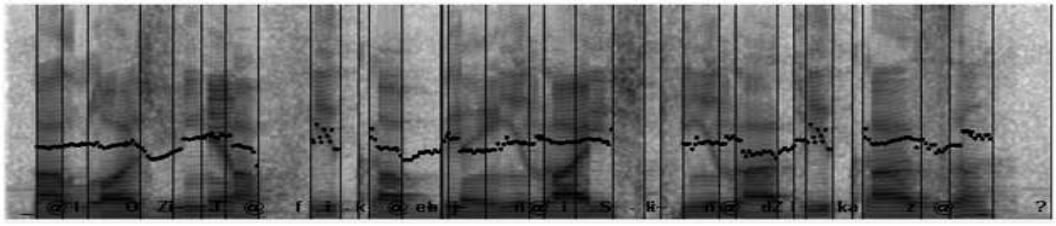
Figura 5.1: Espectograma da frase "Renata Amava" com postura de questionamento: (a) Gravação natural; (b) Resultado da síntese com o Sistema 4.

Para tornar a análise mais ampla, considere os espectogramas da frase "A lojinha fica bem na esquina de casa", que pertence ao banco original de frases, representados na Figura 5.2 com as seguintes variações: (a) Gravação natural da frase; (b) Sistema 1; (c) Sistema 2; (d) Sistema 3; (e) Sistema 4.

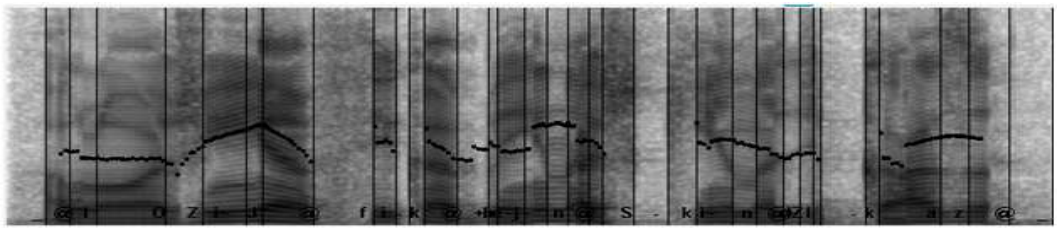
Nesta figura, é simples perceber o perfil de *pitch* quase constante para o sinal sintetizado pelo Sistema 1, resultado da gravação original dos logotomas com o mínimo de entonação possível. Além disto, para o sinal sintetizado pelo Sistema 2, podemos concluir que seria necessária uma manipulação prosódica adicional para tornar o perfil de *pitch* parecido com o do sinal original, o mesmo acontecendo para o sinal do Sistema 3. Finalmente, observando o sinal do Sistema 4, nota-se uma grande semelhança do contorno de *pitch* com o da frase natural, o que indica que o sistema de busca implementado selecionou corretamente as unidades para a etapa



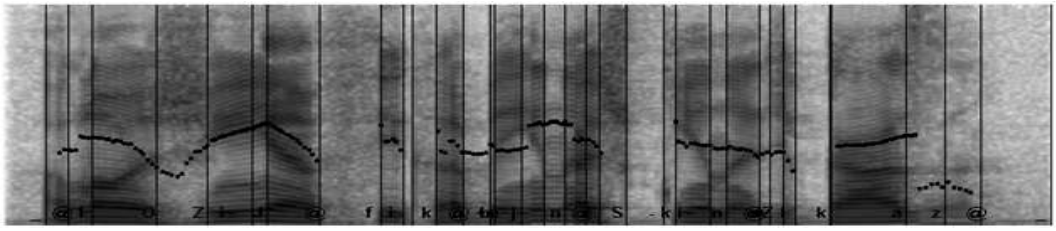
(a)



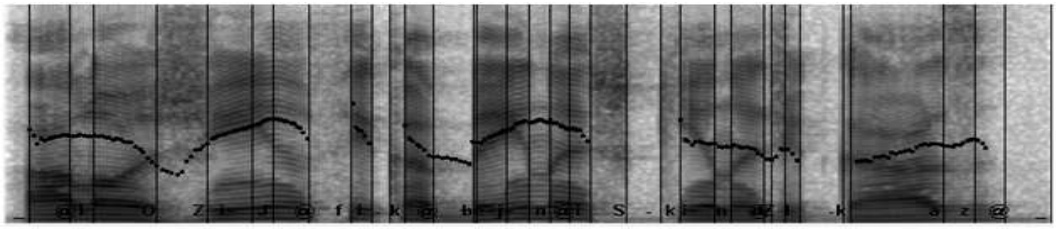
(b)



(c)



(d)



(e)

Figura 5.2: Espectogramas da frase “A lojinha fica bem na esquina de casa”: (a) Gravação natural; (b) Sintetizada pelo Sistema 1; (c) Sintetizada pelo Sistema 2; (d) Sintetizada pelo Sistema 3; (e) Sintetizada pelo Sistema 4.

de síntese.

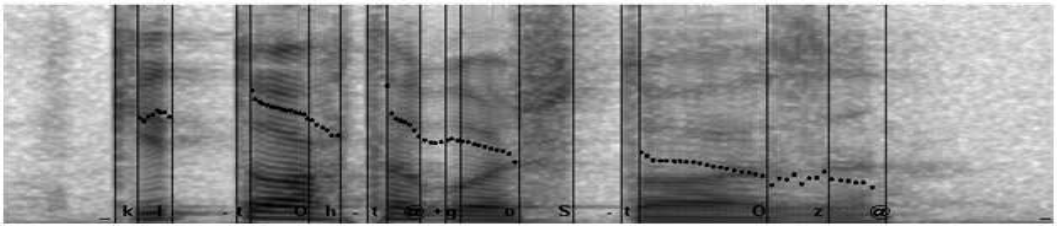
Em seguida, são mostrados os espectogramas da frase "Que torta gostosa!", que não pertence ao banco de frases, com postura exclamativa, como mostrado na Figura 5.2, seguindo a mesma ordem da Figura 5.2.

Como no caso anterior, o Sistema 1 gerou um sinal sem muita variação de *pitch*. Já o Sistema 2, gerou um sinal com algumas descontinuidades percebidas auditivamente, descontinuidades estas ausentes no sinal do Sistema 3, porém que ainda apresentam um contorno de *pitch* bem distinto do original. Por fim, novamente, temos um melhor desempenho do Sistema 4, embora desta vez com pequenas, porém perceptíveis, diferenças no contorno de *pitch*, o que pode ser explicado pela pequena variabilidade de algumas unidades no banco de frases disponível.

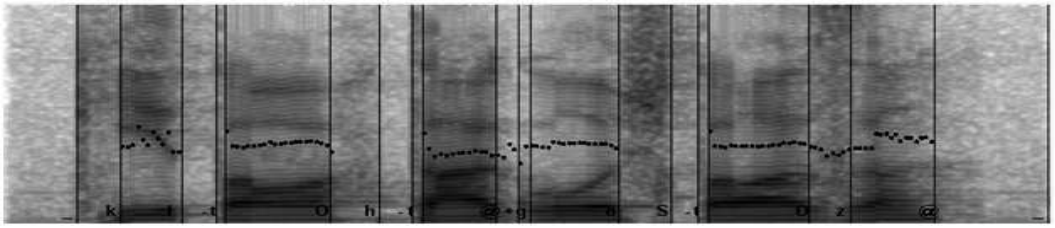
Por fim, na Tabela 5.1 é mostrado um comparativo dos arquivos de definições de unidades gerado a partir de cada sistema. A gravação deste arquivos foi feita através de ferramentas adicionadas ao SASPRO, rodando em um computador Intel Core 2 Duo 2,0 GHZ e 4,0 GB de memória RAM.

Tabela 5.1: Comparação dos arquivos de definições para cada sistema de busca.

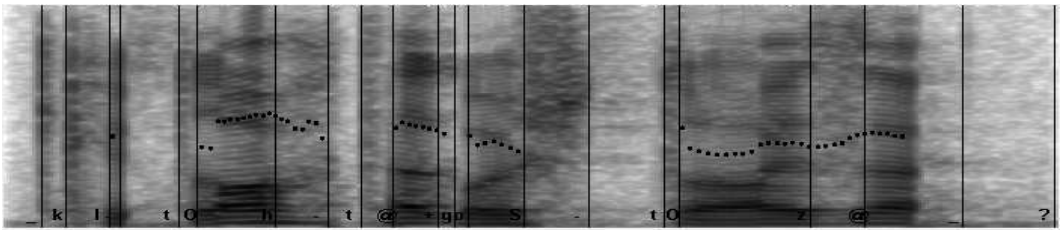
	Sistema 2	Sistema 3	Sistema 4
Tamanho do Arquivo [KB]	850	890	1500
Tempo de Gravação [s]	38	41	58
Número de Unidades Diferentes	3747	5420	11429
Número de Unidades com Uma Realização	1997	3867	0
Número Total de Unidades	11429	11429	11429



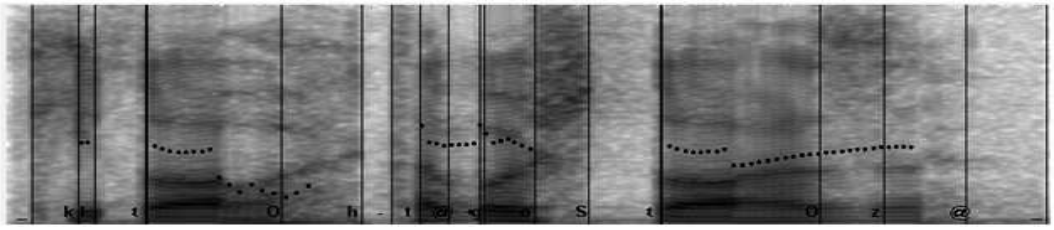
(a)



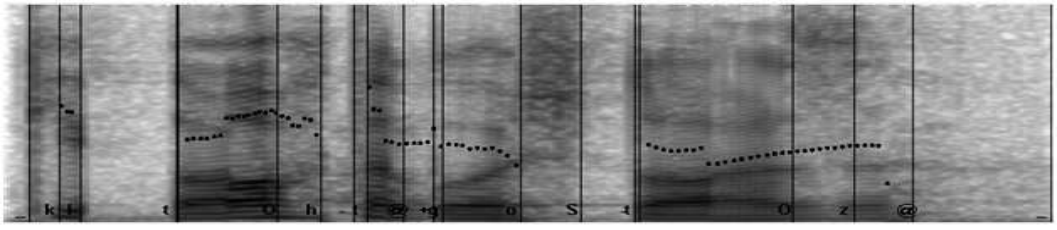
(b)



(c)



(d)



(e)

Figura 5.3: Espectogramas da frase “Que torta gostosa!”: (a) Gravação natural; (b) Sintetizada pelo Sistema 1; (c) Sintetizada pelo Sistema 2; (d) Sintetizada pelo Sistema 3; (e) Sintetizada pelo Sistema 4.

5.3 Avaliação Subjetiva

Para avaliar o desempenho dos sistemas implementados, foram realizados alguns testes subjetivos de qualidade da voz sintetizada. Esses testes consideraram a avaliação de 26 pessoas voluntárias, sob as mesma condições de ambiente e equipamentos. Todos os sistemas foram avaliados sob dois aspectos: inteligibilidade e naturalidade. As palavras e frases utilizadas no teste subjetivo estão listadas no Apêndice A.

5.3.1 Avaliação de Inteligibilidade

O conceito de inteligibilidade, nesse contexto, está relacionado com a capacidade de compreensão do conteúdo do que está sendo falado. Para este teste, o voluntário tinha a possibilidade de avaliar as palavras ditas pelos sistemas em três níveis: compreensível, parcialmente compreensível e incompreensível, aos quais foram atribuídas, respectivamente, as notas de '3', '2' e '1'.

A Figura 5.4 mostra a porcentagem das notas que avaliaram a inteligibilidade de cada sistema. Como pode ser observado nesta figura, na média, nenhum sistema foi considerado incompreensível. Além disso, pode ser visto que sob o aspecto de inteligibilidade, o Sistema 1 foi melhor avaliado que o Sistema 2, o que mostra que a falta de seleção da unidades no processo de concatenação pode gerar sinais de fala que parecem distorcidos para o ouvinte. No método de concatenação utilizado no Sistema 1, essas distorções são menores já que são utilizadas unidades gravadas em circunstâncias praticamente constantes.

5.3.2 Avaliação da Naturalidade

O conceito de naturalidade, conforme já foi dito, está relacionado a quão próximo da fala humana pode ser a fala sintetizada, incluindo aspectos de entonação.

Neste sentido, o teste da naturalidade foi realizado em dois níveis: no primeiro nível, os quatro sistemas foram avaliados com frases curtas; numa segunda fase, envolvendo frases mais longas, ao voluntário era solicitado decidir o sistema de sua preferência. Em ambas as fases, o voluntário avaliava as frases com uma nota que

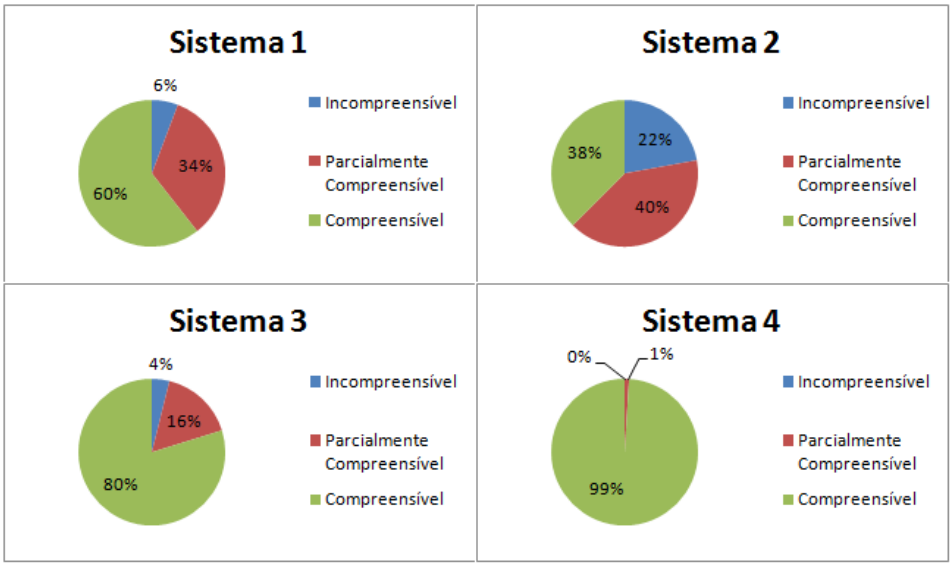


Figura 5.4: Notas da avaliação de inteligibilidade de cada sistema.

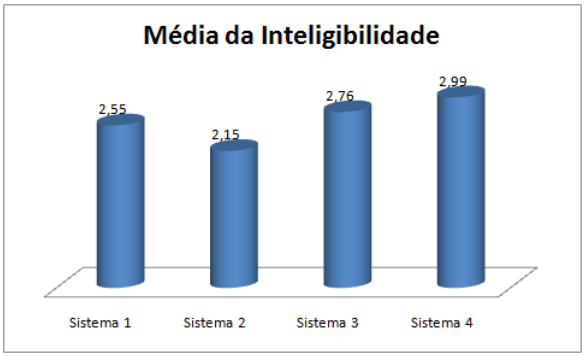


Figura 5.5: Gráfico com a média de inteligibilidade de cada sistema, em que pode ser observado que nenhum sistema foi considerado incompreensível.

variava de '1' a '5', seguindo a legenda da Tabela 5.2.

Tabela 5.2: Legenda das notas para a avaliação da naturalidade.

5	Excelente
4	Bom
3	Razoável
2	Pobre
1	Ruim

A média obtida para o conceito inicial de naturalidade de cada uma dos

sistemas pode ser vista no gráfico da Figura 5.6. Desta figura, pode ser observado que os Sistemas 3 e 4 foram os de melhor avaliação, o que foi efetivamente considerado por todos os voluntários, ou seja, todas as pessoas consultadas preferiram os Sistemas 3 e 4 aos outros dois sistemas.

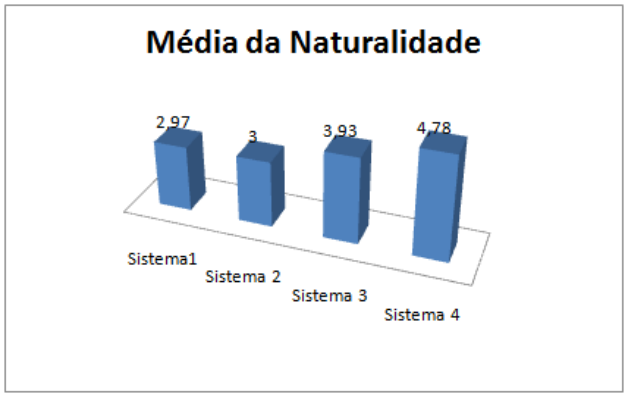


Figura 5.6: Gráfico com a média da naturalidade de cada sistema, em que pode ser vista uma clara preferência pelos Sistemas 3 e 4.

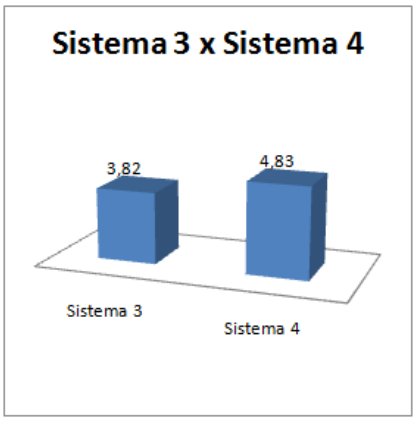


Figura 5.7: Gráfico de comparação direta dos Sistemas 3 e 4, comprovando um melhor desempenho deste último em termos de naturalidade.

A naturalidade dos Sistemas 3 e 4 foi então avaliada utilizando-se frases mais longas do que na primeira etapa. O resultado desta nova etapa é mostrado na Figura 5.7, que ilustra, novamente, a preferência média das pessoas consultadas pelos sinais sintetizados pelo Sistema 4.

5.4 Conclusão

Neste capítulo foram apresentados os resultados obtidos com os sistemas implementados. Foram descritas as avaliações comparativas entre dos diversos sistemas de síntese considerados. Bem como, foram feitas algumas análises subjetivas da inteligibilidade e naturalidade desses sistemas.

De modo geral, pudemos concluir que o método de síntese por seleção de unidades produz sinais de fala mais naturais, quando dispomos de um modelo adequado da prosódia desejada. Entretanto, também foi percebido que essa naturalidade é extremamente dependente da variabilidade do banco de frases utilizado, o que pode (e deve) ser aperfeiçoado no sistema SASPRO atual.

Capítulo 6

Conclusão

6.1 Considerações Finais

Este projeto descreveu as diversas ações adotadas no sentido de incluir um processo de busca com seleção de unidades no sistema SASPRO.

Neste sentido, inicialmente foi proposto um novo método de síntese, em que não mais é utilizado um banco de unidades previamente recortadas de modo manual. No método proposto, utiliza-se um banco de frases e um arquivo de definições que indica todas as possíveis unidades contidas no banco e onde é realizada a busca. Após a busca, a leitura das unidades é realizada diretamente na frase a que pertence. Para isso, o ponto de recorte das unidades passou a ser definido em tempo de execução na etapa final da síntese, de modo que o arquivo de definições passou a ser apenas um arquivo que lista todas as unidades disponíveis no banco.

Além disso, foi mostrado um método para se gerar o arquivo de definições automaticamente, uma vez que definir as unidades manualmente seria inviável na nova metodologia, tendo em vista a quantidade de unidades envolvidas. Neste sentido, foi também realizado um levantamento bibliográfico sobre os tipos ou tamanhos de unidades de síntese a serem utilizadas.

Finalmente, foram implementados diferentes métodos de busca com seleção das unidades. Os critérios de seleção utilizados incluíram, isoladamente ou de modo conjunto, os aspectos de tonicidade da sílaba e seus parâmetros prosódicos (*pitch*, duração e intensidade) também no nível da sílaba.

A idéia principal de se utilizar a busca com seleção é que as unidades já sejam encontradas em seu próprio contexto. Com isso, espera-se que as distorções na concatenação sejam menores, aumentando a qualidade do sinal resultante.

Como resultado dos métodos implementados, percebeu-se que o método de síntese por seleção de unidades é capaz de gerar sinais sintetizados com maior naturalidade, sendo possível, até mesmo, reproduzir certas nuances da postura do processo de fala. Entretanto, concluímos também, que para se atingir este nível de desempenho, são necessários bancos de frases gigantescos e com um extenso planejamento das unidades, de maneira que variabilidade das unidades obtidas também seja muito grande.

Tendo em vista que o objetivo deste trabalho não era gerar um sistema TTS comercial completo, mas sim investigar novas técnicas de síntese com busca elaborada de unidades, considera-se que as metas traçadas foram devidamente atingidas, o que pode ser atestado pelas avaliações subjetivas obtidas para o novo sistema.

6.2 Sugestões de Trabalhos Futuros

O tema deste projeto é bastante amplo e os desafios para o desenvolvimento de um sistema TTS de nível aceitável são quase que infinitos. Como principais formas de dar continuidade ao presente projeto, procurando preencher algumas lacunas aqui observadas, pode-se incluir, por exemplo, os seguintes aspectos:

- Definir o ponto de concatenação das unidades de maneira a minimizar o custo de concatenação. Isto pode ser feito, por exemplo, comparando também as características espectrais das unidades a serem concatenadas procurando maximizar a similariade entre as unidades no ponto em questão;

- Acrescentar um módulo de modelagem prosódica, que, a partir do processamento inicial do texto de entrada, gerasse os perfis de *pitch*, intensidade e duração desejados. Com a adição deste módulo ao sistema SASPRO, não mais necessário seria utilizar a ferramenta de transplante de prosódia, como efetuado no presente projeto;

- Acrescentar, no processo de busca, outros parâmetros de texto, tal como a posição da unidade na sílaba, tamanho da frase em questão, classe gramatical da palavra em questão etc.

- Por fim, mas sem sombra de dúvida não menos importante, incrementar substancialmente o *corpus* de frases do sistema aumentando a variabilidade das unidades de seu banco. Este parece ser o ponto fraco do sistema atual. Vale ressaltar, porém, que esta tarefa, apesar do imenso trabalho associado a ela, seria amplamente facilitada pelo conjunto de funcionalidades já presentes no sistema SASPRO atual.

Referências Bibliográficas

[1] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[2] X. Huang, A. Acero e H.-W. Hon, *Spoken Language Processing - A guide to Theory, Algorithm, and System Development*. Prentice Hall, 2011.

[3] V. L. Latsch, *Desenvolvimento de um Sistema de Conversão Texto-Fala com Modelagem de Prosódia*. Tese de Doutorado, UFRJ/COPPE, 2011.

[4] T. Dutoit, *A Introduction to Text-to-Speech Synthesis*. Academic Publishers, 2011

[5] F. O. Simões, *Implementação de um Sistema de Conversão Texto-Fala para o Português do Brasil*. Tese de Mestrado, Unicamp, 1999.

[6] Projeto ORADOR
https : $//www.linse.ufsc.br/$, acessado em 07/11/2011.

[7] Projeto Aiuruetê
http : $//www.unicamp.br/unicamp/unicamp_hoje/ju/junho2003/ju216pg03.html$,
acessado em 07/11/2011.

[8] Projeto LianeTTS
http : $//intervox.nce.ufrj.br/\ serpro/home.htm$, acessado em 07/11/2011.

[9] A. J. Hunt e A. Black, *Unit Selection in a Concatenative Speech Synthesis system using Large Speech Database. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 373-376, 1996.

[10] L. R. Rabiner e R. W. Schafer, *Introduction to Digital Speech Processing. Foundations and Trends in Signal Processing*, vol. 1, caps. 1-2, pp. 1-194, 2007.

[11] J. V. Santen, K. Alexander, E. Klabbers, e T. Mishra, *Synthesis of Prosody Using Multi-level Unit Sequences*, *Speech Communications*, vol. 46, pp. 365-375, 2005.

[12] E. da S. Morais, *Algoritmo OPWI e LDM-GA para Sistemas de Conversão Texto-Fala de Alta Qualidade Empregando a Tecnologia SCAUS*. Tese de Doutorado, Unicamp, 2006.

[13] *The C++ Resources Network*
http : *//www.cplusplus.com/reference/stl/*, acessado em 23/11/2011.

[14] K. Tokuda, T. Masuko, e T. Yamada, *An Algorithm for Speech Parameter Generation from Continuous Mixture HMMs with Dynamic Features. In Proceedings of Eurospeech*, pp. 757-760, 1995.

[15] E. D. Brill, *Some Advances in Transformation-based Part of Speech Tagging. In: Proceedings of the Twelfth National Conference on Artificial Intelligence*(AAAI-94), v. 1, 1994.

Apêndice A

Conteúdo do *Corpus* de Unidades do Sistema SASPRO

Palavras:

Aceita, Aceitamos, Aceite, Aceito;

Adora, Adoramos, Adore, Adoro;

Fala, Falamos, Fale, Falo;

Transforma, Transformamos, Transforme, Transformo.

Frases:

Joana trazia;

Menina amorosa;

Menina carinhosa;

Que horas começa?

Renata aposta;

Minhas correspondências não estão em casa;

Não fizemos uma viagem muito cansativa;

Queremos discutir o orçamento?

Temos muito orgulho da nossa gente;

Um instituto deve servir a sua meta.