

Sistemas de Síntese de Fala (TTS) Sistemas de Reconhecimento de Fala (ASR)

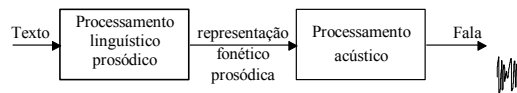
Na perspectiva de:
Ajudas na Reabilitação – TR
Desenvolvimento – APS

João Paulo Teixeira
joaopt@ipb.pt
www.ipb.pt/~joaopt
Departamento de Electrotecnia
ESTiG – Instituto Politécnico de Bragança

Resumo

- ▶ **Sistemas TTS**
 - Blocos Genéricos de um sistema TTS
 - Processamento Linguístico
 - Fonemas do Português
 - Processamento Prosódico
 - Processamento Acústico
 - Formantes
 - LPC
 - Sinusoidal
 - PSOLA (TD-PSOLA)
 - Articulatórios
 - Selecção de Unidades
- ▶ **Sistemas ASR**
 - Reconhecimento do Orador
 - Reconhecimento de palavras Isoladas
 - Reconhecimento de fala contínua
 - Independência do orador
- ▶ Exemplos de produtos comerciais/Empresas ligadas à Tecnologia da Fala
- ▶ Ajudas de Sistemas TTS e ASR na Reabilitação

Conversão Texto-Fala (TTS)

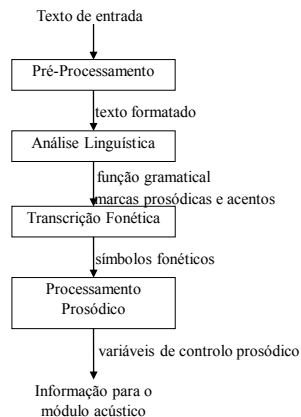


- ▶ **Processamento Linguístico prosódico** – processa informação ao nível segmental e supra-segmental
- ▶ Ao nível segmental corresponde à criação da cadeia de sons (fonemas) que compõem a mensagem
- ▶ Ao nível supra-segmental:
 - reflecte elementos linguísticos, não linguísticos e para-linguísticos.
 - Determina os parâmetros prosódicos:
 - F0
 - durações segmentais
 - intensidade.

JPT

3

Processamento Linguístico-Prosódico



- ▶ **Pré-processamento** – formatação do texto / conversão de números, datas, abreviaturas, acrónimos, etc.
- ▶ **Análise Linguística:**
 - definidas as fronteiras de palavra e de frase,
 - divisão silábica,
 - sílaba tónica,
 - análise morfológica (função gramatical)
 - análise sintáctica
 - marcação do foco

JPT

4

Processamento Linguístico-Prosódico

- ▶ **Transcrição fonética:** consiste na conversão dos grafemas do texto escrito numa sequência de fonemas a serem reproduzidos pelo TTS. Pode ser realizado recorrendo a um conjunto de regras, a um dicionário com transcrição fonética ou com máquinas de estados.
- ▶ **Processamento Prosódico:** com base na informação segmental e supra-segmental determina os contornos de F0 (entoação) as durações dos segmentos (fonemas) e a intensidade do sinal acústico. É fundamental para conferir naturalidade.
- ▶ **Modelo de Prosódia para o Português Europeu (Teixeira, J. P. 2004)**
 - [“Conhece a situação na pele. Aprendeu-a na idade em que se aprende e se não esquece”](#)
 - [“Acusar os trabalhadores é uma chocante demonstração de que afinal a justiça não está acima das classes sociais”](#)

Fonemas do Português Europeu

Vogais

Vogais Orais

[i]	lívro
[e]	Pedro
[ɛ]	terra
[a]	pato
[α]	mano
[ɔ]	gola
[u]	pular
[]	secar
[o]	poço

Vogais Nasais

[i]	pinto
[ɛ]	dente
[α]	canto
[o]	ponte
[u]	fundo

Semi Vogais

[j]	paí
[w]	pau
[j]	mãe
[w]	cão

Fonemas do Português Europeu

Consoantes

Consoantes Oclusivas Orais

[p]	<i>para</i>
[b]	<i>bata</i>
[t]	<i>tarde</i>
[d]	<i>dado</i>
[k]	<i>cão</i>
[g]	<i>gato</i>

Consoantes Oclusivas Nasais

[m]	<i>ama</i>
[n]	<i>nada</i>
[ɲ]	<i>pinho</i>

Consoantes Fricativas

[f]	<i>fado</i>
[s]	<i>sábado</i>
[ʃ]	<i>chão</i>
[v]	<i>vaca</i>
[z]	<i>casa</i>
[ʒ]	<i>jardim</i>

Consoantes Laterais

[l]	<i>lado</i>
[λ]	<i>filho</i>

Consoantes Vibrantes

[r]	<i>porta</i>
[R]	<i>carro</i>

JPT

7

Fonemas do Português Europeu

Ditongos decrescentes orais

/aj/ /saj/

sai

/ej/ /e'nej/

anéis

/ej/ /sej/ (em Lisboa /sej/)

sei

/oi/ /moj/

mói

/oj/ /moite/

moita

/uj/ /fuj/

fui

/iw/ /viw/

viu

/ew/ /mew/

meu

/ɛw/ /vɛw/

véu

/aw/ /maw/

mau

/ow/ /sow/ (no Norte de Portugal)

sou

Ditongos crescentes nasais

/ēj/ /mēj/ *mãe*

/ōj/ /e'nōjj/ *anões*

/ūj/ /mūjje/ *muita*

/ëw/ /mëw/ *mão*

ditongo crescente – semi-vogal seguida de vogal

ditongo decrescente – vogal seguida de semi-vogal

ditongos crescentes orais – só nas situações de qu... ou gu...

JPT

8

Alfabeto Fonético Internacional / Sampa

- ▶ AFI – Alfabeto Fonético Internacional – conjunto de símbolos reconhecidos internacionalmente que identificam cada som/fonema.
- ▶ Sampa – Computer Readable phonetic Alphabet – usa símbolos facilmente utilizados nos computadores.

Vogais / Sampa

Classe	Símbolo IPA	Símbolo SAMPA	Elevação língua	Posição língua	Palavra	transcrição SAMPA
Vogais		6	média	média	cama	k6m6
		a	baixa	média	cara	kar6
		e	média	anterior	pêra	per6
		E	baixa	anterior	sete	sEt@
		@	alta	média	que	k@
		i	alta	anterior	fita	fit6
		o	média	posterior	dou	do
		O	baixa	posterior	corda	kOrd6
		u	alta	posterior	mudo	mudu
		6~	média	média	manta	m6~t6
		e~	média	anterior	menta	me~t6
		i~	alta	anterior	pinta	pi~t6
		o~	média	posterior	ponta	po~ta
		u~	alta	posterior	mundo	mu~du
Glides		w	alta	posterior	pau	paw
		j	alta	anterior	pai	paj
		w~	alta	posterior	cão	k6~w~
		j~	alta	anterior	mãe	m6~j~

Consoantes / Sampa

Classe	Símbolo IPA	Símbolo SAMPA	Vozeado	Ponto de articulação	Palavra	transcrição SAMPA
Oclusivas	p	p0,p	não	média	pai	p0paj
	t	t0,t	não	média	tia	t0ti6
	k	k0,k	não	anterior	casa	k0k6za
	b	b0,b	sim	anterior	bar	b0bar
	d	d0,d	sim	média	data	d0dat6
	g	g0,g	sim	anterior	gato	g0gatu
Fricativas	f	f	não	posterior	férias	fErj6S
	s	s	não	posterior	selo	selu
	ʃ	S	não	posterior	chave	Sav@
	v	v	sim	média	vaca	vak6
	z	z	sim	anterior	azul	6zul~
	ʒ	Z	sim	anterior	agir	6Zir

Consoantes / Sampa (contin.)

Classe	Símbolo IPA	Símbolo SAMPA	Vozeado	Ponto de articulação	Palavra	transcrição SAMPA
Nasais	m	m	sim	bilabial	meta	mEt6
	n	n	sim	apicodental	neta	nEt6
		J	sim	palatal	senha	s6J6
Líquidas	l	l	sim	apicodental	lado	ladu
		l~	sim	apicodental	sal	sal~
		L	sim	palatal	folha	foL6
	R	R		velar	carro	kaRu
		r		apicodental	caro	karu
Silêncio	sil					

Software

- ▶ **Praat: doing phonetics by computer**

<http://www.fon.hum.uva.nl/praat/>



- ▶ **SFS – Speech Filing System**

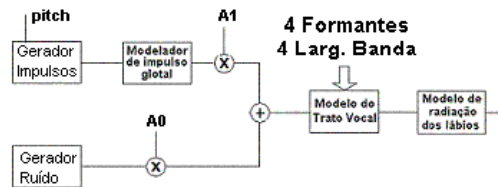
<http://www.phon.ucl.ac.uk/resource/sfs/>



Processamento Acústico – Sintetizador de Fala

- ▶ Converte a sequência de fonemas e as variáveis de controlo prosódico (F0, durações e intensidade) num sinal acústico.
- ▶ Vários modelos:
 - Formantes
 - Predição Linear
 - Modelização Sinosoidal
 - Métodos Psola
 - TD Psola
 - FD Psola
 - Relp Psola
 - MBrola
 - Modelos Articulatorios
 - Modelos de Selecção de Unidades
 - HMM

Sintetizador de Formantes

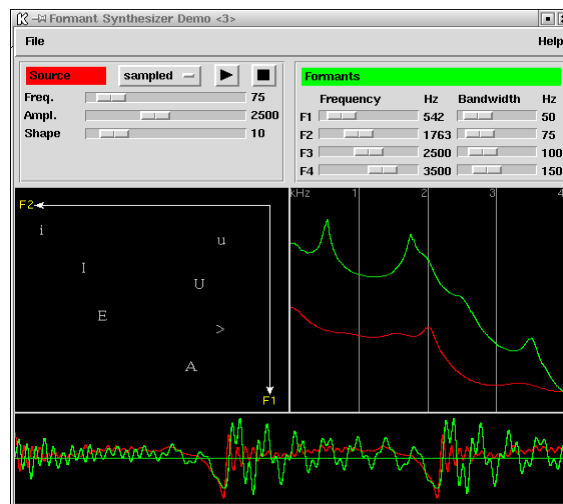


- ▶ Parâmetros prosódicos: pitch (F0), A1 ou A0 e a duração dos parâmetros associados a cada fonema.
- ▶ Parâmetros do trato vocal para cada som (fonema)
 - 4 formantes e respectivas larguras de banda.

JPT

15

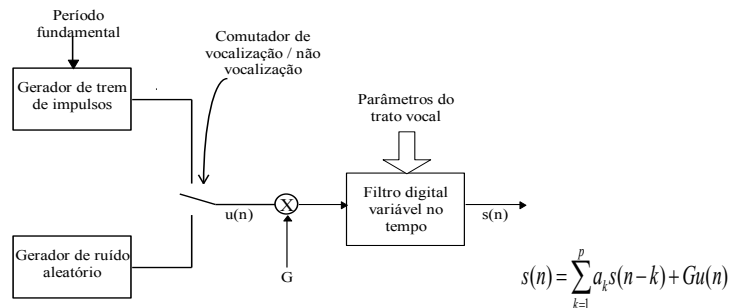
Sintetizador de Formantes



JPT

16

Sintetizador LPC (Linear Predictive Coding)



$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^p a_k z^{-k}}$$

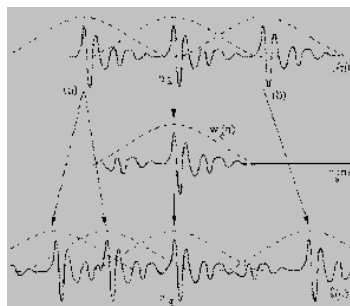
Modelização Sinusoidal

- ▶ Consiste em dividir o sinal em segmentos de curta duração e para segmento decompô-lo num conjunto de sinusóides, representadas pelas suas frequências, amplitudes e fases.
- ▶ As sinusóides dos sucessivos segmentos ora estão ligadas (continuas) ora estão a nascer (nova) ou a morrer (terminam).
- ▶ Este método produziu bons resultados na década de 1990 [Marques, 90], mas foi superada em qualidade por técnicas posteriores.

Métodos Psola

- ▶ Pitch-synchronous-overlap-add (mais usados: TD-Psola, FD-Psola, MBROLA)
- ▶ Os métodos PSOLA (*Pitch Synchronous Overlap and Add*) geram o sinal de fala concatenando segmentos pré-existent de forma síncrona com os períodos de frequência fundamental do sinal original.
- ▶ Produziram excelentes resultados, nomeadamente o TD-Psola, ainda muito actual.
- ▶ Permitem a alteração de parâmetros prosódicos com pouca perda de qualidade.

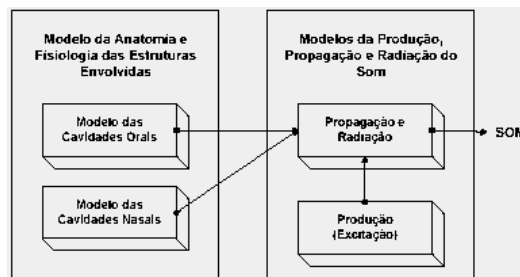
TD-Psola – Modificações prosódicas



- ▶ A análise Psola é realizada com janelas de duração de múltiplos (2 ou 4) períodos fundamentais.
- ▶ A alteração da duração de um segmento (ex. fonema) é realizada pela repetição ou remoção de períodos fundamentais.
- ▶ A alteração da frequência é realizada pela re-sincronização e alteração do número de segmentos.
- ▶ É necessário ter a marcação dos períodos fundamentais.

Modelos Articulatorios

- ▶ Fazem uma modelação físico-acústica do sistema de produção de fala.
- ▶ É um modelo físico muito complexo e difícil de modelar.
- ▶ Há uma grande dificuldade em obter os parâmetros reais do movimento dos articuladores durante a fonação.
- ▶ Conseguem-se reproduzir algumas vogais.
- ▶ O seu interesse prende-se mais com o conhecimento anatómico do processo de produção de fala.



Modelos de Selecção de Unidades

- ▶ Consiste em ter gravada uma enorme base de dados que represente (quase) todas as situações de combinação de parâmetros linguísticos e prosódicos. Depois em cada situação o sistema deve procurar o segmento mais próximo do desejado na enorme base de dados e usá-lo sem necessidade de fazer modificações ao som original.
- ▶ **Problemas:** a construção da base de dados é uma tarefa muito grande; a procura do melhor segmento na base de dados tem que ser muito eficiente para o sistema funcionar em tempo real. A dimensão da base de dados torna o TTS muito 'pesado' em recursos de memória e de capacidade de processamento, limitando a sua utilização.
- ▶ **Vantagens:** teoricamente a qualidade tende para o óptimo; pode ser adequado para sistemas dedicados com um leque reduzido de léxico. Pode ser usado em combinação com outros sistemas de concatenação de unidades.

ASR – Automatic Speech Recognition – extrair informação do sinal de fala



23

ASR – Reconhecimento do Orador

- **Identificação do orador** – Identifica quem está a falar de entre um conjunto de pessoas conhecidas.



- **Verificação do orador** – verifica se quem está a falar é a pessoa que diz ser. Há cooperação do orador.

Ele é realmente o Pedro?

- Em ambos caso é necessário treinar o sistema com a voz da própria pessoa.

JPT

24

Modalidades de Reconhecimento de Fala

- **Reconhecimento dependente do Texto**

- O sistema de reconhecimento conhece o texto que a pessoa pode falar
- Exemplos: frases conhecidas, frases de solicitação.
- Usado em aplicações com grande controlo sobre o texto de entrada
- O conhecimento prévio do texto melhora o desempenho do sistema

- **Reconhecimento independente do Texto**

- O sistema não conhece o texto que pode ser dito
- Exemplos: texto à escolha do utilizador, conversação
- Usado em aplicações sem controlo do texto de entrada
- Sistema mais flexível mas que envolve uma maior dificuldade
- O sistema pode prover conhecimento sobre a construção frásica

Reconhecimento de Palavras Isoladas

- ▶ Reconhece uma de um leque de várias palavras com elevada taxa de acerto.
- ▶ Usada para controlo de sistemas através da fala.
- ▶ Permite uma interface falada usando uma navegação em menus de opções.
- ▶ Pode ser independente do falante ou dependente do falante, neste caso com melhores performances.

Empresas que desenvolvem Aplicações Texto-Fala

▶ Empresas de *Speech*:

- Português Dígitro – http://www.digitro.com/pt/tecnologia_tecnologias_sintese-de-fala.php
- Italiana Loquendo: http://www.loquendo.com/en/demos/demo_emb_tts.htm
- Belga Acapela: <http://www.acapela-group.com>
- Catalã Verbio Speech Technologies (anterior Atlas): http://www.verbio.com/webverbio3/html/demos_ttsonline.php

▶ Empresas americanas:

- ▶ Cepstral: <http://cepstral.com>
- ▶ At&T: <http://www.research.att.com/~ttsweb/tts/demo.php>
- ▶ Aculab: http://www.mcct.com/aculab_text.html
- ▶ Fonix: <http://www.fonixspeech.com/tts.php>

Empresas que desenvolvem Aplicações Texto-Fala

▶ Multinacionais:

- Siemens:
- Nokia:
- France Telecom:
- Nuance (anterior Scansoft): <http://www.nuance.com/>
- Oddcast http://vhost.oddcast.com/vhost_minisite/demos/tts/tts_example.html
- IBM: http://www-306.ibm.com/software/pervasive/voice_server/demos/?S_CMP=rnav
- Microsoft: <https://research.microsoft.com/speech/tts.asp>

Aplicações

- TTS
 - Ajuda a deficientes visuais (na interface com PC).
 - Interface de surdo-mudo na comunicação (escreve e o sistema TTS converte em fala para o interlocutor).
- ASR
 - Ajuda a ditado para relatório médico (em ambiente de cirurgia).
 - Interface com deficiente auditivo na comunicação em balcão público (sistema converte a fala do balconista em texto para o deficiente ler).

Trabalho 1

1. Instalar o Praat e/ou o SFS
2. Gravar um texto de jornal com cerca de 10 segundos.
3. Fazer a anotação fonética do sinal de fala usando o Praat.

Trabalho 2 (Apenas APS)

Desenvolvimento de um sintetizador de fala em Matlab usando diversos modelos de síntese.

O trabalho deve ter um desenvolvimento teórico em relatório e uma aplicação em Matlab que realize a síntese de pelo menos uma vogal. No relatório deve ser explorada a possibilidade de modificações prosódicas no que concerne às durações e a F0.

Trabalhos:

1. Modelo de formantes
2. Modelo LPC
3. TD-Psola