MULTIVOX - Conversor Texto Fala Para Português

- João Paulo Ramos Teixeira, Escola Superior de Tecnologia e Gestão de Bragança (ESTiG) - Quinta de Santa Apolónia, Apartado 134, 5300 Bragança, Portugal; e Centro de Estudos de Física Acústica e Telecomunicações (CEFAT) - Faculdade de Engenharia da Universidade do Porto (FEUP), Rua dos Bragas, 4099 Porto Codex, Portugal (e_mail: joaopt@ipb.pt).
- Diamantino Rui da Silva Freitas, CEFAT e FEUP (e_mail: dfreitas@fe.up.pt)
- Paulo Duarte Ferreira Gouveia, ESTiG (e_mail: pgouveia@ipb.pt)
- Gábor Olaszy, Phonetics Laboratory, Linguistics Institute, Budapeste, Hungary
- G Németh, Departement of Telecommunications and Telematics, Technical University of Budapest, Hungary

Resumo:

Apresenta-se neste artigo uma visão genérica sobre sistemas de conversão texto-fala. Aborda-se de uma forma mais detalhada o sistema MULTIVOX de conversão texto fala para o português como um modelo de síntese de formantes baseado em regras. Este sistema foi desenvolvido no âmbito da cooperação entre o CEFAT e a Universidade Técnica de Budapeste em duas versões. Uma primeira versão, desenvolvida até 1996, que será descrita por blocos para permitir uma melhor compreensão deste sistema. Uma segunda versão, ainda em desenvolvimento, será apresentada nas suas potencialidades. O sistema corre num processador 386 ou superior com recurso a uma placa de circuito impresso dedicada para realizar a síntese [1], em sistemas de mínimo porte ou num processador 486 ou superior apenas com placa de som, sendo desta forma a síntese realizada inteiramente por software.

Conversão Texto-Fala

De um sistema de conversão texto-fala espera-se que converta um texto escrito de forma electrónica em fala sintetizada.

Uma primeira motivação para o estudo e desenvolvimento destes sistemas é a ajuda preciosa que podem dar a pessoas com alguns tipos de incapacidades. As principais maneiras que uma pessoa cega tem para comunicar com um computador são o terminal "braille", com custos muitas vezes inacessíveis, e o conversor texto-fala que lhe pode proporcionar a saída de som necessária para dar uma "imagem auditiva" do que acontece no écran do computador. Isto permite-lhe ter acesso a livros, mensagens e jornais em forma electrónica nomeadamente os que estejam disponíveis na rede pública de dados. É igualmente importante, por permitir exercer as suas capacidades profissionais na escrita e desenvolvimento de programação. Para pessoas com deficiência temporária ou definitiva ao nível da fala, permite que o sintetizador de fala associado a um PC se façam ouvir com recurso a um programa específico de onde podem seleccionar e compor rapidamente e com facilidade um grande número de mensagens pré-gravadas ou escritas no momento, podendo assim estabelecer comunicação ou telecomunicação.

Genericamente, este tipo de sistemas tem hoje um crescente interesse nos mais diversos campos de aplicação, como sejam: serviço de atendimento a clientes em bancos e estações de informação, dicionários multilíngua para turistas, correio de voz, instruções em simuladores, telecomunicações, etc..

Um sistema de conversão texto-fala é composto por dois módulos claramente distintos, que requerem para a sua realização uma metodologia e conhecimentos de base radicalmente distintos: o processamento linguístico-prosódico e o processamento acústico [2] como se representa na figura 1.

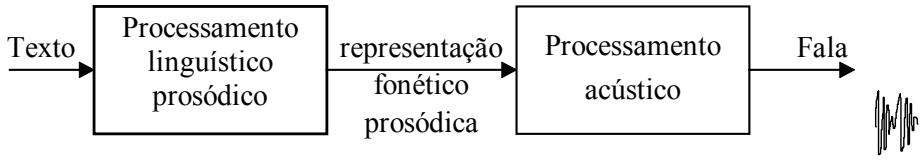

Figura 1 - Diagrama de blocos genérico de um sistema de conversão texto-fala.

Processamento Linguístico-Prosódico

É objectivo do processamento linguístico-prosódico determinar, a partir de um texto, dois tipos de informação necessária para proporcionar ao processamento acústico dados que lhe permitam gerar uma fala o mais possível natural. Estes dois tipos de informação são conhecidos como informação segmental e informação suprasegmental:

- A informação segmental está associada aos sons elementares que compõem a mensagem. Para cada língua existe um conjunto limitado de sons base ideais que permitem produzir, quando correctamente combinados, todas as particularidades da fala nessa língua. Criam-se assim uma série de representações abstractas denominadas por fonemas cuja variedade depende da língua em causa.

- 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 é considerada por muitos autores [2] a chave para conseguir uma elevada naturalidade na fala sintetizada. A informação suprasegmental vem geralmente codificada através de três parâmetros acústicos do sinal de fala:

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

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

c) Curva 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, e cujos blocos realizam as funções a seguir descritas.

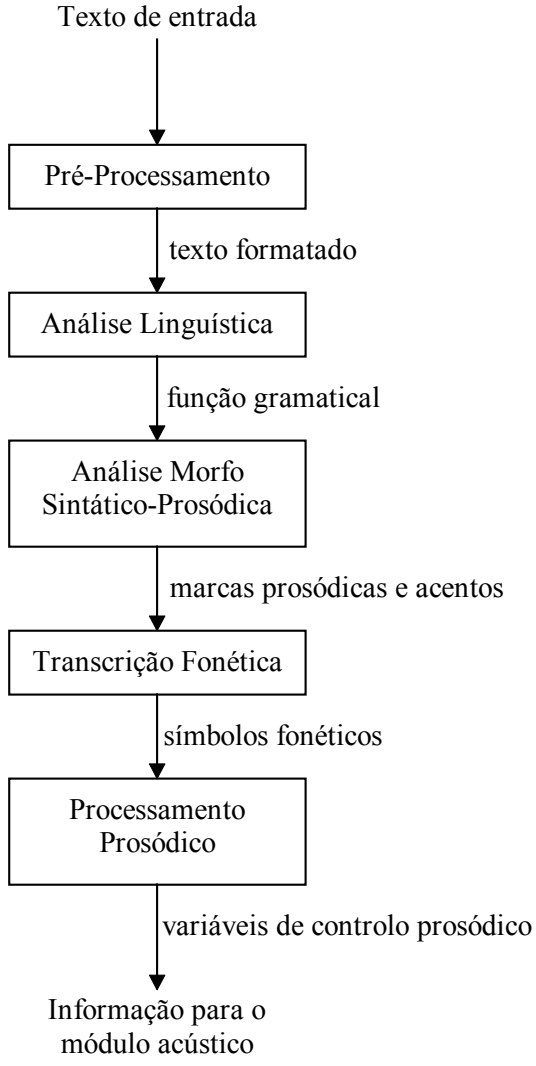

Figura 2 - Diferentes tarefas do processamento linguístico-prosódico.

Pré-processamento - A primeira tarefa a realizar no processamento linguístico é a formatação do texto representando adequadamente na sua forma textual números, abreviaturas, etc. Esta é uma tarefa bastante dependente da aplicação para que se desenvolverá o conversor.

Análise linguística - Depois do pré-processamento realiza-se a análise linguística que abarca tanto uma análise sintáctica como uma análise semântica, no intuito de encontrar o

foco (segmento com maior conteúdo semântico) da oração e tentar modelar aspectos como a ênfase. Esta tarefa é bastante complexa e muito dependente do idioma. A análise gramatical é normalmente realizada sobre um dicionário com um léxico relevante (formas verbais, expressões comuns) e uma tabela de prefixos e sufixos. Normalmente, podem incluir-se regras de conteúdo gramatical para determinar as categorias gramaticais das palavras que não tenham sido encontradas no dicionário.

Análise morfosintáctico-prosódica - Nesta tarefa pretende-se, a partir da análise anterior, marcar, por um lado, fronteiras sintáctico-prosódicas e, por outro lado, os acentos de palavra. As fronteiras sintáctico-prosódicas ficam definidas pela sua natureza (relação lógica entre duas estruturas consecutivas) e força relativa (pausas e alargamento de sílabas). Os acentos obedecem melhor a aspectos rítmicos e de ênfase principal.

Transcrição fonética - A transcrição fonética automática do texto cuja saída é uma sequência de códigos fonéticos em vez de um texto com eventuais marcas impostas pelas tarefas anteriores, é realizada, geralmente, mediante regras dependentes do contexto que devem ter em conta, por efeito de coarticulação, também, a existência de pausas e acentuação. Para a língua portuguesa estas regras são particularmente complexas no que diz respeito à transcrição das vogais, visto que, à mesma vogal do alfabeto natural correspondem várias vogais do alfabeto fonético, consoante a sua posição na palavra, acentuação e fonemas adjacentes.

Processamento prosódico - A última tarefa a realizar denomina-se processamento prosódico e recolhe a informação suprasegmental e segmental extraída dos últimos passos (marcas prosódicas e transcrição fonética) para as traduzir em variações de duração segmental (ritmo), frequência fundamental (entoação) e inserção de pausas com duração adequada.

Processamento Acústico

O objectivo do processamento acústico é converter a sequência fonética e as variáveis de controlo prosódico na forma de onda associada à voz sintetizada.

Um diagrama de blocos típico para o processamento acústico é o representado na figura 3.

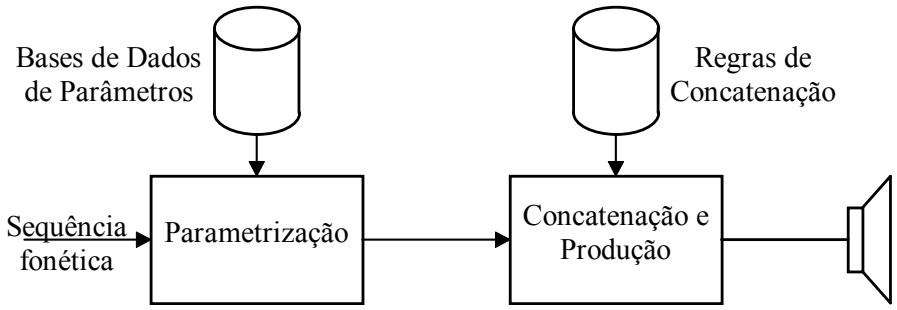

Figura 3 - Diagrama de blocos do processamento acústico.

Um aspecto a realçar é a existência de um compromisso entre, por um lado, o número de regras de parametrização e concatenação, destinadas a evitar transições bruscas desagradáveis ao ouvido e, por outro lado, o tamanho da base de dados de parâmetros. Assim, do ponto de vista do processamento acústico pode-se estabelecer uma ampla gama de sistemas de conversão texto-fala que abarca desde os sistemas comandados por regras aos sistemas comandados por dados. De uma maneira concisa num sistema "puro" comandado por regras, estas geram a representação paramétrica que alimentará um sintetizador de fala, e num sistema comandado por dados, estes representam directamente segmentos de fala. Entre estes dois casos podem-se encontrar sistemas intermédios.

Em qualquer caso o modelo de produção de fala deve ser flexível para o controlo prosódico e deve ter uma alta qualidade na geração de fala sintética. Assim, utilizam-se actualmente três grupos principais de modelos de sintetizador [2]:

1) **Sintetizadores de formantes:** nestes sintetizadores as sequências fonética e prosódica controlam as ressonâncias e a excitação do sintetizador de formantes. O sintetizador de formantes consiste numa composição de filtros que modelam as ressonâncias e anti-ressonâncias das cavidades vocal e nasal. A configuração mais genérica para o modelo destes filtros é a sua ligação em série e em paralelo. Trata-se de um procedimento, com enorme flexibilidade que mediante ajuste manual dos parâmetros do sintetizador sintetiza a fala com elevada qualidade. Sem dúvida que é necessário um número enorme de regras para a síntese automática, o que requer compiladores cada vez mais sofisticados, capazes de integrar todo o conhecimento adquirido com a experiência de trabalho com sintetizadores.

2) **Sintetizadores mediante modelos articulatórios:** trata-se de simular a propagação das ondas acústicas no trato vocal. Os segmentos e as variáveis prosódicas traduzem-se em parâmetros de um modelo simplificado do aparelho fonador humano que explicitamente restringem a dinâmica do sistema, podendo produzir voz da mais alta qualidade. Surgiram para fazer corresponder explicitamente os sintetizadores de formantes a um modelo mais explicito do trato vocal. O seu interesse centra-se no facto de as restrições implícitas neste modelo permitirem ver a fala como um contínuo acústico pelo que se evitam os problemas de concatenação de segmentos. A dificuldade principal destes tipos de sistemas é que ainda não se conhece totalmente o processo de produção da fala humana.

3) **Sintetizadores baseados em concatenação de unidades:** nestes sistemas realiza-se a concatenação de um conjunto de unidades extraídas da produção humana. Neste tipo de sintetizadores deve estar presente um algoritmo que permita, além da concatenação de unidades, modificar prosodicamente os segmentos a concatenar. Adicionalmente, nestes sintetizadores podem-se usar técnicas de codificação de voz para reduzir a necessidade de armazenamento da base de unidades acústicas. Também existe a possibilidade de incluir no modelo de codificação de voz as tarefas de concatenação e modificação prosódica, sempre que o codificador parametrize o sinal de fala com uma flexibilidade suficiente para a modificação prosódica das unidades.

O Conversor Texto-Fala MULTIVOX - Versão 1.0

O MULTIVOX, de origem Húngara, é um sistema multilíngua, com blocos dependentes da língua [1] [3], que conta já com a implementação de 9 idiomas entre os quais o Húngaro, Alemão, Italiano e Espanhol.

Trata-se de um sistema baseado em regras que realiza a transcrição fonética de um texto escrito e, a concatenação, no domínio das frequências, da sequência de sons que alimentam um sintetizador de formantes que produzirá um sinal acústico de fala.

Este sistema é bastante flexível relativamente à alteração dos parâmetros frequência fundamental, energia e duração dos segmentos de fala, permitindo uma grande facilidade de exercitar a informação suprasegmental. Estes aspectos tornam o sistema atraente numa primeira fase de aquisição de "Know how", tendo sido uma das razões para a escolha do sistema.

As ferramentas usadas nesta implementação foram:

- O próprio sistema de desenvolvimento que, ao nível do processamento acústico, com alguma facilidade, permite ouvir o som sintetizado de uma lista de parâmetros, alterar a base de sons e as regras de concatenação.

- O programa Phonovox que permite sintetizar ficheiros com sequências de parâmetros de fala natural previamente analisada, ou converter um texto escrito, já que este sistema faz uso da estrutura e bases de dados do conversor MULTIVOX. Isto permite, por comparação auditiva ou inspecção visual analítica ou gráfica da sequência de valores dos formantes, larguras de banda, frequência fundamental

e amplitude, aproximar os parâmetros da fala sintetizada dos parâmetros da fala natural. Este sistema é extraordinariamente útil no estudo de uma sequência de parâmetros, já que permite a audição continuadamente repetida, ou esporádica de partes desta sequência de segmentos de parâmetros (desde a sequência completa passando por apenas uma parte seleccionada dessa sequência até à audição de apenas um segmento), ao mesmo tempo que se podem alterar valores dos parâmetros desta sequência e ouvir os seus efeitos. É também útil no estudo de regras de entoação de frases já que permite acompanhar a evolução da frequência fundamental ao longo de uma oração.

- Um conjunto de programas em linguagem C que realizam parte do processamento linguístico.

Neste trabalho será relatada a implementação da versão da língua Portuguesa.

Do ponto de vista do desenvolvimento fonético este sistema difere dos restantes por possuir módulos dependentes do idioma, representados na figura 4 a traço interrompido.

A entrada do sistema é o texto numa sequência de códigos ASCII. Esta sequência de códigos é convertida por um "filtro de pré-processamento" numa sequência de caracteres pertencentes à representação interna de caracteres do MULTIVOX.

Neste módulo foi desenvolvido, também para o português, uma rotina em linguagem C, de conversão de números em códigos fonéticos desses mesmos números até 1 bilião.

Ainda neste módulo, e especialmente para o português, foi desenvolvida uma rotina de conversão de abreviaturas e acrónimos mediante uma tabela de regras de conversão facilmente editável e passível de alteração, acrescento ou remoção das regras. Nesta tabela é também possível editar regras para a correcta transcrição fonética de palavras homógrafas (embora escritas da mesma maneira, têm formas diferentes de pronúncia consoante a sua categoria gramatical e portanto a sua posição na oração ou fonemas adjacentes), já que este sistema não faz a distinção gramatical das palavras (exemplo: "frango no esp**e**to" - sendo "espeto" pertencente à categoria gramatical substantivo masculino com transcrição fonética [ʃpetu] em oposição a "esp**e**to a agulha" - sendo neste caso "espeto" presente do indicativo do verbo espetar conjugado na primeira pessoa do singular com transcrição fonética [ʃpɛtu]).

A delimitação de frases é realizada unicamente com base na pontuação.

Esta sequência de caracteres internos do MULTIVOX é a entrada do módulo, "Nível de conversão grafema-fonema", que realiza a transcrição fonética, regido por uma tabela de cerca de 600 regras ("Regras de conversão grafema-código de fonema na forma tabular") para a conversão de grafemas em códigos de fonemas, com base num "Grupo de sons de fonemas base". Para a língua portuguesa esse conjunto de sons base é apresentado na tabela 1 [4] [5]. Deve notar-se que alguns destes sons não são verdadeiramente fonemas (casos dos códigos 17, 31 e 35) e o seu uso nesta tabela como sons base resulta de se conseguirem melhores resultados considerando-os um só elemento em vez de os produzir por concatenação de duas outras unidades da tabela (respectivamente [α̃ u], [ʧ] e [kw]).

Um exemplo de uma das regras de conversão grafema em código de fonema é o seguinte:

agem \=202 2 28 36 38 0

Do lado esquerdo encontra-se a sequência de grafemas (terminada com espaço, indicando final de palavra) e do lado direito a sequência de códigos de fonemas que realizam a correspondente transcrição fonética. O código 202 é uma marca prosódica que indica a silaba tónica.

Neste bloco e no bloco seguinte são impostas algumas funções prosódicas. Estas funções são representadas por códigos das marcas prosódicas (números superiores a 40) que na transcrição fonética se misturam com os códigos de fonemas (números inferiores a 40). Estas marcas e respectivas funções prosódicas são apresentadas na tabela 2.

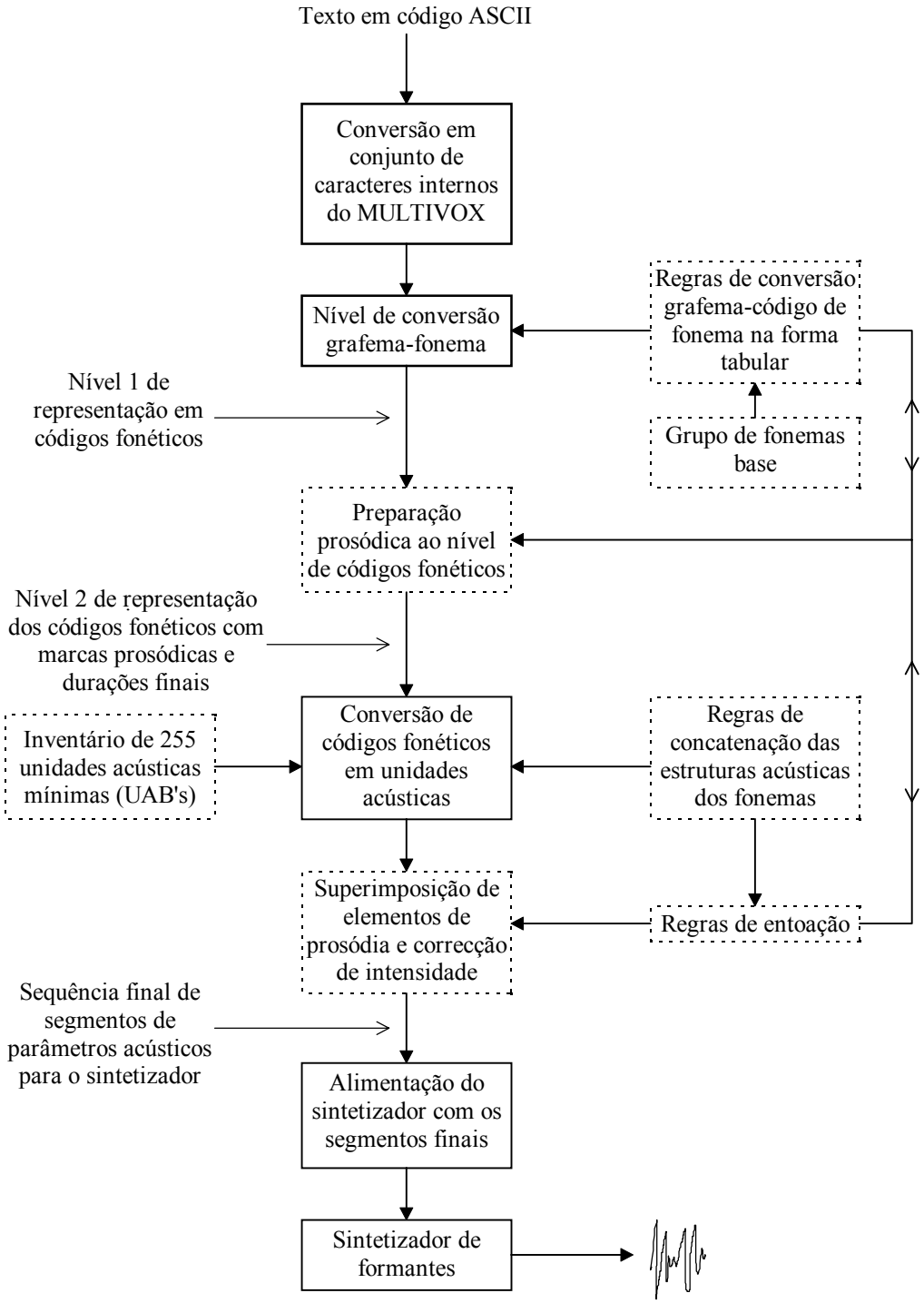

Figura 4 - Módulos constituintes do conversor MULTIVOX.

A saída deste sistema é o nível 1 de representação em códigos fonéticos. Este nível de representação consiste em códigos de fonemas e algumas marcas prosódicas inerentes já dos módulos anteriores.

O módulo seguinte, "Preparação prosódica ao nível de códigos fonéticos", baseado nos códigos fonéticos e algumas marcas prosódicas realiza, recorrendo a algoritmos desenvolvidos em linguagem C, um ajuste de códigos fonéticos para algumas regras de excepção não possíveis de implementar de forma tabular (exemplo: códigos fonéticos de [e] [m] seguidos de código fonético de uma consoante, troca a sequência de códigos fonéticos de [e] [m] pelo código fonético de [e͂]).

Tabela 1 - Símbolos fonéticos e respectivos códigos usados no conversor texto-fala MULTIVOX (versão portuguesa).

Símbolo	**Represent. no A.F.I.**	**Código do fonema**	**Exemplo**
	(pausa)	1	(pausa)
a	[a]	2	p**a**to
oo	[ɔ]	3	g**o**la
o	[o]	4	p**o**ço
u	[u]	5	p**u**la
@	[ɐ]	6	s**e**car
i	[i]	7	l**i**vro
e^	[e]	8	P**e**dro
a-	[ɑ]	9	bal**a**
e	[ɛ]	10	t**e**rra
b	[b]	11	**b**ata
p	[p]	12	**p**ara
d	[d]	13	**d**a**d**o
t	[t]	14	**t**erra
g	[g]	15	**g**ato
k, c	[k]	16	**c**asa
ão		17	c**ão**
lh	[ʎ]	18	fi**lh**o
m	[m]	19	a**m**a
n	[n]	20	**n**ada
nh	[ɲ]	21	pi**nh**o
ii	[j]	22	pa**i**
livre	livre	23	
v	[v]	24	**v**aca
f	[f]	25	**f**ilho
z	[ʒ]	26	ca**s**a
s	[s]	27	**s**ábado
j	[ʒ]	28	**j**ardim
uu	[w]	29	pa**u**
ch	[ʃ]	30	**ch**ama
tch	[ʧ]	31	**tch**eco
l	[l]	32	a**l**a
rr	[r]	33	ca**rr**o
r	[r̃]	34	ca**r**o
qu	[k] e [w]	35	**qu**ando
an	[α̃]	36	c**an**to
on	[õ]	37	p**on**te
in	[ĩ]	38	p**in**to
un	[ũ]	39	f**un**do
en	[ẽ]	40	d**en**te

Tabela 2 - Função prosódica relativa a cada código de marca no nível 1 de representação do conversor texto-fala MULTIVOX.

Código da marca prosódica	Função Prosódica
45	marca um elemento não acentuado
49	palavra interrogativa
52	aumenta a frequência fundamental (tom) durante a próxima palavra
60	aumenta intensidade da próxima palavra
61	reduz intensidade da próxima palavra
64	aumenta a frequência fundamental (tom) no texto a partir desta marca
65	reduz a frequência fundamental (tom) no texto a partir desta marca
80	palavra mais rápida
81	palavra mais lenta
85	Impede que os códigos fonéticos da palavra sejam alterados
88	Junção à próxima palavra (elimina pausa posterior)
89	Junção à palavra anterior (elimina pausa anterior)
202 a 240	Sílaba tónica. O fonema (vogal) acentuado(a) fica com o código 200 + código do fonema.
254	Vírgula

Ainda neste módulo é desenvolvida uma preparação prosódica ao nível de marcação da sílaba tónica, entoação das orações e programação rítmica.

A marcação da sílaba tónica obedece às seguintes três regras com prioridade decrescente:

1. - O texto escrito que contenha uma das seguintes letras será convertido com a marca de acentuação da palavra. {á, é, í, ó, ú, à, è, ì, ò, ù, ã, õ, â, ê, ô}. As marcas de acentuação correspondentes a estas regras são inseridas nas regras de conversão da forma tabular.

2. - Quando não há nenhuma marca de acentuação nas palavras terminadas por uma das seguintes sequências {al, el, il, ol, ul, ar, er, ir, or, ur, az, ez, iz, oz, uz}, é colocada uma marca de acentuação na última sílaba.

3. - Quando não existe ainda na palavra nenhuma marca de acentuação, é seguida a regra geral de acentuação na penúltima sílaba.

As 2ª e 3ª regras são implementadas ao nível da preparação prosódica.

As excepções a estas regras são colocadas na tabela de conversão grafema-fonema com a correcta transcrição fonética, acentuação de palavra e com o código 85, para não haver alterações.

Relativamente à entoação, as orações foram estudadas e são classificadas em três tipos: declarativa, interrogativa sem palavra interrogativa e interrogativa com palavra interrogativa.

É reconhecida uma entoação diversa para frases com tamanhos relativamente diferentes. Foi já iniciado o estudo para o contorno da frequência fundamental em frases de diferentes tamanhos, não tendo sido ainda implementada nenhuma regra desta natureza.

Ainda ao nível de entoação, quando é detectada uma virgula insere-se uma pausa antes desta e processa-se um tratamento especial da frequência fundamental na palavra anterior, impondo uma descida na penúltima sílaba até ao final da palavra.

A preparação prosódica ao nível da programação rítmica é realizada neste bloco e no bloco de "Nível de Conversão Grafema-Fonema" pela inclusão das marcas da tabela 2 nas regras de conversão na forma tabular. Neste

bloco, "Preparação prosódica ao nível de códigos fonéticos", são programadas algumas regras simples que tornam a fala sintética menos robótica e mais fluente. Por exemplo, a junção de alguns artigos com a palavra vizinha (tanto a anterior como a próxima), alterando ou suprimindo alguns códigos de fonemas na reunião de algumas palavras, especialmente nos casos de algumas vogais e fricativas. Noutros casos quando a partícula "e" é usada, produz-se um efeito idêntico ao de uma virgula.

A saída deste módulo é o nível 2 de representação dos códigos fonéticos com marcas prosódicas e durações finais.

A partir deste nível 2 de representação os códigos fonéticos são convertidos pelo módulo "Conversão de códigos fonéticos em unidades acústicas", baseado nas regras de concatenação das estruturas acústicas dos fonemas (dependentes do idioma), em unidades acústicas de um inventário de 255 elementos de unidades acústicas (também dependente da língua).

Cada unidade acústica corresponde a um segmento de fala com duração entre 10 e 50 ms e composto pelos parâmetros do modelo do sintetizador (4 formantes, 4 larguras de banda, amplitude, duração e indicação de vocalização ou não). A frequência fundamental não é um parâmetro destas unidades acústicas porque é imposta por elementos de prosódia consoante a posição destas na palavra e na oração.

Esta imposição é realizada no módulo seguinte, "Superimposição de elementos de prosódia e correcção de intensidade", baseada nas regras de entoação bem como nas marcas prosódicas existentes no nível 2 de representação dos códigos fonéticos. Este módulo não só impõe a variação da frequência fundamental, mas também faz correcções de amplitude e de duração dos segmentos acústicos. Nomeadamente a acentuação da palavra (associada à marca prosódica) é realizada pelo aumento para o dobro da duração da vogal e por uma ligeira subida da frequência fundamental. A entoação das frases é realizada de acordo com os tipos de frase:

- Para as frases declarativas foi programado um decaimento da frequência fundamental antes do fim da frase. Depois da acentuação da penúltima palavra foi implementada uma tendência decrescente da frequência fundamental. Na última palavra da frase não é realizada nenhuma variação de frequência fundamental associada à marca de acentuação.

- Para frases interrogativas com palavra interrogativa (exp.: como, onde, que, qual, quando, quantos) reconhece-se um contorno da frequência fundamental diferente das mesmas frases sem palavra interrogativa. Assim foram implementados os seguintes contornos da frequência fundamental para estas frases: uma subida no início da palavra interrogativa e uma descida brusca no início da palavra seguinte. No final da frase uma nova subida da frequência fundamental é imposta na última palavra com uma eventual descida, caso a última palavra não seja acentuada na última sílaba.

- Nas frases interrogativas sem palavra interrogativa não acontece a subida e descida da frequência fundamental no início da frase, como no caso anterior, e apenas a última palavra é afectada pela variação da frequência fundamental, como no caso anterior.

A saída deste bloco é a sequência final de segmentos de parâmetros acústicos que são enviados para o sintetizador de formantes.

O sintetizador de formantes é implementado em "hardware" ou em “software” e tem um modelo cujos parâmetros já foram referidos, 4 ou 5 formantes e respectivas larguras de banda, duração, amplitude e indicação da frequência fundamental, no caso de segmentos vocalizados, ou indicação de ruído, em segmentos não vocalizados.

Um exemplo de uma sequência de segmentos para produção de fala que chega ao sintetizador na forma paramétrica, é apresentado na tabela 3. Estes segmentos correspondem a uma parte da transição entre os fonemas (o) e (l) na palavra “olá”. F1, B1, F2, B2, F3, B3, F4 e B4 são os valores das frequências formantes e respectivas larguras de banda em Hz. Duraç. corresponde à duração do respectivo segmento em milisegundos. Pitch indica a frequência fundamental em Hz ou "noise" para segmentos não vocalizados. Amplitude indica a energia do segmento em dB.

Tabela 3 - Parâmetros dos segmentos de sinal de fala.

F1	B1	F2	B2	F3	B3	F4	B4	Duraç.	Pitch	Amplit.
612	77	1025	126	2414	190	3528	265	20	88	9
433	153	1079	433	2719	190	3528	700	31	88	11
398	77	1136	433	2719	600	3528	265	31	88	10

Poderemos apontar alguns aspectos positivos desta versão e outros aspectos que funcionam como limitações.

Como aspectos positivos pode-se referir os baixos recursos de memória usados por este sistema (140 Kbytes de memória RAM) e o facto da transmissão entre o computador e o sintetizador se realizar à insignificante taxa de 1 Kbit/seg. Estes aspectos são importantes no contexto em que este sistema foi desenvolvido, como uma aplicação para deficientes visuais que têm a necessidade de em simultâneo também terem em funcionamento um leitor de écran e um processador de texto usando os mesmos recursos de memória RAM (640 Kbytes em MS-DOS). Para aplicações na industria de telecomunicações, a baixa taxa de transmissão é um aspecto relevante já que consome poucos recurso de largura de banda.

Estas características são conseguidas à custa da economia de recursos usada pelo modelo: apenas 255 unidades acústicas base (UAB) e um máximo de 6 UAB para realizar a transição entre qualquer par de dois fonemas. A suavização das transições entre UAB's foi conseguida por meio de repetidas sessões de correcção e é beneficiada pela realização intrínseca da produção do som por via de uma interpolação linear entre os valores dos parâmetros vizinhos realizada pelo sintetizador.

Alguns destes aspectos surgem também como limitações importantes do sistema. Pois, o facto de existirem apenas 255 UAB's para a realização de qualquer som acarreta fortes limitações em termos de inteligibilidade do sistema. Por outro lado, o facto de todas as transições entre fonemas terem de ser realizadas com um máximo de 6 destes sons, traz também dificuldades na realização de sons mais complexos (variações acústicas mais complexas). Em termos de possibilidade de implementação prosódica, o sistema comporta algumas regras e permite a programação de outras. Estas restrições impõem, obviamente, uma limitação final da qualidade da fala sintetizada a nível fonético e prosódico. Resultando, contudo, uma fala inteligível que, para as aplicações inicialmente previstas, ajuda a pessoas com limitações, cumpre os objectivos pretendidos.

MULTIVOX versão 2.0

A segunda versão deste sistema, actualmente em desenvolvimento, suprime as limitações da 1ª versão.

A conversão grafema-código de fonema é realizada também de uma forma tabular, baseada no conjunto de regras da versão anterior, mas com a possibilidade de introduzir marcas prosódicas de separação de silabas e de palavras, de controlo dos parâmetros acústicos, etc…

O conjunto de fonemas base é o mesmo acrescido do fonema mudo “h”. A base de dados fonética é agora conseguida à custa de fala humana num processo que passou pela gravação por um locutor de todas as combinações de fonemas, pela posterior análise e parametrização destes sinais de fala, segundo o modelo usado pelo sintetizador (4 formantes e respectivas larguras de banda) e, finalmente, pela construção de uma base fonética contendo todas as transições entre fonemas baseada nos sinais de fala adquiridos e analisados. Isto leva a uma melhoria significativa da qualidade fonética do sistema e, portanto, da sua inteligibilidade.

Em termos prosódicos, esta segunda versão permite a implementação de um modelo prosódico mais elaborado [6].

REFERÊNCIAS:

[1] - Olaszy, Gábor, G. Gordos e Németh, Géza "The MULTIVOX multilingual text-to-speech converter", Talking Machines 1992.

[2] -López, Eduardo "Estudio de Técnicas de Processado Linguístico y Acústico Para Sistemas de Conversión Texto-Voz en Espanhol Baseado en Concatenatión de Unidades", Tesis Doctoral - Universidad Politécnica de Madrid 1993.

[3] -Teixeira, João Paulo "Modelização Paramétrica de Sinais Para Aplicação em Sistemas de Conversão Texto-Fala", dissertação de Mestrado em Engenharia Electrotécnica e de Computadores na Faculdade de Engenharia da Universidade do Porto, 1995.

[4] -Martins, M. R. Delgado "Ouvir Falar - Introdução à Fonética do Português", segunda edição, Caminho Colecção Universitária série Linguística 1992.

[5] -Mateus, M. H. Mira, Andrade A., Viana, Maria do Céu, Villalva A. "Fonética, Fonologia e Morfologia do Português", Universidade Aberta, Lisboa 1990.

[6] -Olaszy, Gábor e Németh Géza "Prosody Generation for German CTS/TTS Systems - from theoretical intonation patterns to practical realisation".