SISTEMAS DE CONVERSÃO TEXTO-FALA

João Paulo Teixeira, Maria João Barros, Diamantino Freitas*

joaopt@ipb.pt, mjbarros@ipb.pt, dfreitas@fe.up.pt,

Instituto Politécnico de Bragança e *Faculdade de Engenharia da Universidade do Porto

Resumo: *No presente trabalho apresenta-se a constituição de um sistema genérico de Conversão Texto-Fala, explicando as funções dos vários módulos constituintes e descrevendo as respectivas técnicas bem como o enquadramento científico com as áreas da linguística, comunicação e computação. Dada a grande importância da interface homem-máquina por voz, a função do conversor Texto-Fala, que aí tem um papel central, merece a atenção da comunidade de engenharia, em particular a electrotécnica e a informática. A presente comunicação tem como primeiro objectivo sensibilizar esta comunidade e contribuir para dinamizar a actividade neste domínio.*

1 INTRODUÇÃO

Os sistemas de conversão texto-fala (CTF), vulgarmente denominados TTS (Text-to-Speech), realizam a leitura automática de um texto. Têm vindo a ser objecto de crescente investigação desde há 4 décadas. O grande desenvolvimento dos computadores em termos de capacidade de processamento e de armazenamento abriu caminho para novas técnicas de processamento, mais elaboradas e com outros requisitos computacionais. Surgiram novas ferramentas computacionais, mais acessíveis que permitiram analisar com mais profundidades os sinais de fala. O progresso da linguística computacional trouxe também um melhor apoio linguístico à síntese da fala.

Fazer um sistema com uma qualidade de leitura fraca é relativamente simples, mas é extremamente complexo desenvolver um sistema que faça a leitura de qualquer tipo de texto com boa qualidade. A fala humana é muito complexa e apresenta um elevado número de distintas situações pouco frequentes, o que a torna difícil de modelar e muito exigente em termos da dimensão de *corpus* de fala analisada.

O progresso da qualidade da fala sintetizada passa obrigatoriamente por um elaborado bloco de análise linguístico-prosódico, dependente da língua em causa, e de modelos prosódicos.

Torna-se então necessário que as equipas que se dedicam a estes sistemas possuam conhecimentos multidisciplinares, em sistemas, programação de computadores, processamento de sinal, linguística (fonética, fonologia, prosódia, semântica, morfologia), interface homem-máquina, etc.

Existem já vários sistemas TTS para muitas línguas, incluindo o Português, alguns comerciais outros de domínio público, abertos para investigação. Alguns para aplicações específicas, com melhor qualidade nesse domínio restrito que os sistemas de aplicação geral, mas não aplicáveis a outros domínios. Sistemas com melhor ou pior qualidade, usando diversos

modelos descritos neste trabalho ainda que com pequenas modificações.

Listam-se a título de exemplo alguns endereços URL onde podem ser testados outros sistemas de síntese TTS:

http://www.elantts.com/ - Sistema comercial com várias línguas, incluindo Português do Brasil.

http://tcts.fpms.ac.be/synthesis/ - Sistema de desenvolvimento disponível, baseado no modelo MBROLA. Demonstrações em Português do Brasil e Português Europeu.

http://www.speech.inesc.pt/~lco/dixi/ Sistema TTS em Português do INESC – Lisboa.

http://www.scansoft.com/realspeak/demo/ Sistema comercial para diversas aplicações em várias línguas, incluindo Português Europeu e do Brasil.

http://www.nuance.com/prodserv/demo_vocalizer.html Sistema comercial para algumas variações do Inglês.

http://www-gth.die.upm.es/research/synthesis/synth-form-concat.html Sistema da Universidade Politécnica de Madrid, para o Espanhol.

Os sistemas TTS dividem-se em dois grandes blocos de processamento: o bloco de processamento linguístico-prosódico e o bloco de processamento acústico do sinal de fala, como mostra a Figura 1.

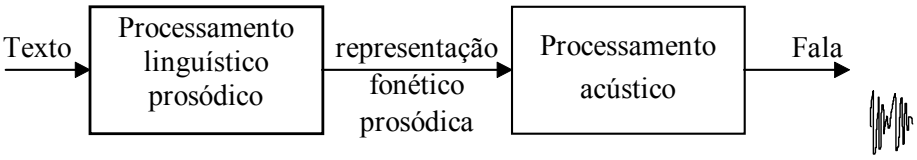

Figura 1: Blocos de processamento de um sistema TTS.

No presente trabalho pretende-se descrever sob a perspectiva da implementação alguns sub-blocos do bloco linguístico-prosódico e as técnicas mais importantes do bloco de processamento acústico. A fundamentação linguística pode encontrar-se em (Braga, D. et. al., 03).

No bloco de análise linguístico-prosódico discutir-se-ão as dificuldades inerentes a cada um dos diferentes componentes deste. Apresentam-se alguns trabalhos realizados ao nível linguístico, nomeadamente a necessidade de uma primeira filtragem do texto, o pré-processamento por conversão para extenso de acrónimos, abreviaturas e numerais, e a representação simbólica dos fonemas. Descreve-se, em seguida o bloco de transcrição fonética. Será também apresentado o enquadramento da análise sintáctica para estruturar as frases que vão ser lidas. A análise linguística retira informação do texto para prover os modelos prosódicos. A informação linguística pode, se assim for concebido, também ajudar a descriminar algumas situações no bloco de transcrição fonética.

Ainda neste bloco, discute-se a importância de análise prosódica, fundamental para uma entoação natural. Referem-se os parâmetros acústicos modelados pela prosódia, nomeadamente, frequência fundamental, durações e intensidade, por ordem de relevância.

Este bloco serve o seguinte, processamento acústico, com informação essencial, sobre a sequência de fonemas ou outras unidades acústicas que podem também ser difones, partes de palavras, palavras ou partes de frases. A esta sequência de unidades estão associados

parâmetros prosódicos com informação sobre as durações de cada unidade e os valores da frequência fundamental e da intensidade.

Seguidamente, no bloco de processamento de sinal, apresentar-se-ão os modelos de formantes, modelização sinusoidal, modelos de predição linear, modelos de concatenação no domínio dos tempos, TD-PSOLA, MBROLA, modelos articulatórios e finalmente os sistemas de selecção de unidades. Para cada um, mencionam-se as capacidades de manipulação prosódica, nomeadamente das alterações de F0 (frequência fundamental) e de duração dos segmentos.

Finalmente, na secção 4 referem-se brevemente algumas aplicações para estes sistemas.

2 PROCESSAMENTO LINGUÍSTICO-PROSÓDICO

É objectivo do processamento linguístico-prosódico determinar, a partir do texto, dois tipos de informação necessários para proporcionar ao processamento acústico dados que lhe permitam gerar fala natural. Estes dois tipos de informação são conhecidos como informação segmental e informação supra-segmental.

A informação segmental está associada à cadeia de sons que compõem a mensagem. Para cada língua existe um conjunto limitado de sons base idealizados que permitem produzir, quando correctamente combinados, todas as particularidades da fala nessa língua. Cria-se assim uma série de representações abstractas denominadas fonemas cujo número depende da língua em causa.

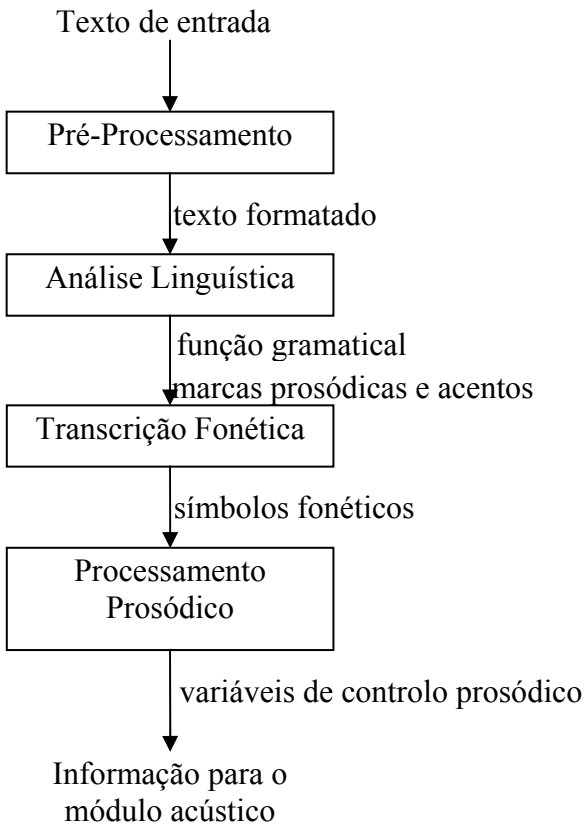

Figura 2 (ao lado): Diferentes tarefas do processamento linguístico-prosódico.

A informação suprasegmental está associada à prosódia. Reflecte tanto elementos linguísticos, tais como tipos de frase, pausas, acentuação e agrupamento de elementos de significado, como elementos não linguísticos. Esta informação vem geralmente codificada através de três parâmetros acústicos do sinal de fala:

a) A evolução temporal da frequência fundamental (F0), que é o aspecto mais importante do ponto de vista perceptivo.

b) Durações dos segmentos de som que compõem a frase.

c) Curva temporal de energia do sinal acústico.

Nos conversores texto-fala actuais estes dois tipos de informação são extraídos por uma sequência de tarefas que genericamente se representam pela Figura 2. Em cada bloco de processamento vão sendo adicionadas marcas à informação que passa para o bloco seguinte.

2.1 Pré-processamento do Texto

A primeira tarefa do processamento linguístico é a formatação do texto representando adequadamente por extenso, números, abreviaturas, acrónimos e eventuais caracteres ou conjunto de caracteres que não sejam texto, como por exemplo, o caractér €.

Estas tarefas são realizadas, com recurso a tabelas contendo as listas de acrónimos e de abreviaturas e a correspondente forma escrita. No caso dos números o programa que é utilizado é um pouco mais complexo e converte as formas numerais para extenso.

A conversão dos numerais pode ser uma tarefa relativamente simples se se considerarem apenas os numerais na sua forma natural. Contudo a situação complica-se quando se considera que estes podem ser representados de outras formas como decimais, fracções, representações na forma exponencial, ordinais, etc. E complica-se ainda mais quando se pretender reproduzir esses números de forma natural. Ora, um número de telefone não é lido da mesma forma que uma data ou uma quantidade monetária. Este caso obriga a considerar as diferentes formas de representar todas as diferentes situações e a identificá-las. Para a sua correcta identificação pode não ser suficiente a forma em que o número esta representado, tornando-se então necessário recorrer ao contexto para desambiguar estas situações.

2.2 Análise Linguística

Neste bloco, são colocadas mas marcas nas fronteiras de palavras, é realizada a divisão silábica, marcada a sílaba acentuada, ou eventuais graus de acento. Se as fronteiras de palavra são facilmente identificadas já não se passa o mesmo para as sílabas. Estas podem ser identificadas recorrendo a algoritmos que implementem um conjunto de regras de divisão silábica (Gouveia, 00). O mesmo se passa para a identificação da sílaba tónica.

Ainda neste bloco se podem incluir: a análise morfológica, identificando a função gramatical de cada palavra, recorrendo a grandes dicionários de palavras e algoritmos de desambiguação; a análise sintáctica, que recorrendo a regras por valências ou árvores generativas e à informação fornecida pela análise morfológica, marca as fronteiras sintáctico-prosódicas e tenta identificar o foco da frase, como o elemento mais relevante em termos semânticos.

2.3 Transcrição Fonética

Neste bloco o texto de entrada é transcrito foneticamente para uma sequência de fones (ou códigos de fones). Uma representação destes fones cada vez mais usada é o código SAMPA, alfabeto fonético para leitura por computador, (Wells, John).

Será importante aqui referir a distinção entre representações fonética e fonológica. Enquanto a representação fonética é o resultado de uma transcrição clássica do texto para fonemas, a representação fonológica usa os fones, que são variações dos fonemas, que são efectivamente usados numa determinada realização. E aqui reside uma diferença importante, especialmente para o português Europeu, em que a realização fonológica diverge consideravelmente da realização fonética, sendo a mais importante a redução ou elisão de muitas vogais, e de silêncios entre palavras.

Este bloco pode ser realizado por um conjunto de regras como em (Teixeira, 95), recorrendo a um dicionário com a transcrição fonética das palavras, ou com máquinas de estados (Trancoso, 02). A informação sobre a sílaba tónica é fundamental para um bom desempenho de qualquer tipo de algoritmo baseado em regras ou em máquinas de estados. A informação morfológica é também importante para desambiguar algumas situações como em <espeto>, que, se for um verbo é lido com <e> aberto, mas se for substantivo é lido com <e> fechado. A experiência do grupo neste assunto sugere ainda a utilização de uma tabela com palavras que são excepção às regras implementadas.

2.4 Processamento Prosódico

O processamento prosódico recolhe a informação supra-segmental e segmental extraída dos últimos passos (marcas prosódicas e sequência de fones) para traduzi-las em variações de duração segmental e inserção de pausas com uma duração adequada (ritmo), da frequência fundamental (F0-entoação) e da intensidade sonora, que são os parâmetros fundamentais.

Os resultados deste bloco são comprovadamente determinantes na naturalidade dos sistemas TTS, e a naturalidade é também determinante na aceitação destes sistemas pelos possíveis utilizadores. A naturalidade pode ser definida como a proximidade à forma natural, ou humana, como os sistemas "falam". Este aspecto é hoje objecto de intensa investigação pela comunidade científica que trabalha com sistemas TTS, por ainda não haver sistemas com naturalidade suficiente. Não que essa falta se deva apenas à insuficiência prosódica, mas deve-se também à falta de qualidade segmental bem como às alterações por vezes menos aceitáveis impostas pela modificação prosódica dos segmentos realizadas no bloco acústico.

Há sistemas que usam um modelo para cada parâmetro e sistemas que usam um modelo estatístico integrado para todos os 3 parâmetros (Mixdorff, 02).

2.4.1 Modelização das Durações Segmentais

O termo duração refere-se ao tempo que dura um determinado segmento de fala. Os segmentos de fala considerados são definidos de formas distintas consoante o modelo usado. Alguns autores usam unidades de mais alto nível, como as sílabas (Campbell and Isard, 91), ou grupos entre centros perceptuais, do Inglês *Inter-Perceptual-Centre-Group - IPGC*, (Barbosa e Bailly, 94). Estes algoritmos, usam, numa segunda etapa, as expressões (1) e (2) para distribuir a duração da sílaba pelos seus fonemas com um valor de z para cada sílaba. Em que μ_i, e σ_i são, respectivamente, a média e desvio padrão, do logaritmo das durações, do segmento i.

$$Dur_i = \exp(\mu_i + z\sigma_i); \qquad \sum_i Dur_i = \text{duração da sílaba} \qquad (1), (2)$$

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

A principal dificuldade deste tópico reside no conjunto de factores que podem influenciar a duração de um segmento, bem como o seu grau de influência e a forma como cada um destes factores se correlaciona com os outros. Claramente é reconhecido que as durações dos segmentos em posição de sílaba tónica são maiores. Este é então um factor a ter em conta juntamente com outros que, com maior ou menor pormenor, tentam caracterizar o contexto, tais como a identidade dos segmentos vizinhos, posição na palavra, dimensão da oração, etc. Podem também ser usados factores de natureza semântica como a proeminência da palavra, a consideração de grupos entoacionais ou tipos de frase, bem como factores prosódicos tais como níveis de acento de 'pitch', ou ainda outros factores de natureza linguística como classes de palavras.

Os modelos tradicionalmente existentes podem-se distinguir pela forma como tratam esses factores. Assim, claramente se distinguem os modelos de regras, como o de Keller-Zellner (Zellner, 94), que pela aplicação de regras mais ou menos complexas alongam ou encurtam a duração dos segmentos, os modelos matemáticos, como o de Klatt (Klatt, 76) e de Van Santen, (van Santen, 94. que pela consideração de diversos factores os combinam numa expressão, normalmente uma soma de produtos, que estabelece a duração dos segmentos, e, finalmente, modelos estatísticos, que pela aplicação de ferramentas genéricas como árvores de classificação e regressão (*Classification And Regression Trees* – CART) ou redes neuronais (*Artificial Neural Networks*), estimam também a duração dos segmentos. Alguns modelos combinam algumas destas funcionalidades, como é o caso do modelo de Barbosa e Bailly

(Barbosa e Bailly, 94) que combina redes neuronais com modelos matemáticos.

2.4.2 Modelização da Frequência Fundamental

A maioria dos sistemas TTS divide a tarefa de geração de entoação (F0), em duas componentes, a linguística e a geração das componentes de F0. As componentes linguísticas são obtidas a partir da análise do texto, sumariamente descrita atrás e aprofundada em (Braga, D. et. al., 03), deduzindo eventos entonacionais que resultam em marcas. Essas marcas são então codificadas numa representação abstracta. Exemplos dessas representações são ToBI (Silverman et al. 92), Tilt (Taylor, 98), INTSINT (Hirst et al., 00), Fujisaki (Sagisaka et al, 97), entre outras. As componentes de geração de F0 resultam do processo de descodificação dessas representações linguísticas, para curvas de F0.

Na representação ToBI cada acento é representado por não mais de dois pontos, que de forma abstracta especificam o contraste relativo entre alto (H-high) e baixo (L–low). A representação Tilt permite mais amostras que ToBI junto do pico de acento e deixa as outras regiões por especificar. A representação INTSINT consiste num sistema de entoação que define o acento por apenas um ponto. As curvas de F0 são geradas fazendo passar splines quadráticas por esses pontos.

O modelo de Fujisaki separa F0 nas componentes de entoação de frase e de acento entonacional localizado representadas por comandos de frase e comandos de acento que, de *grosso modo*, caracterizam a entoação devida à frase e à palavra, respectivamente. As curvas de F0 são geradas pela sobreposição de funções logarítmicas devidas a cada comando. Trata-se de um modelo de origem fisiológica que modela a parte do aparelho fonador humano responsável pelas variações de F0.

3 PROCESSAMENTO ACÚSTICO DO SINAL DE FALA

Neste bloco, as sequências de segmentos determinadas anteriormente são seleccionadas de uma base de dados contendo todos os possíveis segmentos, e sucessivamente concatenados de acordo com cada tipo de modelo acústico.

3.1 Modelo de Formantes

A síntese de formantes, ilustrada na figura 3, usa um modelo fonte-filtro, onde o filtro é caracterizado por variar suavemente as frequências formantes ao longo do tempo.

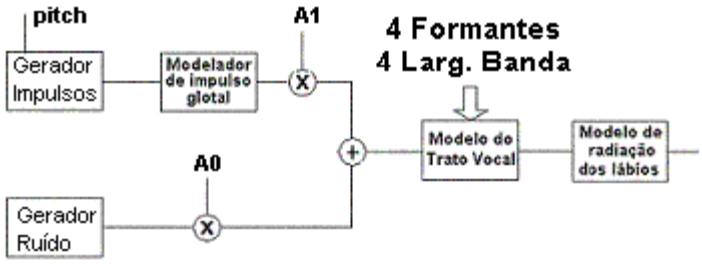

Figura 3: Esquema de um sintetizador de formantes

Este método modela a fonte glotal de som e as frequências formantes. O tracto vocal é descrito através das suas frequências de ressonância (formantes) e respectivas larguras de banda (Teixeira, 95). Para uma boa qualidade de síntese é necessário usar pelo menos quatro formantes. A qualidade finalmente obtida na prática situa-se ao nível de Bom.

O sintetizador é implementado por um filtro variante no tempo cujos parâmetros se obtêm da decomposição do polinómio do denominador de um filtro de predição linear (Barros, 02) (3) sendo F_k as frequências formante, B_k as larguras de banda e T_s o período de amostragem:

$$\frac{G}{1+\sum_{i=1}^{p} a_i z^{-i}} = G \prod_{k=1}^{p/2} \frac{1}{(1 - r_k e^{j\omega k} z^{-1})(1 - r_k e^{-j\omega k} z^{-1})} \quad (3)$$

em que:

$r_k e^{j\omega k}$ são as raízes do polinómio e $r_k e^{-j\omega k}$ as suas conjugadas,

$\omega_k = 2\pi T_s F_k$ são as fases das raízes

3.2 Modelos de Predição Linear

Este tipo de síntese, figura 4, estima parâmetros do sinal de fala em segmentos de sinal em que este é considerado estacionário devido à variação lenta das características do tracto vocal.

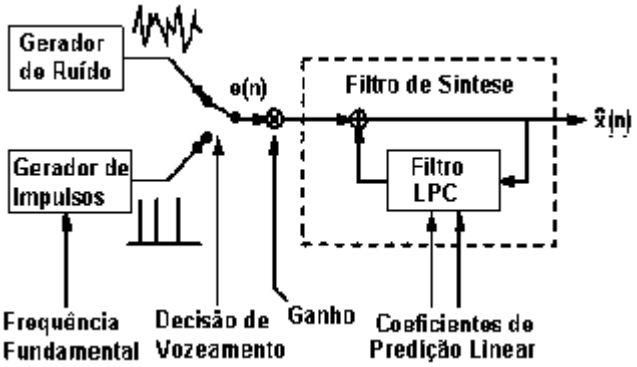

Figura 4: Esquema de um sintetizador LPC

A técnica consiste em prever o valor de uma amostra por combinação linear dos valores das amostras anteriores. Os pesos da combinação linear, denominados coeficientes LPC, são estimados por minimização do erro quadrático entre a amostra real e a sua predição, por meio de um método de autocorrelação, de covariância ou em forma de malha, e devem ser actualizados num máximo de 30ms. Normalmente escolhem-se os segmentos de fala para extracção dos coeficientes usando-se múltiplos (1,2 ou 4) do período fundamental (Barros, 02). A ordem utilizada para o filtro varia entre 10 e 18, sendo 13 um número corrente.

O filtro resultante, só com pólos, simula o tracto vocal e é alimentado, ora com um trem de impulsos que simulam os pulsos glotais e cuja periodicidade representa a frequência fundamental a produzir, no caso dos sons vozeados, ora com ruído gaussiano para os não vozeados.

O sinal à saída do filtro resulta então em:

$$s[n] = -\sum_{k=1}^{p} a_k s_q[n-k] + e_q[n] \quad (4)$$

onde e_q é o sinal excitador.

Neste tipo de síntese é preciso ter atenção tanto ao facto dos trens de impulsos terem uma composição espectral plana, como também ao facto de se impor a mesma fase na origem para todas as harmónicas, o que irá produzir um sinal sintetizado com um pico maior que o da fala natural. A qualidade finalmente obtida situa-se ao nível de Bom.

3.3 Modelização Sinusoidal

Um modelo sinusoidal é uma aproximação do sinal de fala que utiliza a sua decomposição num espaço de sinusóides. Este tipo de síntese consiste em dividir o sinal em segmentos e descrever cada um desses segmentos por uma soma de sinusóides. Se se determinarem as frequências, as amplitudes e as fases de cada uma das sinusóides é possível sintetizar o sinal.

Assim sendo, a um segmento, por exemplo j, está associado um conjunto de funções de base,que são usadas para o representar.

Estas funções de base são sinusóides com amplitudes e frequências variantes no tempo, sendo a k-ésima função de base do segmento j dada por:

$$o_k^j(t) = a_k^j(t)\cos\varphi_k^j(t) \quad (5)$$

em que:

$a_k^j(t)$ é a amplitude;

$\varphi_k^j(t)$ é a fase;

$$\varphi_k^j(t) = \int_{t_j}^{t} \omega_k^j(t)\,dt + \varphi_{ok}^j \quad (6)$$

$\omega_k^j(t)$ é a frequência instantânea da sinusóide;

t_j é o início do segmento;

e $\varphi_{ok}^j(t)$ é a fase inicial.

Os conjuntos de sinusóides de um dado segmento não são independentes das sinusóides usadas nos segmentos adjacentes. Quando uma sinusóide não está ligada a nenhuma sinusóide de segmentos anteriores ou seguintes, diz-se a "nascer" ou a "morrer", respectivamente (Marques, 90).

É necessário impor a continuidade de frequência, fase e amplitude sempre que duas sinusóides estão ligadas. Um método possível em relação à evolução das frequências, amplitudes e fases das sinusóides é determiná-las pelos seus valores instantâneos nos extremos de cada segmento e interpolar os valores intermédios. De outra forma teríamos de estimar as suas trajectórias por técnicas mais complexas (Barros, 02). A qualidade finalmente obtida pode situar-se ao nível Muito Bom.

3.4 Métodos PSOLA

Os métodos PSOLA (*Pitch Synchronous Overlap and Add*) geram o sinal de fala concatenando segmentos pré-existentes de forma síncrona com os períodos de frequência fundamental do sinal original, para o que requerem a etiquetagem do mesmo (Barros, 02).

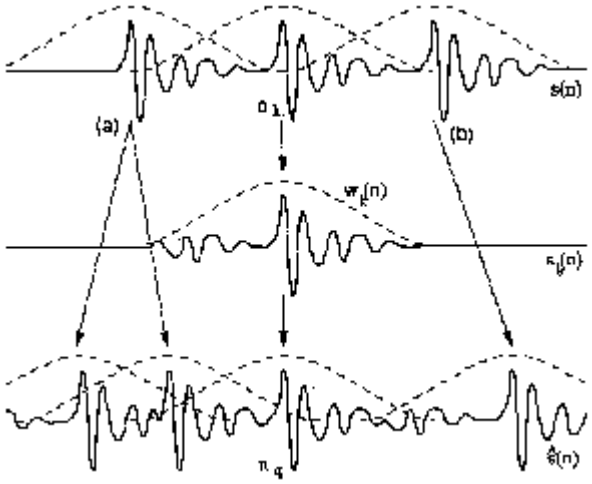

Figura 5: Modificações prosódicas usando o método TD-PSOLA

A análise PSOLA é realizada em janelas de duração correspondente a múltiplos (geralmente 2 ou 4) do período de F0, logo, de comprimento variável dependendo do segmento de sinal em questão, mas não muito variável, dado o carácter razoavelmente estacionário do sinal da fala (Barros, 02).

As modificações à frequência fundamental e à duração são realizadas pela re-sincronização dos segmentos janelados, num diferente mapeamento das marcas de pitch e, eventualmente,

pela alteração do número de segmentos. A qualidade obtenível com estes métodos pode situar-se ao nível Muito Bom.

3.4.1 Método TD-PSOLA

No método TD-PSOLA (*Time Domain PSOLA*) as modificações prosódicas são feitas directamente no sinal da fala. Para tal, é feito um mapeamento entre as marcas de frequência fundamental de análise e as marcas de síntese, para escolha dos segmentos do sinal de análise a usar em cada marca de síntese. Dependendo da situação, como se observa na figura 5, podem ter que ser repetidos ou retirados alguns segmentos de análise (Barros, 02) (Oliveira, 96). O sinal é fragmentado em segmentos s_k, em que:

$$s_k(n) = w_k(n - n_k)\, s(n) \tag{7}$$

e ω_k é uma janela com comprimento proporcional ao período de F0.

É feito um mapeamento entre os instantes de análise k e os novos instantes de síntese q, de forma a escolher que segmentos de análise deverão ser usados nos instantes n_q, com atraso adequado:

$$\hat{s}_q(n) = s_k(n + n_k - n_q) \tag{8}$$

A equação de síntese é obtida a partir de um estimador de mínimos quadrados, ficando:

$$\hat{s}(n) = \frac{\sum_q \alpha_q \hat{s}_q(n)\, w_q(n_q - n)}{\sum_q w_q^2(n_q - n)} \tag{9}$$

em que α_q é um factor compensador da diferença de energia devido à variação da distância entre segmentos.

3.4.2 Método FD-PSOLA

O modelo FD-PSOLA (*Frequency Domain PSOLA*) é idêntico ao TD-PSOLA, mas realizado no domínio da frequência.

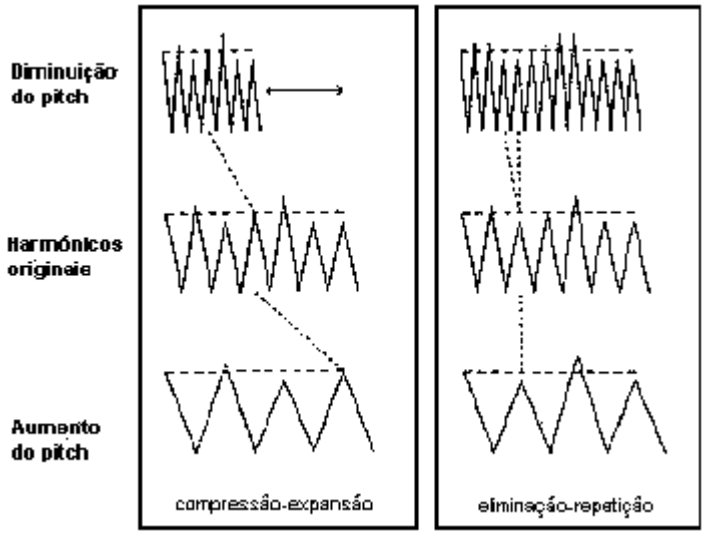

Figura 6: Modificações prosódicas usando o método FD-PSOLA

Este método começa por calcular a transformada de Fourier dos vários segmentos, obtendo depois uma estimativa da envolvente espectral através, por exemplo, de predição linear.

O quociente entre o espectro localizado e a envolvente do espectro dá-nos um espectro de riscas de amplitudes aproximadamente constantes, espaçadas da frequência fundamental. Variando o seu espaçamento e introduzindo ou removendo riscas, como se pode observar na

figura 6, e voltando a multiplicar pela envolvente, obtemos o sinal com a frequência fundamental e a duração pretendidas, mantendo as características espectrais razoavelmente inalteradas (Charpentier, 90), (Oliveira, 96).

3.4.3 Método RELP-PSOLA

O modelo RELP-PSOLA (*Residual Excitation Linear Prediction- PSOLA*) foi desenvolvido para ajudar a resolver os problemas das descontinuidades espectrais que ocorrem nas fronteiras das janelas, suavizando os sinais no domínio espectral, usando o método TD-PSOLA no resíduo de predição linear. Este modelo consiste em associar um sintetizador LPC com um algoritmo TD-PSOLA. As modificações prosódicas são realizadas usando TD-PSOLA no sinal de excitação do filtro de predição linear, em vez de realizadas directamente no sinal da fala (Barros, 02). Embora esta técnica devesse reduzir as descontinuidades espectrais, na prática a sua implementação ainda não produz melhorias de qualidade significativas (Huang et al, 01).

3.4.4 MBROLA

A técnica MBROLA (*Multiband Resynthesis Overlap and Add*) tenta resolver os problemas de fase provocados pela posição de sobreposição das janelas no processo de soma dos segmentos de sinal.

Esta técnica usa a técnica PSOLA no domínio do tempo para modificação prosódica, mas pré-processa os ciclos de pitch de modo a apresentarem uma fase fixa. A vantagem do método é que o alisamento espectral pode ser feito por interpolação directa dos ciclos de pitch no domínio do tempo, sem se adicionar qualquer complexidade extra como acontece na RELP-PSOLA. A desvantagem é que, ao mudar as fases para uma fase fixa, introduz-se ruído extra. A qualidade obtenível com estes métodos pode situar-se ao nível Muito Bom.

3.5 Modelos Articulatórios

A síntese articulatória usa um modelo físico-acústico da produção da fala, que inclui todos os articuladores. Num sintetizador articulatório, ilustrado na figura 7, a naturalidade da fala depende da modelação de alguns sub-sistemas inerentes ao mecanismo de produção da fala (Silva, 01), nomeadamente das Cordas vocais: que funcionam como fonte de sinal, ou seja, excitam o tracto vocal; do Tracto vocal: que é correctamente modelado como um sistema linear que transmite o sinal de velocidade volumétrica à saída da glote.

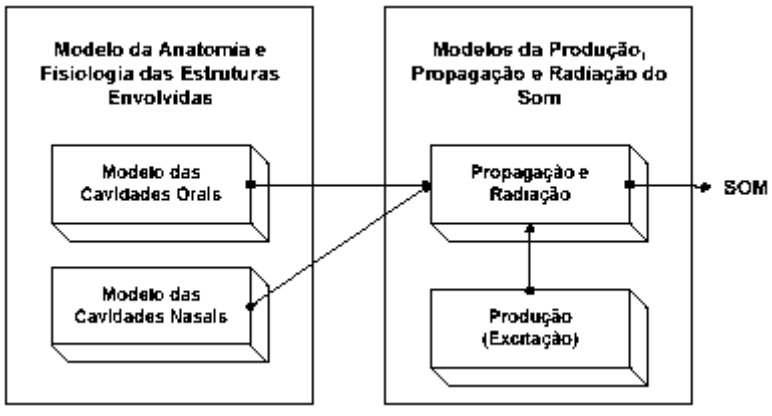

Figura 7: Módulos constituintes dos sintetizadores articulatórios

A síntese articulatória modela directamente o sistema em vez de modelar o sinal ou as suas características acústicas. Os articuladores são normalmente modelados por um conjunto de funções correspondentes às pequenas áreas de secção transversal. O modelo da corda vocal é usado para gerar um sinal de excitação apropriado (Barros, 02).

As dificuldades dos sintetizadores articulatórios prendem-se à obtenção de informação morfológico dimensional sobre o tracto e as cordas vocais durante a produção da fala, à obtenção de informação sobre a dinâmica dos articuladores e à morosidade e complexidade dos cálculos necessários (Teixeira, 01). A qualidade obtenível com estes métodos pode situar-se ao nível Muito Bom.

3.6 Sistemas de Selecção de Unidades

Mais recentemente, e na procura de evitar os problemas de distorção do sinal resultantes da sua manipulação prosódica realizada pelos métodos atrás descritos, caminha-se no sentido da construção de sistemas que não necessitem de fazer essa modificações nos segmentos, à custa de bases de dados muito grandes de segmentos para fazer a selecção. A ideia será a de encontrar na base de dados segmentos com as características prosódicas próximas das pretendidas e que produzam a menor distorção harmónica possível quando concatenados com os segmentos vizinhos.

A dificuldade destes sistemas está na construção da própria base de dados, na grande quantidade de memória necessária para esta, no tempo de processamento para escolha do melhor segmento e em encontrar algoritmos eficientes que determinem a distorção harmónica para todos os possíveis segmentos candidatos. A qualidade obtenível com estes métodos pode situar-se ao nível Muito Bom.

4 APLICAÇÕES

Os sistema de conversão texto-fala encontram já, presentemente uma série de aplicações de grande utilidade, nomeadamente, como elementos de uma interface por voz para computador, assegurando a função de saída acústica para avisos ou informações ao utilizador dos conteúdos das mensagens que o sistema procurar transmitir-lhe. Este tipo de interface tem especial interesse no acesso a sistemas de informação de qualquer natureza, tais como em estações de transporte e em comércios e tem particular utilidade para os utilizadores que tenham dificuldade em ler as indicações escritas equivalentes (cegos, amblíopes, idosos, iletrados, etc) em particular nos lugares públicos. Os sistemas de comércio electrónico podem também utilizar com vantagem a síntese da voz (Freitas, 02). Também no ensino de línguas há lugar à utilização deste tipo de sistema, nomeadamente, em complemento com o ensino personalizado, para as actividades autónomas dos alunos, destinadas a adquirir a vocalização correcta da língua, quer para crianças de tenra idade, na própria língua materna (Freitas, 03), quer para pessoas estrangeiras.

5 CONCLUSÕES

A presente comunicação dedicou-se a apresentar a constituição geral de um sistema de Conversão Texto-Fala e as funções e soluções técnico-científica dos seus componentes principais, no sentido de proporcionar uma visão de conjunto e também o enquadramento deste tipo de sistemas no âmbito da engenharia dos sistemas de informação ou electrotécnica. Assinalaram-se as fecundas ligações com as áreas da computação, da línguística, e mesmo da comunicação. Pretendeu-se que esta apresentação, na sua visão globalizante fosse informativa dos contornos técnico-científicos do problema, e simultaneamente estimuladora e convidativa a novas acções e projectos de engenharia para realização em conversão texto-fala em particular no âmbito da Comunidade dos Países de Língua Portuguesa.

6 REFERÊNCIAS:

Barbosa P., Bailly G.. Characterisation of rhythmic patterns for text-to-speech synthesis, in Speech Communication, 15: 127-137. 1994.

Barros, M. J., Estudo Comparativo e Técnicas de Geração de Sinal para Síntese da Fala, Tese de Mestrado, DEEC- Faculdade de Engenharia da Universidade do Porto, 2002.

Braga D., Freitas, D., Ferreira, H., Processamento Linguístico Aplicado à Síntese da Fala, 3º Congresso Luso-Moçambicano de Engenharia, Maputo, 2003.

Campbell, W. N. and Isard, S. D., Segment durations in a syllable frame. Journal of Phonetics, 19 :37-47. 1991.

Caseiro, D., Trancoso, I., Oliveira, L. e Viana, C., Grapheme-to-phone Using Finite-State Transducers. IEEE 2002 Workshop on Speech Synthesis, Califórnia USA. 2002.

Charpentier, F e Moulines, E., Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication, 9 (5/6):452-467. 1990.

Freitas, D., et. al., A Project of Speech Input an Output in an e-Commerce Application, actas do congresso PORTAL – Portugal for Speech Processing, Faro, 2002.

Freitas, D. Ferreira, H., A prototype application for teaching number, HCII-2003, Creta, Grécia, 2003.

Fujisaki, H., Prosody, Models and Spontaneous Speech, in Sagisaka, Y., Campbell, N.;Higichi, N.; (editors), Computing Prosody, Springer-Verlag, 1997.

Gouveia, P. D., Teixeira, J. P. e Freitas, D., Divisão Silábica Automática do Texto Escrito e Falado, nas actas do V PROPOR, Processamento Computacional da Língua Portuguesa Escrita e Falada, Atibaia – S. Paulo, 2000.

Hirst, D. J., Di Cristo, A., and Espesser, R. Levels of representation and levels of analysis for the description of intonation systems. In Horne, M., (ed.), *Prosody: Theory and Experiment. Studies Presented to Gösta Bruce*, pages 51-87. Kluwer Academic Publishers, Dordrecht, 2000.

Huang, X., Acero, A., Hon, H. W., Spoken Language Processing, a Guide to Theory, Algorithm and System Development, Prentice Hall PTR, 2001.

Klatt, D. H., Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. Journal of Acoustic Society of America, 59, 1208-1220, 1976.

Klatt, D. H., Software for a cascade/parallel formant synthetizer. Journal of the Acoustical Society of America, 67:971-995, 1980.

Marques, J. S. S., Modelamento Sinusoidal da Fala – aplicação à codificação a ritmos médios e baixos. Tese de Doutoramento – Instituto Superior Técnico 1990.

Oliveira, L. C., Síntese de Fala a partir de Texto, Tese de Doutoramento, DEEC- Instituto Superior Técnico da Universidade Técnica de Lisboa, 1996.

Silva, C. A., Automatic Extraction of the Parameters of an Articulatory Model for Speech Synthesis, Tese de Doutoramento, DEI- Universidade do Minho, 2001.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J., Tobi: A standard for labeling english prosody. *Proceedings of the International Conference on Spoken Language Processing*, volume 2, 1992.

Taylor, P. A., The Tilt intonation model. In Proceedings of the International Conference on Spoken Language Processing, Sydney, Austrália, 1998.

Teixeira, A. e Vaz, F., *European Portuguese Nasal Vowels: An EMMA Study*, Eurospeech 2001, Aalborg, Dinamarca.

Teixeira, J. P. *Modelização Paramétrica de Sinais para Aplicação em Sistemas de Conversão Texto-Fala*, Tese de Mestrado, DEEC- Faculdade de Engenharia da Universidade do Porto, 1995.

Van Santen, J. P. H., Assignment of segmental duration in text-to-speech synthesis. Computer Speech and Language, 8, 95-128, 1994.

Wells, John http://www.phon.ucl.ac.uk/home/sampa/home.htm

Zellner, B., Pauses and the Temporal Structure of Speech, chapter 3 in Fundamentals of Synthesis and Speech Recognition, Basic Concepts, State-of-the-Art and Future Challenges, by Eric Keller, John Wiley & Sons, Chichester, 1994.