UM SISTEMA DE CONVERSÃO TEXTO-FALA PARA O PORTUGUÊS FALADO NO BRASIL

Flávio Olmos Simões[1], Fábio Violaro[1]
Plínio A. Barbosa[2], Eleonora C. Albano[2]

[1]DECOM-FEEC-UNICAMP, Caixa Postal 6101, 13083-970 Campinas, SP, Brasil
[2]LAFAPE-IEL-UNICAMP, Caixa Postal 6045, 13083-970 Campinas, SP, Brasil
fosimoes@yahoo.com, fabio@decom.fee.unicamp.br
plinio@iel.unicamp.br, albano@iel.unicamp.br

Resumo – A síntese de fala a partir de texto é o objeto de estudo deste trabalho. As dificuldades relacionadas a essa tarefa são colocadas em questão e uma estratégia de implementação de um sistema de conversão texto-fala baseado em síntese concatenativa para o português do Brasil é apresentada. Tal sistema utiliza um inventário de 2.450 segmentos de fala pré-gravados, sendo capaz de empregar as técnicas de síntese híbrida e TD-PSOLA.

A adoção de critérios lingüísticos cuidadosos, sobretudo na etapa de transcrição fonética e na elaboração do inventário de unidades, constitui o ponto chave deste trabalho. A notação fonética utilizada diferencia dois tipos de segmentos fonéticos (plenos e reduzidos), que se distinguem no grau pelo qual estão sujeitos a fenômenos de coarticulação. O inventário de unidades foi construído de forma a preservar segmentos reduzidos e encontros vocálicos. No intuito de reduzir o tamanho do inventário, alguns cortes no interior de segmentos reduzidos foram efetuados. Mais uma vez, nesse caso, utilizaram-se critérios lingüísticos cuidadosos, a fim de minimizar descontinuidades espectrais após a concatenação.

Abstract - Text-to-speech synthesis is the subject treated in this work. Most of the difficulties related to this task are discussed, and an implementation of a Brazilian Portuguese text-to-speech concatenative synthesis system is presented. The system uses an inventory of 2,450 pre-recorded speech segments, and is able to employ both TD-PSOLA and hybrid synthesis techniques.

The use of carefully chosen linguistic criteria, mainly during phonetic transcription and also during the creation of the speech segments inventory, is the main contribution of this work. The phonetic notation employed distinguishes two kinds of phonetic segments (full and reduced), on the basis of the extension of coarticulation phenomena. The main criterion underlying the building of the speech segments inventory was to preserve reduced segments and vowel clusters. Nevertheless, some of the reduced segments were split, aiming at reducing the size of the inventory. Once again, in this case, specific linguistic criteria were employed, in order to minimize spectral discontinuities after concatenation.

Palavras-chave: síntese de fala, conversão texto-fala, síntese da voz, sistemas de processamento da fala, interação homem-máquina, modelos lingüísticos.

1 INTRODUÇÃO

A utilização da fala como instrumento de comunicação entre o homem e o computador é, sem dúvida, uma das maneiras de tornar a interação homem-máquina mais simples e natural. Essa característica torna-se cada vez mais desejável nos sistemas computacionais de hoje, sobretudo devido à crescente participação dos computadores nas nossas atividades diárias. Uma das maneiras de tornar possível a utilização da fala como instrumento de interação homem-máquina é por meio da conversão texto-fala.

Pode-se definir conversão texto-fala como sendo a geração automática de fala a partir de texto escrito. Esse tipo de tecnologia possui aplicações as mais diversas possíveis, como o acesso a sistemas de informação de forma automática pela linha telefônica, sistemas de auxílio a deficientes visuais e vocais, revisão de textos, leitura de correio eletrônico, dentre outros, além de facilitar a comunicação em situações em que os olhos estejam ocupados com outras tarefas.

A utilização de sistemas de conversão texto-fala é extremamente abrangente, pois a representação textual trabalha com vocabulário irrestrito e, portanto, qualquer tipo de mensagem pode ser sintetizada. A complexidade dos sistemas de conversão texto-fala advém justamente do tipo de entrada com o qual eles trabalham, pois deve-se gerar um sinal acústico de fala única e exclusivamente a partir de texto escrito.

Por essa razão, a conversão texto-fala é uma das tarefas mais complexas dentre aquelas envolvendo síntese de fala. A derivação do sinal acústico a partir do texto de entrada se faz em etapas. Cada uma dessas etapas é responsável por sucessivas transformações na forma de representação que se tem na entrada do sistema, de forma que, ao final do processamento, o texto é transformado em um sinal acústico de fala correspondente à mensagem a ser sintetizada.

Neste trabalho será descrita uma estratégia de implementação de um sistema de conversão texto-fala para o português falado no Brasil [13][6]. O sistema foi desenvolvido no Laboratório de Processamento Digital de Fala (LPDF) da Faculdade de Engenharia Elétrica e de Computação da UNICAMP, num trabalho conjunto com o Laboratório de Fonética Acústica e Psicolingüísitca Experimental (LAFAPE) do Instituto de Estudos da Linguagem, também da UNICAMP. Trata-se de um sistema de síntese concatenativa, pois a geração do sinal de fala é

feita a partir da concatenação de segmentos de fala pré-gravados. Serão apresentadas cada uma das etapas do sistema, que incluem pré-processamento do texto de entrada, transcrição fonética, processamento prosódico e a síntese do sinal propriamente dita, bem como os critérios lingüísticos que foram adotados na elaboração de cada uma dessas etapas.

2 SÍNTESE CONCATENATIVA

Antes de partir para a descrição da estrutura geral do sistema, apresentaremos algumas considerações gerais a respeito dos mecanismos que podem ser utilizados para a obtenção de um sinal de fala artificial, em especial o método de síntese concatenativa, no qual se baseia o sistema que será aqui apresentado.

Os sistemas de síntese de fala em geral (dentre eles os sistemas de síntese a partir de texto) devem ser capazes de gerar, de maneira automática, um sinal de fala tão próximo quanto possível de um sinal produzido por um ser humano. Existem diversos *métodos de síntese* que podem ser utilizados para a geração do sinal acústico de fala.

Dentre os métodos de síntese mais utilizados pelos diversos sistemas de síntese de fala podemos destacar a *síntese por regras*, a *síntese articulatória* e a *síntese concatenativa.*

A síntese por regras [10] tem seu funcionamento baseado no modelo fonte-filtro da teoria acústica de produção da fala. É um método capaz de gerar sinais de fala de alta qualidade; no entanto a determinação de parâmetros de controle para um sintetizador desse tipo é bastante complexa. Os melhores resultados para esse método de síntese tem sido obtidos para a língua inglesa.

O método de síntese articulatória [8], por sua vez, procura simular mais realisticamente o processo de produção da fala através de um modelo físico do aparelho fonador humano. Trata-se de uma área de estudos bastante recente e que apresenta resultados animadores, mas onde há ainda muito a avançar até que se possa obter um desempenho equiparável aos demais métodos de síntese.

O sistema de síntese a partir de texto apresentado neste trabalho baseia-se no método de síntese concatenativa. Trata-se de um método de síntese capaz de gerar sinais de fala de qualidade similar à dos melhores sistemas baseados em síntese por regras, e muito embora seja menos flexível no tocante às variações possíveis da qualidade de voz gerada, é de implementação muito menos complexa. Por esse motivo, é bastante utilizado em sistemas que trabalham com línguas diferentes do inglês.

A idéia por trás da síntese concatenativa é a de gerar um sinal de fala artificial a partir da concatenação de segmentos pré-gravados de fala natural. Tais segmentos devem ser selecionados a partir de um inventário de unidades previamente construído, e o conteúdo desse inventário deve ser tal que seja possível sintetizar todas as seqüências fonéticas possíveis de serem realizadas dentro de uma determinada língua.

A principal decisão a ser tomada ao se projetar um sistema de síntese concatenativa diz respeito ao tamanho das unidades básicas de fala que irão constituir o inventário para concatenação. A utilização de palavras inteiras não é uma solução conveniente, pois o número de palavras que pode ocorrer na língua é muito grande. A utilização de fones (que são as unidades básicas da fala) como unidades para concatenação, muito embora pareça ser uma saída razoável, produz resultados de concatenação bastante pobres. A principal razão para isso é que essa estratégia não leva em conta a ocorrência do fenômeno de *coarticulação* entre fones. As características espectrais de um segmento fonético são fortemente influenciadas pelos segmentos subjacentes, notadamente nas transições entre um fone e outro. Por isso, a utilização de um fone dentro de um contexto fonético muito diferente do qual ele foi extraído pode gerar descontinuidades espectrais significativas nas junções entre unidades durante o processo de concatenação [9].

Uma maneira de contemplar a existência dos fenômenos coarticulatórios é a utilização de unidades maiores denominadas *difones.* Um difone inicia-se na porção espectralmente estável de um fone e termina na porção espectralmente estável do fone seguinte. Como as junções serão sempre efetuadas em porções espectralmente estáveis de fones idênticos, as descontinuidades nas junções entre unidades serão muito menores. A transição entre fones, que é mais sujeita à coarticulação, estará sempre inteiramente contida no interior da unidade. Considere o exemplo abaixo:

/pato/ -> /p + pa + at + to + o/

Note que as junções ocorrem entre fones idênticos (/p+p, a+a, t+t, o+o/). A notação aqui utilizada é arbitrária, e o símbolo "/" representa um silêncio no início e no final da palavra.

Muitas vezes os fenômenos de coarticulação se estendem para além do segmento fonético vizinho. Nesse caso é interessante utilizar unidades maiores contendo três ou mais fones. Tais unidades são genericamente chamadas de *polifones.*

Um sistema de síntese concatenativa deve ser capaz de selecionar, a partir do inventário de unidades, uma seqüência de unidades que corresponda à sentença a ser sintetizada, e concatenar essas unidades minimizando as descontinuidades espectrais nas junções. No entanto, não basta simplesmente concatenar as unidades, uma vez que os parâmetros prosódicos (como duração e freqüência fundamental) dos fones que compõem essas unidades não são necessariamente os mesmos dos fones constituintes da sentença. Por isso, essas unidades devem ser manipuladas de forma a fornecer ritmo e entonação naturais à sentença final.

Há várias técnicas que podem ser utilizadas para a manipulação do sinal (concatenação das unidades e alteração dos parâmetros prosódicos). Dentre elas podemos citar as técnicas PSOLA (*Pitch Synchronous Overlap and Add* [7]) e a técnica híbrida [14].

3 ESTRUTURA GERAL DO SISTEMA

O sistema de conversão texto-fala descrito neste trabalho é estruturado de acordo com o diagrama de blocos da Figura 1. O papel de cada um desses blocos dentro do sistema de conversão texto-fala é apresentado a seguir.

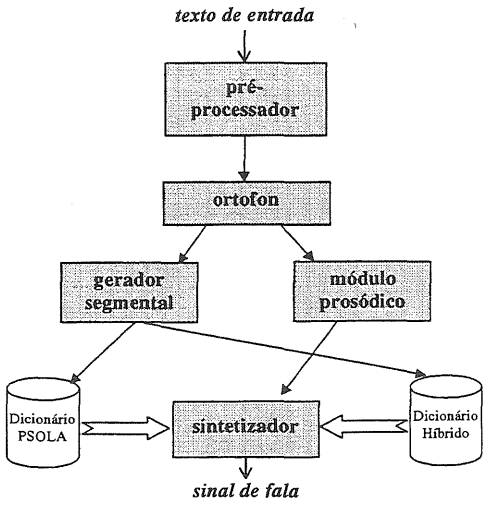

Figura 1. Estrutura geral do sistema de conversão texto-fala

4 PROCESSAMENTO LINGÜÍSTICO

A primeira fase do processo de conversão texto-fala é denominada de *processamento lingüístico.* O objetivo principal desse processamento é o de transformar a mensagem textual de entrada em uma representação simbólica que indique a seqüência de unidades básicas de fala correspondente à mensagem a ser sintetizada. Essa tarefa pode ser dividida em duas etapas distintas: o *pré-processamento* do texto e a *transcrição ortográfico-fonética* (ortofon).

4.1 PRÉ-PROCESSAMENTO

O processamento lingüístico inicia-se com a etapa de *pré-processamento.* Trata-se de uma normalização do texto de entrada, onde os elementos não-lexicais são substituídos por seus equivalentes "por extenso", de forma que o texto de saída seja formado única e exclusivamente por palavras e sinais de pontuação.

Números, siglas, abreviaturas e símbolos não alfabéticos são exemplos típicos de elementos a serem tratados durante a etapa de pré-processamento. Considere os exemplos a seguir:

Ela faz 25 anos no dia 1°.

Ela faz vinte e cinco anos no dia primeiro.

A Av. Brasil mede 5 km..

A avenida Brasil mede cinco quilômetros.

A CUT e o PT apóiam o MST.

A cut e o pê-tê apóiam o eme-ésse-tê.

2+2=4

Dois mais dois é igual a quatro.

A saída do módulo de pré-processamento é ainda textual, e dará entrada na etapa seguinte do processamento lingüístico, que é o módulo de transcrição fonética.

4.2 TRANSCRIÇÃO FONÉTICA

A função do módulo de transcrição ortográfico-fonética (ortofon) é determinar uma seqüência de unidades básicas da fala correspondente ao texto normalizado gerado após o pré-processamento. Para isso, ele deve transformar a seqüência de símbolos ortográficos em uma nova seqüência de símbolos, onde cada símbolo corresponda a uma unidade básica de fala.

No sistema apresentado neste trabalho, o módulo de transcrição fonética é dividido em duas partes distintas:

- um aplicador de *regras de transcrição,* desenvolvido no LAFAPE (Laboratório de Fonética Acústica e Psicolingüística Experimental – UNICAMP) que trata das correspondências regulares entre letras e sons;

- um *dicionário de exceções,* que contém as palavras da língua para as quais as regras de transcrição falham.

O dicionário de exceções é atualmente constituído por 1383 verbetes do minidicionário Aurélio (representação ortográfica + transcrição fonética) para os quais a versão atual das regras de transcrição ainda falham. O texto de entrada é varrido em busca de palavras contidas no dicionário. Caso uma delas seja encontrada, é substituída pela transcrição fonética correspondente. As regras de transcrição são aplicadas apenas para aquelas palavras que não constam do dicionário de exceções.

A principal função do dicionário de exceções é aumentar a taxa de acerto do módulo de transcrição, que atualmente, com a presença do dicionário, é da ordem de 96%.

A notação fonética usada pelo módulo de transcrição segue critérios lingüísticos cuidadosamente elaborados [1]. Como princípio básico na criação dessa notação, assumiu-se a existência de duas classes de fonemas distintas: os *plenos* e os *reduzidos.*

Os segmentos plenos são aqueles que ocorrem em ambientes prosódicos fortes, e são menos sujeitos à coarticulação. Dentre eles, podemos destacar as vogais tônicas e pré-tônicas, bem como as consoantes de início de sílaba.

Os segmentos fracos ou reduzidos, por sua vez, ocorrem em ambientes prosódicos mais fracos. Tais segmentos possuem, em geral, duração menor que a dos segmentos fortes e são mais suscetíveis à coarticulação com segmentos subjacentes. Eles apresentam, portanto, maior variabilidade de acordo com o contexto em que estão inseridos. Como exemplos de segmentos reduzidos podemos citar as vogais pós-tônicas e semivogais, as consoantes de final de sílaba e as líquidas de encontros consonantais, como o *'"r"* de *"prato"* e o *"l"* de*"placa".*

A notação empregada utiliza letras minúsculas para representar os segmentos plenos e letras maiúsculas para representar os reduzidos. Considere o exemplo abaixo:

casa -> kazA

A primeira vogal (*"a"*) é uma vogal tônica, e portanto constitui um segmento pleno. Já a segunda (*"A"*) é uma vogal pós-tônica, e portanto corresponde a um segmento reduzido. No exemplo, as duas consoantes são plenas, por isso são representadas por letras minúsculas.

O segundo exemplo, a seguir, ilustra com mais detalhes o processamento efetuado pelo módulo de transcrição fonética:

o senhor reginaldo rossi comprou uma dúzia de bananas por nove reais e cinqüenta centavos

O senhoR rezhinaUdU rohsI coNpRoU umA duzIA dI banaNnas poR dezenohvI reaIS I sinkUeNtA seNtavUS

A saída do módulo de transcrição fonética corresponde a uma seqüência de fonemas, que dará entrada na etapa seguinte do processo de conversão texto-fala: o módulo de processamento prosódico.

5 PROCESSAMENTO PROSÓDICO

A etapa de transcrição fonética trata tão somente da determinação da seqüência de sons que irá constituir o sinal de fala correspondente ao texto de entrada. Ela não fornece, no entanto, nenhuma informação a respeito do *ritmo* e da *entonação* das sentenças a serem sintetizadas. Essas informações são determinadas durante a etapa de *processamento prosódico.* Tais informações são essenciais para garantir não só a inteligibilidade mas também a naturalidade da fala sintetizada.

A função básica do módulo de processamento prosódico é determinar valores de *parâmetros prosódicos,* associados a cada um dos segmentos fonéticos obtidos na etapa de transcrição fonética. Existem dois tipos de parâmetros que devem ser determinados durante essa etapa:

- *duração:* corresponde ao intervalo de tempo entre o início e o final de um segmento fonético. Cada segmento possui uma duração média que varia de acordo com o contexto prosódico em que este está inserido. É essa variação que deve ser calculada pelo módulo de processamento prosódico.

- *freqüência fundamental (F0):* é um valor instantâneo que está diretamente associado à taxa de vibração das pregas vocais. Quanto mais agudo um som, maior o valor de F0. O módulo de processamento prosódico deve determinar uma curva de F0 para cada um dos segmentos fonéticos *sonoros* da sentença, de forma que o valor de F0 final de um segmento coincida com o valor de F0 inicial do segmento seguinte.

A determinação de parâmetros prosódicos deve refletir diversos aspectos da sentença (estrutura sintática, acentos lexicais, acentos frasais, além das características individuais do falante). Além disso, é altamente dependente da língua com a qual se está trabalhando. Por esse motivo, a determinação automática da prosódia deve levar em conta diversos critérios de natureza lingüística.

O módulo prosódico do sistema apresentado neste trabalho está em fase de elaboração. Há trabalhos em progresso a respeito do modelo entoacional [11][12]. Um modelo de duração [4][5], desenvolvido no Instituto de estudos da Linguagem da Unicamp, já apresenta resultados satisfatórios e deverá ser, muito em breve, incorporado ao sistema.

6 O INVENTÁRIO DE UNIDADES

O inventário de unidades é constituído pelo conjunto de segmentos de fala pré-gravados que são utilizados para a geração do sinal de fala sintetizada. Esse inventário deve ser projetado de tal forma que seja possível formar qualquer combinação de sons existente na língua a partir das unidades nele contidas.

A elaboração do inventário de unidades é uma das etapas mais importantes do projeto de um sistema de síntese concatenativa, pois somente um inventário bem construído garantirá a qualidade final do sinal de fala gerado.

O inventário de unidades do sistema descrito neste trabalho foi desenvolvido no LAFAPE e é constituído por 2.450 polifones. Dois critérios lingüísticos básicos foram adotados na sua elaboração[2], visando principalmente a evitar descontinuidades na região de concatenação entre unidades:

- Não efetuar cortes no interior de encontros vocálicos, ou seja, mantê-los intactos dentro das unidades, pois estes são extremamente coarticulados e não possuem região espectralmente estável.

- Não efetuar cortes no interior de segmentos reduzidos, por serem estes os segmentos mais sujeitos à coarticulação.

A utilização à risca dos critérios citados levaria à elaboração de um inventário de dimensões impraticáveis, com cerca de 20.000 unidades. A solução encontrada para minimizar o tamanho do inventário foi a de utilizar critérios lingüísticos que permitissem efetuar cortes no interior de alguns segmentos reduzidos, sem comprometer de forma significativa a qualidade da concatenação.

Os seguintes critérios foram aplicados:

1. Vogais pós-tônicas em posição de núcleo silábico foram segmentadas no final da transição com a consoante precedente. Isso torna as descontinuidades espectrais na região de concatenação menos perceptíveis.

   Ex.: *ótimo* -> */ohtImO/* ->

   /oh + oht + tI + Im + mO +O/

As unidades do tipo *cv* (consoante-vogal) contêm apenas o *onset* (início) da vogal (transição com a consoante precedente). A porção estável da vogal encontra-se totalmente contida na unidade seguinte.

2. Vogais nasais e ditongos nasais são concatenados com *onsets* orais. Evitou-se assim a necessidade de criar unidades específicas contendo *onsets* nasais.

   Ex.: *bomba* -> */boNbA/* ->

   /b + bo + oNb + bA + A/

põe -> /poNI/ ->

/p + po + oNI/

Tal estratégia não funciona bem para os segmentos nasais *aN, aNI, aNU,* pois o segmento oral *"a"* apresenta características espectrais muito diferentes das do início oral de *"aN"*, *"aNU "*e *"aNI"*. Nesses casos, faz-se a junção não com a oral tônica *"a"*, mas sim com a oral pós-tônica *"A"*.

Ex.: *mão -> /maNU/ ->*

/m + mA + aNU/

3. As consoantes reduzidas *"S"*, *"R"* e *"L"* são segmentadas quando ocorrem em início de encontros consonantais.

   Ex.: *pasta -> /paStA/ ->*

   /p + pa + aS + St + tA + A/

   porta -> /pohRtA/ ->

   /p + poh + ohR + Rt + tA + A/

4. *"L"* em posição de coda silábica (final de sílaba) é transformado na semivogal *"U"* (característica do português brasileiro).

   Ex.: *calma -> /kaLmA/ -> /kaUmA/ ->*

   /k + ka + aUm + mA + A/

A utilização desses critérios permitiu a redução do tamanho do inventário de 20.000 para 2.450 unidades.

7 SÍNTESE DO SINAL

A síntese do sinal de fala é a última etapa do processo de conversão. A função do módulo de síntese é fazer a concatenação das unidades e efetuar as modificações prosódicas determinadas durante a etapa de processamento prosódico.

Antes de promover a concatenação é preciso determinar a seqüência correta de unidades a serem concatenadas, ou seja, transformar a seqüência *fonética* gerada pelo módulo de transcrição em uma *seqüência de polifones*. Tal tarefa é efetuada pelo *gerador segmental*.

O exemplo abaixo ilustra o processamento efetuado pelo gerador segmental:

- *João mora na fronteira com o Paraná.*

- */zhoaNU mohRA na fRoNteIRA koN o paRana/*

- */zh-zho-oA-aNUm-moh-ohRA-An-na-af-fRo-oNt-te-eIRA- Ak-ko-oNo-op-pa-aRa-an-na-a/*

Uma vez determinada a seqüência de polifones, parte-se para a síntese do sinal propriamente dita. O processo de síntese consiste em fazer a concatenação das unidades, minimizando a ocorrência de descontinuidades nas junções, bem como efetuar as alterações prosódicas apropriadas em cada um dos segmentos fonéticos da sentença a ser sintetizada.

Existem diversas *técnicas de síntese* que podem ser utilizadas para realizar essa tarefa. O sistema aqui apresentado é capaz de efetuar a síntese por meio de duas técnicas distintas: *síntese híbrida* e *síntese TD-PSOLA*. Ambas as técnicas trabalham de forma síncrona com o período de *pitch* do sinal.

A técnica híbrida é baseada na decomposição do sinal original em duas componetes distintas: uma componente harmônica e uma componente de ruído. A componente harmônica é modelada como uma soma de senóides múltiplas da freqüência fundamental. A componente de ruído, por sua vez, é modelada como uma excitação aleatória aplicada a um filtro LPC (*linear predictive coding*) [3]. Cada uma dessas componentes é submetida a processamentos distintos, a fim de promover as alterações prosódicas apropriadas. No caso da componente de ruído, faz-se apenas alteração de duração (ressíntese LPC com novo fator de duração). Já no caso da componente harmônica faz-se alterações de duração e de F0 (modelo de síntese senoidal com interpolação da envoltória espectral e correção de fase para cada marca de síntese). Por fim, as duas componentes são somadas a fim de obter o sinal sintetizado. Apesar de permitir maiores variações prosódicas com maior naturalidade, sua complexidade computacional torna a síntese mais lenta que o TD-PSOLA. Uma descrição detalhada a respeito da síntese híbrida pode ser encontrada em [14].

A técnica de síntese TD-PSOLA é descrita a seguir.

7.1 SÍNTESE TD-PSOLA

A técnica TD-PSOLA (*Time-domain Pitch Synchronous Overlap and Add*) [7], é uma técnica de síntese conceitualmente simples e de custo computacional baixo, pois trabalha diretamente com a forma de onda do sinal de fala. Não obstante, é capaz de gerar sinais de fala de alta qualidade.

A técnica pode ser dividida em duas etapas: *análise* e *síntese*. Na etapa de análise o sinal é submetido a um algoritmo de marcação de *pitch* : as marcas de *pitch* são colocadas nos picos das porções sonoras do sinal, ao passo que nas porções não-sonoras elas são igualmente espaçadas em cerca de 10 ms. Em seguida, o sinal original é particionado em sinais elementares, através de uma seqüência de janelamentos de Hamming síncrona com o período de *pitch*. A sobreposição entre janelas adjacentes é de 50%. A Figura 2 ilustra esse processo:

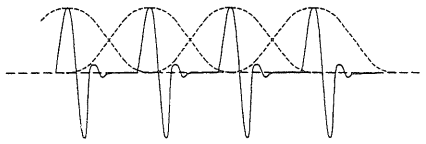

Figura 2. Janelamento do sinal de análise

O sinal assim particionado é denominado *seqüência de análise*.

Na etapa de síntese uma nova seqüência de sinais, denominada *seqüência de síntese*, é gerada a partir da seqüência de análise. Para isso, os sinais elementares são

manipulados, de forma a alterar os parâmetros prosódicos do sinal original. Há dois tipos de manipulações que podem ser efetuadas: alteração da duração e da freqüência fundamental F0 (freqüência de *pitch*).

O procedimento básico para alterar a duração do sinal consiste em omitir ou duplicar alguns dos seus sinais elementares. A omissão é utilizada quando se deseja diminuir a duração, ao passo que a duplicação permite aumentar a duração do sinal. Em ambos os casos o número de sinais elementares omitidos (ou duplicados) determina a nova duração do sinal. O processo de alteração da duração pode ser feito tanto com as porções sonoras como com as porções não sonoras. As figuras 3 e 4 ilustram o processo descrito: na Figura 3, o terceiro sinal elementar é omitido, de forma que duração total do sinal é reduzida; já na Figura 4, o sinal sintetizado possui uma duração maior que a do sinal original, uma vez que o terceiro sinal elementar foi duplicado no processo de síntese.

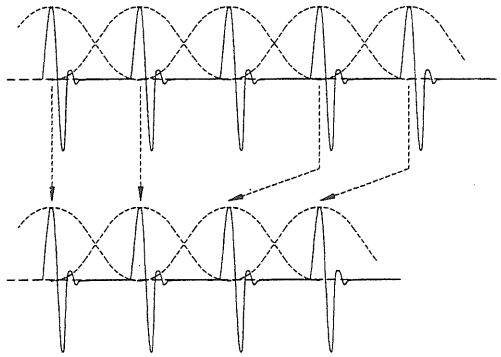

Figura 3. Redução da duração de um sinal de fala por omissão de sinais elementares

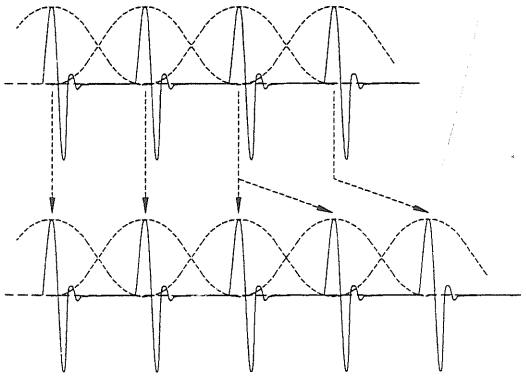

Figura 4. Aumento da duração de um sinal de fala por duplicação de sinais elementares

Para alterarmos a freqüência fundamental do sinal original devemos modificar o espaçamento entre as janelas de análise. Ao aumentarmos o espaçamento, diminuímos a freqüência, e vice-versa. Matematicamente temos:

Δt_a = intervalo de tempo entre 2 sinais elementares do sinal de análise;

Δt_b = intervalo de tempo entre 2 sinais elementares do sinal de síntese;

β = fator de alteração da freqüência fundamental.

$$\Delta t_b = \Delta t_a / \beta$$

Nesse caso, a freqüência fundamental do sinal resultante é β vezes a freqüência fundamental do sinal original.

Diferentemente do que ocorre no caso da alteração da duração, a alteração da freqüência fundamental é feita apenas com as porções sonoras do sinal de análise.

As figuras 5 e 6 ilustram o processo descrito. Na Figura 5, a freqüência fundamental (F0) do sinal foi aumentada, ao passo que a Figura 6 ilustra um sinal resultante com uma freqüência fundamental menor que a do sinal original.

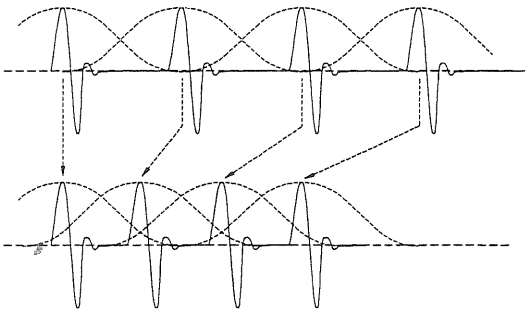

Figura 5 - Aumento da freqüência fundamental de um sinal

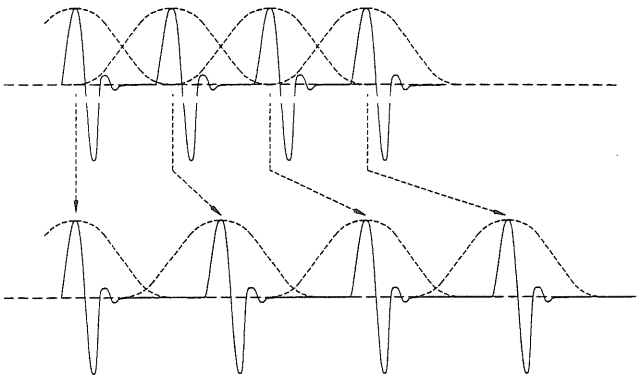

Figura 6. Redução da freqüência fundamental de um sinal

8 O SISTEMA IMPLEMENTADO

O sistema implementado foi desenvolvido no Laboratório de Processamento Digital de Fala, em conjunto com o Laboratório de Fonética Acústica e Psicolingüística Experimental, ambos da Unicamp. Ele funciona em microcomputador do tipo PC, equipado com uma placa de áudio, e opera em ambiente Windows.

O sistema foi desenvolvido em linguagem C++ e conta atualmente com todos os módulos apresentados neste trabalho, com exceção do módulo prosódico, que ainda está em fase de implementação. Para suprir a ausência momentânea deste módulo, o sistema permite a utilização de textos no formato de *entrada prosódica*, onde a duração e as curvas de freqüência fundamental dos segmentos fonéticos do texto a ser sintetizado já vêm especificadas. O exemplo abaixo ilustra o formato do arquivo de entrada prosódica, correspondente à sentença "bom dia".

0: / [120] 333 [120]

1: b [95] 97 [89]

2: oN [89] 163 [79]

3: d [79] 90 [77]

4: i [77] 149 [74]

5: A [74] 107 [52]

6: / [120] 330 [120]

Os valores entre colchetes correspondem a F0 inicial e F0 final, respectivamente. O outro valor é a duração do segmento fonético.

Os resultados obtidos com o sistema implementado foram bastante animadores. Arquivos de fala gerados *sem prosódia*, ou seja, apenas com a prosódia intrínseca das unidades utilizadas para concatenação apresentaram um grau de inteligibilidade de praticamente 100%.

A inserção manual de prosódia, como já era esperado, melhorou sensivelmente a qualidade das sentenças geradas. A presença de um módulo prosódico atuante no sistema sem dúvida implicará em ganhos consideráveis na qualidade da fala sintetizada.

O conversor texto-fala foi em seguida comparado com um protótipo anterior desenvolvido no LPDF e que empregava um dicionário de cerca de 1.500 polifones. Essa comparação mostrou uma melhora significativa de qualidade, melhora esta que pode ser atribuída a dois fatores principais:

- O uso de um dicionário com um número maior de unidades de concatenação (2.450).

- A maneira como foi feita a segmentação das unidades, seguindo uma abordagem voltada para demissílabas e não para polifones tradicionais. No caso de uma unidade *vc* (vogal-consoante), por exemplo, a segmentação foi feita do início da vogal até o meio da consoante. Trata-se pois de uma unidade longa. No caso de uma unidade *cv* (consoante-vogal), por outro lado, a segmentação foi feita do meio da consoante até o início (*onset*) da vogal. Trata-se pois de uma unidade curta. Na concatenação pode-se verificar que o ouvido humano é menos sensível a descontinuidades espectrais na região de transição. Assim, o critério usado na segmentação das unidades diminui o efeito perceptual das descontinuidades espectrais.

O conversor texto-fala descrito neste trabalho foi também apresentado no Eurospeech'99 [6], em Budapeste, Hungria, e os portugueses e brasileiros que ouviram a demonstração em fita cassete tiveram oportunidade de atestar a qualidade da síntese produzida. Para uma avaliação final foram também feitas algumas comparações informais com outros sistemas concatenativos desenvolvidos na Inglaterra e na França e disponíveis na Internet. Essas comparações permitiram mais uma vez comprovar a qualidade de síntese do sistema descrito no presente trabalho.

9 CONCLUSÕES

Este trabalho procurou apresentar uma estratégia de implementação de um sistema de conversão texto-fala para o português falado no Brasil, baseado em síntese concatenativa. Critérios lingüísticos foram utilizados na elaboração das diversas etapas do processo de conversão, notadamente na transcrição fonética e na criação do inventário de unidades, visando garantir a alta qualidade da fala sintetizada. A etapa de processamento prosódico deve também incorporar modelos lingüísticos específicos, sem os quais não será possível garantir a naturalidade do sinal gerado.

REFERÊNCIAS

[1] Albano, E., Moreira, A.A., Silva, A.H.P., Aquino, P.A., Kakinohana, R.K.; "Um conversor ortográfico-fônico e uma notação prosódica mínima para síntese de fala em língua portuguesa"; Scarpa, E.(Org.), *Estudos de prosódia*, Campinas, Editora da UNICAMP, pp. 85-109, 1999.

[2] Albano, E., Aquino, P.A.; "Linguistic criteria for building and recording units for concatenative speech synthesis in Brazilian Portuguese"; 5th European Conference on Speech Communication and Technology (Eurospeech 97), vol.2, pp.729-732, 1997.

[3] Atal, B.S., "Linear predictive coding of speech" em *Computer Speech Processing*, Fallside,F., Woods, W.A., Prentice-Hall International, University of Cambridge, 1983.

[4] Barbosa, P.A.; "A model of segment (and pause) duration generation for Brazilian Portuguese text-to-speech synthesis"; 5th European Conference on Speech Communication and Technology (Eurospeech 97), vol.3, pp.2655-2658, 1997.

[5] Barbosa, P.A.; "Revelar a estrutura rítmica de uma língua construindo máquinas falantes: pela integração de ciência e tecnologia de fala"; Scarpa, E. (Org.), *Estudos de prosódia*, Campinas, Editora da UNICAMP, pp. 21-52, 1999.

[6] Barbosa, P.A., Violaro, F., Albano, E.C., Simões, F.O., Aquino, P., Madureira, S., Françozo, E.; "Aiuruetê: a high-quality concatenative text-to-speech system for brazilian portuguese with demisyllabic analysis-based units and a hierarchical model rhythm production"; 6th European Conference on Speech communication and Technology (EuroSpeech 99) vol. 5, pp. 2059-2062, 1999.

[7] Charpentier, F., Moulines, E.; "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones"; *Speech Communication,* 9 (5/6), pp. 453-467, 1990.

[8] Cocker, C.H.; "A model of articulatory dynamics and control"; Proceedings of the IEEE 64(4), pp. 452-460, Abril de 1976.

[9] Harris, C.M.; "A study of the building blocks in Speech"; Journal of the Acoustical Society of America 25, pp. 962-969, Maio de 1953.

[10] Klatt, D.H.; "Software for a cascade/parallel formant synthesizer"; Journal of the Acoustical Society of America 67, pp. 971-995, 1980.

[11] Madureira, S.; "Pitch patterns in Brazilian Portuguese: an acoustic phonetics analysis"; Vth Australian International Conference of Speech Science and Technology, pp. 156-161, 5 a 9 de Dezembro, Perth, Austrália, 1994.

[12] Madureira, S., Fontes, M.; "Fundamental contours in Brazilian Portuguese words"; em Botinis, A., Kouroupetroglou, G., Carayannis, G., (Eds.) *Proceedings of the ESCA workshop Intonation: Theory and Applications,* September 18-20, Athens-Greece, University of Athens, pp. 211-214, 1997.

[13] Simões, F. O.; "Implementação de um sistema de conversão texto-fala para o português do Brasil"; Tese de Mestrado, Faculdade de Engenharia Elétrica e de Computação da UNICAMP, maio de 1999.

[14] Violaro, F., Böeffard, O.; "A hybrid model for text to speech synthesis", IEEE Transactions on Speech and Audio Processing 6(5), pp 426-434, 1998.

Flávio Olmos Simões nasceu em Santos, São Paulo, em 1974. Graduou-se em Engenharia de Computação em 1996 e obteve o título de Mestre em Engenharia Elétrica em maio de 1999, pela UNICAMP, na área de Processamento Digital de Fala. Atualmente é pesquisador em telecomunicações pela Fundação CPqD. Suas áreas de interesse incluem síntese de fala, codificação de voz e de áudio, avaliação subjetiva e objetiva de codecs de voz e áudio.

Fábio Violaro nasceu em Campinas, São Paulo, em 8 de dezembro de 1950. Graduou-se em Engenharia Elétrica, obteve os títulos de Mestre e Doutor em Engenharia Elétrica, todos pela Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas (FEEC/UNICAMP), em 1973, 1975 e 1980 respectivamente. Atualmente é professor titular do Departamento de Comunicações da FEEC e coordenador do Laboratório de Processamento Digital de Fala. Suas áreas de interesse se concentram em Processamento Digital de Sinais de Fala: Análise, Codificação, Reconhecimento e Síntese.

Plínio A. Barbosa nasceu em Itabuna, Bahia, em 1966. Formou-se em Engenharia Eletrônica pelo ITA, em São José dos Campos, em 1988 e recebeu o título de Mestre em Ciência pelo mesmo instituto em 1990. Doutorou-se na área de Ciência da Fala em 1994 pelo Institut de la Communication Parlée, em Grenoble, França. É professor do Departamento de Lingüística do Instituto de Estudos da Linguagem (IEL/UNICAMP) e pesquisador do Laboratório de Fonética Acústica e Psicolingüística Experimental (LAFAPE) onde desenvolve pesquisa sobre a Estrutura Rítmica do Português do Brasil com aplicações em Síntese de Fala. Seus interesses também cobrem as áreas de Fonética e Fonologia do Português do Brasil, Fundamentos Cognitivos do Ritmo, Produção de Fala, Bases Fonéticas para Síntese de Fala e Redes Neurais.

Eleonora Cavalcante Albano é professora titular do Departamento de Lingüística do Instituto de Estudos da Linguagem da UNICAMP. Coordena o Laboratório de Fonética Acústica e Psicolingüística Experimental (LAFAPE) e edita a revista Cadernos de Estudos Lingüísticos. Tem atuado nas áreas de Fonética, Fonologia e Psicolingüística, nas quais já publicou cerca de 50 artigos e 3 livros. Tem como interesse de pesquisa central a sensorimotricidade envolvida na aquisição e no uso da fala.