



Instituto Politécnico de Castelo Branco  
Escola Superior de Tecnologia

Escola Superior de Tecnologia  
Instituto Politécnico de Castelo Branco

# Processamento da Voz Tecnologia da Fala

**Material de Apoio à Disciplina – NÃO OFICIAL**

## Exames e Resoluções 2005 e 2006

Disponível Online:  
<http://pvtf.areasdeservico.com>



Impressão e Scan: [workjunior.com](http://workjunior.com)



Site Web: [areasdeservico.com](http://areasdeservico.com)

|                                                                                   |  |                                                           |                                        |
|-----------------------------------------------------------------------------------|--|-----------------------------------------------------------|----------------------------------------|
|  |  | <b>Processamento da Voz</b><br><b>Tecnologias da Fala</b> | <b>Exame</b><br>27/05/2003<br>14h30min |
| Duração: 2h00min                                                                  |  | Sessão: 1ª                                                |                                        |

Tempo: 8 m

- 1) Explique a razão da difficuldade da análise da fala, através de um simples espectro. (1.5)

há variação de  
distribuição de  
freq. e amplitude  
num certo  
momento no  
tempo

Tempo: 8 m

- 2) Diga quais as vantagens ou desvantagens na utilização de um alfabeto fonético. (1.5)

Vantagens:

Desvantagens

Tempo: 8 m

- 3) A produção da fala é um processo integrado. Comente. (1.5)

Tempo: 8 m

- 4) Quais os elementos do aparelho fonador envolvidos na articulação e Clareza do Som? (1.0)

Lábios, Língua, dentes, faringe, palato.

Tempo: 8 m

- 5) Faça a transcrição Sampa da seguinte frase: "Já estava de férias... se não fosse este exame". (1.0)

Tempo: 10 m

- 6) Diga se a afirmação é falsa ou verdadeira e justifique "O segundo formante (F2), é suficiente para que uma vogal seja perceptível". (2.0)

Sim, o segundo formante é suficiente para q. 1 vogal seja perceptível, pois o 1º e 2º formantes são suficientes para q. a qualidade da vogal, na maioria dos casos, e...

Tempo: 10 m

- 7) Tendo presente a seguinte frase: "Se passar no exame, vou logo de férias"

Indique a palavra ou palavras que contém:

- a) Uma consoante velar,  
 b) Uma consoante bilabial,  
 c) Uma consoante oclusiva: /b/, /d/, /g/, /k/, /p/, /t/,  
 d) Uma consoante fricativa (1.5) /v/, /z/, /ʒ/, /ʃ/, /s/, /ʎ/ → vou, férias
- Nasais: /m/, /n/, /ɲ/ → no  
 Líquidas: /l/, /ʎ/, /ʝ/, /ʒ/, /ʎ/.

Tempo: 10 m

- 8) Classifique as seguintes consoantes quanto ao modo e ponto de articulação. T, S, C, N, V (1.5)

+>

os detalhes importantes são preservados numa forma simples para representar os caracteres importantes da fala

Tempo: 10 m

- 9) O que são, e para que servem espectrogramas? Caracterize um espectrograma "waterfall" (1.5)

um espectrograma é um conjunto de espectros numa fila, visto acima para baixo e onde os eixos do espectro estão representados com pontos escuros no espectrograma.

waterfall - outra forma de se ver a evolução do tempo, da freq. e amplitude q. 3D

Tempo: 10 m

10) O seguinte quadro apresenta as frequências dos formantes de três oradores, a pronunciar a mesma vogal, /i/.

| Orador | F1(Hz) | F2(Hz) |
|--------|--------|--------|
| A      | 340    | 3530   |
| B      | 220    | 2200   |
| C      | 260    | 2760   |

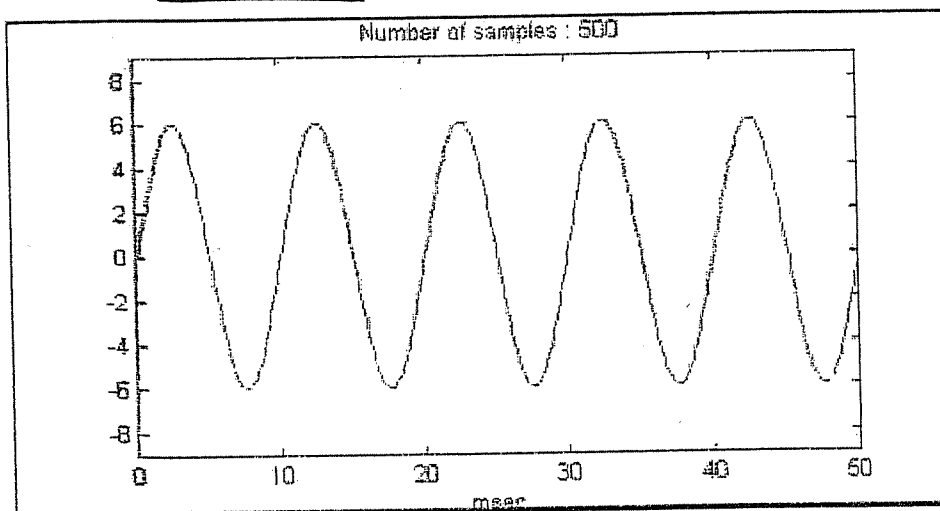
ceiças  
Adulto masculino  
adulto feminino

- As ceiças se ce ã têm as freq. altas e os adultos masculinos
- a) Identifique cada um dos oradores e justifique as suas conclusões. (1.0) Se os 5 têm menores valores
- b) Qual o principal factor que faz com que os valores de F1 e F2 sejam tão diferentes. (1.0) de freq.

Tempo: 10 m

11) Considere a seguinte forma de onda:

- a) De que tipo de forma de onda se trata?
- b) Calcule o período.
- c) Calcule a frequência.
- d) Calcule o comprimento de onda



(1.0)

Tempo: 10 m

12) A síntese da fala envolve um conjunto de operações de processamento linguístico. Diga quais são e qual o objectivo de cada uma (2.0)

- Parsing Sintático
- Interpretação Semântica
- Estrutura de discurso com diátophones - hipótese

Tempo: 10 m

13) O que entende por Prosódia? Qual a sua finalidade num sistema de síntese? Caracterize o Ritmo, Entoação

Stress (2.0)

erros de variação da freq. fundamental



## Processamento da Voz Tecnologias da Fala

Frequência

10/05/2005

LEI/LEII

Duração: 2.00h

Tempo: 8 m

- 1) A fala é um evento contínuo. De que forma é verdadeira esta afirmação? (1.0)

Tempo: 8 m

- 2) O aparelho fonador é constituído por sensores e articuladores. Diga quais as suas componentes e as suas funções. (1.0)

respiratório  
fonador  
articulador

aparelho respiratório  
aparelho digestivo

dentes maxilares  
palato, faringe, língua  
clara e com

Tempo: 8 m

- 3) Que propriedades dos ditongos, é claramente visível nos espectrogramas? Porque é que os espectrogramas dos ditongos têm um aspecto diferente dos monotongos (vogais)? (1.0)

Tempo: 10 m

- 4) O processo de audição e percepção é um processo complexo, se se considerar a natureza dos sinais de voz, e outros factores associados, tais como a continuidade dos sinais de fala, variabilidade, linearidade e ruido. Caracterize cada um destes factores e diga qual a sua contribuição para a complexidade da percepção da fala. (1.5)

Tempo: 6 m

- 5) A predição linear é uma das técnicas de processamento de sinal, e um dos métodos mais poderosos de análise e caracterização de sinais de fala. Na análise de fala por predição linear, tem que se utilizar uma ordem mínima de  $p=8$  (geralmente utiliza-se  $p=10$  ou  $p=12$ ). Explique porquê. (1.5)

preciso de 2 p para cada período

para obter o ganho e constante

Tempo: 10 m

- 6) Faça a transcrição Sampa da seguinte frase: "E a chuva não bate assim!". (1.0)

Tempo: 10 m

- 7) Imagine que está a desenhar um implante para a cóclea, para deficientes auditivos profundos. A miniaturização do sistema é importante, e você quer que o implante transmita informação suficiente para a cóclea. Qual o conjunto mínimo de formantes a serem transmitidos, de modo a que não existam ambiguidades na percepção das vogais por parte do paciente? Justifique a sua resposta. (1.5)

$f_1, f_2$

Tempo: 10 m

- 8) Tendo presente a seguinte frase: "É a chuva não bate assim!"

c h

Indique a palavra ou palavras que contém:

- a) Uma consoante bilabial:  
b) Uma consoante líquida: l r  
c) Uma consoante fricativa: ç z ʃ s → chuva, assim,  
d) Uma consoante oclusiva: (1.5) c b g k p t → bate,

Tempo: 5 m

- 9) Enuncie os métodos que conhece para a detecção de palavras isoladas numa trama de fala. (1.5)

Tempo: 10 m

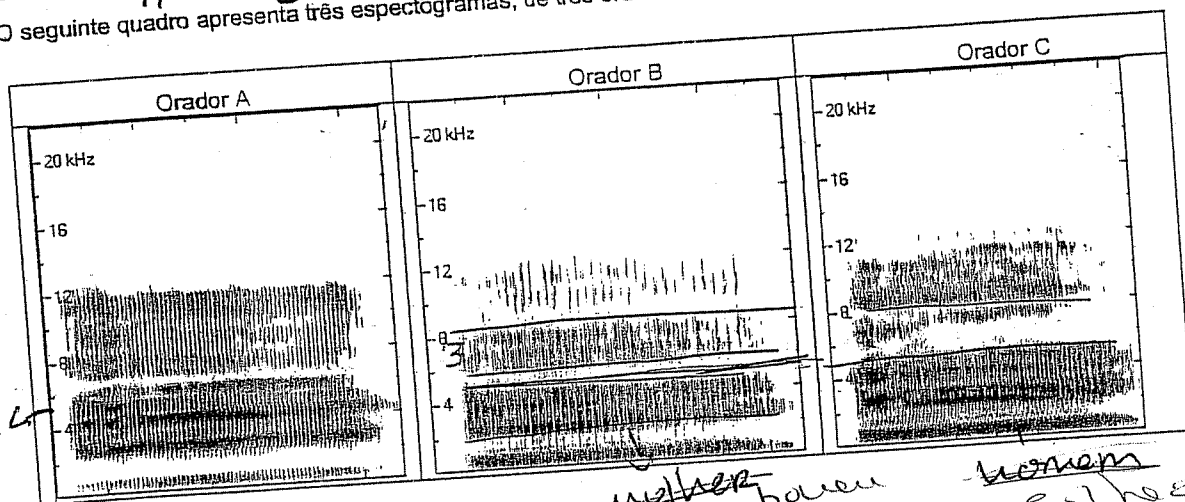
- 10) Um dos métodos de estimação da  $F_0$  e decisão de vozeamento é o método da autocorrelação. Enuncie os fenómenos em que se baseia esta técnica para estimar a  $F_0$  e decidir se uma trama é ou não vozeada. (2.0)

Tempo: 5 m

- 11) Enuncie e explique as técnicas de processamento de sinal, para a identificação do género do orador. (2.0)

Tempo: 10 m

- 12) O seguinte quadro apresenta três espectrogramas, de três oradores, a pronunciar a mesma vogal, /i/.  
crianças → pitch maior  
baixos → homem → pitch baixo



- a) Identifique cada um dos oradores e justifique as suas conclusões. (0.5)  
b) Qual o principal factor que faz com que os valores de  $F_1$  e  $F_2$  sejam tão diferentes. (0.5)  
c) Que valores para  $F_0$  poderiam ser estimados para cada um destes oradores? Justifique a sua resposta. (0.5)

Tempo: 10 m

- 13) O que entende por síntese de fala? Enuncie resumidamente as técnicas de síntese de fala que conhece. (1.5)

gravação de frases.  
texto → som

- 14) Descreva sucintamente como funciona um sistema típico de Text-to-Speech. Faça o esquema da sua arquitectura básica. (1.5)

exemplo de síntese de fala  
texto → voz/som.

Requerida de 10/5/2005

1

1- A fala é um evento contínuo. De que forma é verdadeira esta afirmação?

R. A fala é um evento contínuo humano por natureza e o que a distingue de todas as outras séries vivas para permitir a troca de ideias, expressões de opinião e revelações do nosso pensamento. O estudo da fala envolve alguns aspectos tais como fonética, e mediana, psicologia, física, matemática e a computação. A fala humana normalmente contém contradições. Importa não ter medo e desconfiar e pela própria língua.

2- O aparelho fonador tem a função de produzir sons e é constituído por 2 grupos de órgãos:

O digestivo e o respiratório. Este é constituído pelos órgãos, abridor, Resonadores, Articuladores e Sensores. Os produtores são os pulmões, músculos abdominais, diafragma, músculos intercostais, músculos extensores de coluna. ~~Estes são os órgãos~~ Este tem a função de produzir a coluna de ar que pressuriza - começa produzindo sons nos cordos vocais. O abridor é a língua pois produz a frequência fundamental. Os Resonadores são - cavidade nasal, boca e faringe no seu todo a função de amplificar o som. Os Articuladores são Labios, Língua, Palato mole, Palato duro, Mandíbula. Estes articulam e dão sentido ao som, transformando sons em letras e palavras. Os sensores são o ouvido e o cérebro que tem a função de analisar e captar, selecionar e interpretar o som.

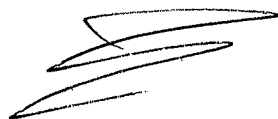
3- A propriedade dos órgãos, que se chamam vibrantes e que têm menor duração que os órgãos respiratórios e maior continuidade são os órgãos vibrantes.

4- Variabilidade e grande as unidades linguísticas, fonemas, letras, sílabas e as palavras no contexto fonético - segmentos de um sinal de voz e tem origem na comunicação, troca de voz, diáfragma, no trato vocal, séptimo e 10.

A Linearidade  
Devido à natureza do processo de produção de voz, não é possível obter um segmento único entre os segmentos linguísticos e as propriedades do sinal acústico. Os órgãos ao produzirem os segmentos fonéticos em série, ao contrário do que sucede no texto escrito. Os segmentos podem sobrepor-se no tempo, em vez disso, o sinal de voz tem pouca probabilidade e pode aparecer distorcido, com ruído sobreposto por outros sons em contextos os ouvintes são capazes de distinguir a informação.

②

@ Suva naiv BATH 621m



b) uma constante ligada - Assim

c) Um wassoer frohen - chov<sup>2</sup>

d) 1. Occlusor - Bate

9) Enumere os métodos que conheço para a detecção de pilhas isoladas num traço de feto. O método que conheço é o (U40) método de Detecção de A. B. dos de 1902, Detecção por pilhas de energia, atingindo da vida de feto a zona de transição de 02, pilhas de energia deturba do por mais 25 ans

10) Para estimar o FO os limites possíveis ~~de~~ das ondas de uma sístole se aproximam de zero e são  
são com 8-20 Hz em gr e freq mais baixas são mais numerosas e as mais altas são  
tipicamente com 2-4 Hz. O FO é ativado com o valor correspondente a 1 Hz sobre o mínimo  
e para definir a ~~base~~ uma taxa e velocidade ~~constante~~ no primeiro e último minuto  
de fusão de Atrialização, normalizada pelo conteúdo do sinal e produzido dentro dos limites de  
frequência possíveis e a largura das ondas com a presença e valor de normal, se não é possível por  
zero e  $\geq 2$

1d) A estratégia para melhorar o reconhecimento e a identificação e verificação da ordem é dividir o espaço do sinal em 2: fornecida por ondas femininas e masculinas. A decisão através da função de nossa discriminação. A frequência do fio se estende entre 50-250 Hz é masculina, se estiver entre 120-300 Hz é feminina e se estiver entre 300 Hz é criança.

12) a) o oscilador é um sistema de ordem 4 com dois ~~graus~~ graus de liberdade e o modo B é um modo flexional  
b) os principais fatores da amplitude do sinal são os modos A e B de 120 e 180 Hz  
modo C ressonância B 50-250 Hz e para o E um regime ~~de~~ excitação de 100 Hz

Qual o que entra no sistema de fala? Enumere as componentes e técnicas de síntese de fala.

A síntese de fala diz respeito ao gerar o sinal de fala por computador e poder compará-lo com o original e categorizá-lo.

- Os que reproduzem fala previamente gravada
- Os que, com base em regras, sintetizam o sinal de fala a partir do texto

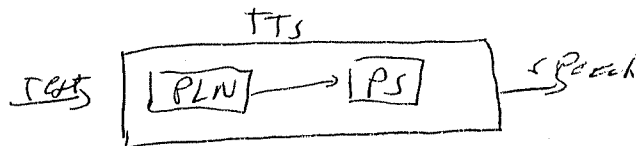
As sínteses de fala se caracterizam por: síntese por produção linear, síntese sinusoidal em que o objetivo é modelar o sinal por sobreposição de sinusóides, síntese de formantes, síntese por concatenação em que existem pequenas palavras pré-gravadas e se são concatenadas numa ordem.

14) O sistema TTS é um sistema modular ou seja ~~é~~ é dividido em vários módulos com funções específicas. Este é substituído pelo o módulo de Processamento de Linguagem Natural e pelo o módulo de processamento de sinal.

O módulo de Processamento de Linguagem Natural tem as funções de: Tokenization, análise morfológica, Eliminação de Ambiguidade, Análise sintática, normalização do texto, Correção do texto em português, S.O. Inglês e Espanhol, etc.

No módulo de Processamento do Sinal podem aplicar 2 abordagens: 1. Sistema baseado em regras ou síntese por concatenação. A síntese baseada em regras é uma forma de síntese de formantes em que modela as frequências das formantes juntamente com as frequências dos sons nos blocos de texto. Quando sons sintetizados periódicos e aperiódicos, tem como modelo natural.

A síntese por concatenação utiliza unidades de som natural de tamanho variável.



Frequência 10/Maio/2005

c.1 1) A fala é o meio de comunicação humana por excelência e o que nos distingue de todos os outros seres vivos. Esta permite a troca de ideias, a expressão de opiniões ou revelações do nosso próprio pensamento. O estudo da fala envolve alguns aspectos da fonética, medicina, psicologia, física, matemática e computação. A fala continua normalmente contém constrições impostas pela taxa e desempenho e pela própria língua.

c.1 (1-7) 2) O aparelho fonador tem a função de produzir sons. É constituído por dois aparelhos: o aparelho digestivo e o aparelho respiratório. Este é constituído pelos produtores, vibrador, ressoadores, articuladores e sensor/coordenador.

Os produtores são os pulmões, músculos abdominais, diafragma, músculos intercostais e músculos extensores da coluna. Tem a função de produzir a coluna de ar que pressiona a laringe, produzindo som nas cordas vocais.

O vibrador é a laringe que produz a frequência fundamental.

Os ressoadores são a cavidade nasal, a boca e a faringe. Têm por função ampliar o som.

Os articuladores são os lábios, a língua, o palato duro e mole e as mandíbulas. Estes articulam e dão sentido ao som, transformando sons em sons orais e sons nasais.

Por último, o sensor/coordenador são o ouvido e o cérebro. Estes captam, seleccionam e interpretam o som.

c.4 3) Um ditongo é a combinação de uma vogal e de uma semi vogal que se pronuncia numa só emissão de voz.

As propriedades dos ditongos que são claramente visíveis num espectrograma é que estes possuem uma menor duração que as vogais respectivas. Os formantes num ditongo alteram a sua frequência à medida que a qualidade sonora do ditongo se altera enquanto que os formantes de uma

vogal são mais ou menos constantes.

#### c3 4) → Processo de Audição e Percepção da fala:

Ouvir é um dos sentidos humanos que transmite a informação auditiva ao cérebro que permite a sua compreensão, ou seja, percepção.

As ondas sonoras produzidas pelos mecanismos articulatórios e transmitida através do ar, são:

- Processadas pelo sistema auditivo
- e decodificadas como mensagem linguística na ligação final da cadeia de fala.

O estudo do processo de percepção da fala é dividido em duas partes:

- processamento Auditivo, que é de natureza psico-fisiológica
- processamento linguístico, baseado na acústica que é codificada na onda sonora.

A continuidade dos sinais de fala confronta o ouvinte com um problema de segmentação visto que no domínio contínuo os segmentos entre as palavras são difíceis de determinar.

A variabilidade ocorre devido às unidades linguísticas abstratas, como os fonemas e as palavras, não corresponderem exactamente a segmentos de um sinal de voz. Este fenómeno tem origem na co-articulação, taxa de voz, diferenças do tracto vocal e sotaques.

A linearidade ocorre visto que devido à natureza do processo de produção de voz, não é possível um mapeamento um-para-um entre os segmentos linguísticos e as propriedades do sinal acústico. Os oradores não produzem os segmentos fonéticos em série, ao contrário do que sucede no texto escrito. Os segmentos produzidos podem sobrepor-se no tempo.

Em relação ao ruído o sinal de voz é transmitido, frequentemente, com pouca qualidade. Este pode parecer distorcido, com ruído ou sobreposto por outras vozes. Contudo, os ouvintes são capazes de distinguir a informação.

c.5 5) cada formante é modelado com um circuito ressonante de 2ª ordem, ou seja, com preditores de 2ª ordem.

Até à banda telefónica (4 KHz) são produzidas no máximo 4 formantes.

A ordem mínima deve ser 2 vezes o número de formantes, ou seja,  $2 \times 4 = 8$ . sendo 8 a ordem mínima, p.

ordens mínimas superiores a 8 têm um ganho constante.

c.4 6) "é a chuva não bate assim"

Transcrição SAMPA: @ 6 56uv6 n6w "bat@ 65w

?7) F1 e F2

c.4 8)

a) consoante bilabial: (p, b, m)

Bate

b) consoante líquida: (l, L, r, R)

c) consoante fricativa: (f, v, s, z, ʃ, ʒ)

chuva, assim

d) consoante oclusiva: (t, p, d, k, g)

Bate

c.5 9) o método que conheço para a detecção de palavras isoladas é o método de detecção de Actividade da Voz (VAD). Esta detecção pode ser efectuada, segundo a proposta de Lori Lamel, através da detecção de pulsos de energia, estimação do ruído de fundo em zonas de ausência de

actividade de voz, e por pulsos de energia distanciados de menos de 75 ms.

c.5 10) A frequência fundamental ( $F_0$ ) é a frequência de vibração das cordas vocais, sendo os limites possíveis de vibração das cordas situados entre os 50 e os 500 Hz. As frequências mais baixas são típicas das vozes graves masculinas e as mais altas são típicas das crianças.

Os métodos de estimação da  $F_0$  e a decisão de vozeamento existentes são:

- Método de autocorrelação
- Método da correlação cruzada normalizada
- Método no domínio da frequência
- Método do Pitch fraccionário

⇒ Método de autocorrelação

Neste método uma trama é considerada vozeada se o primeiro valor máximo da função de autocorrelação, normalizado pela energia do sinal e procurado dentro dos limites possíveis de vibração das cordas vocais apresentar um valor razoável (20,35).

A frequência fundamental é estimada com o valor correspondente a esse valor.

A decisão de vozeamento pode tomar em conta outras características, nomeadamente o número de passagens por zero (se for alta, > 2 passagens/ms, é não vozeada) e a relação entre a energia de alta e baixa frequência.

c.5 11) Devido a diferentes dimensões das cordas vocais os sinais produzidos por oradores femininos correspondem a frequências fundamentais mais elevadas.

uma estratégia para melhorar o reconhecimento e a identificação e verificação do orador é dividir o espaço dos sinais em dois:

- Os produzidos por oradores femininos
- Os produzidos por oradores masculinos

A decisão do tipo de orador é dada através da função de máxima verossimilhança.

Trata-se de um orador masculino se a frequência fundamental se encontrar entre os 50 e os 250 Hz, de um orador feminino se se encontrar entre os 120 e os 300 Hz e de uma criança se for cerca de 500 Hz.

? 12)

a) orador A:

orador B:

orador C:

b) sexo, idade, formas anatómicas, género.

Amplitude do sinal.

c) criança - 500 Hz

orador feminino - 120 a 300 Hz

orador masculino - 50 a 250 Hz

ca 13) A síntese da fala é a conversão do texto em fala, ou seja, é a geração de fala por computador.

Os sistemas de síntese de fala podem agrupar-se em duas categorias:

- os que reproduzem fala previamente gravada
- os que sintetizam o sinal da fala a partir de texto, baseando-se em regras.

As técnicas de síntese de fala que conheço são: síntese por predição linear, síntese sinusoidal, cujo objectivo é modular o sinal por sobreposição de sinusóides, síntese de formantes e síntese por concatenação em que os segmentos fonéticos são concatenados num modelo.

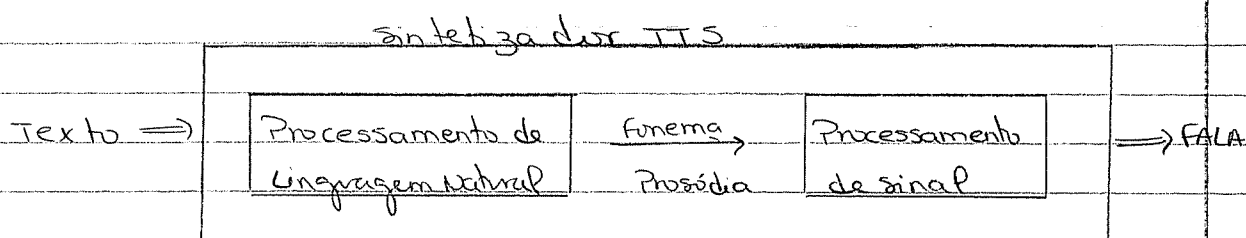
14) o sistema Text-to-Speech (TTS) é um sistema modular, ou seja, possui vários sub-módulos com funções específicas. Este é constituído por dois módulos: o módulo de Processamento de Linguagem Natural (PLN), que produz a transcrição fonética através de vários passos e produz a prosódia, e o modelo de processamento de sinal, que transforma a informação simbólica num sinal de fala audível.

O modelo de PLN possui as funções de tokenization, análise morfológica, eliminação de ambiguidade, análise sintáctica, normalização do texto, conversão texto-fonema, sibilização e duração.

O modelo de processamento de sinal possui duas abordagens:

- Síntese baseada em regras. É uma forma de síntese de formantes. Esta modela as frequências dos formantes juntamente com as frequências dos sons não vozeados, gerando sons sintetizados periódicos e aperiódicos, combinando-os de modo a que se pareçam com sons de fala.

- Síntese por concatenação. Esta é utilizada nas aplicações modernas. Utiliza amostras de som natural de tamanho variável. Pode ser efectuada através da síntese de difones ou por selecção de unidades.





## Processamento da Voz Tecnologias da Fala

Frequência

22/05/2006

Duração: 2:00 h

LEI / LEITI

Tempo: 10 m

- ✓ 1) Nas ciências do processamento da voz e tecnologias da fala, tanto as técnicas de síntese como de reconhecimento, têm por vezes como modelo base o sistema fonador e auditivo humano. Explique a razão deste facto. (1,0)

Tempo: 10 m

- ✓ 2) A produção de fala pode ser vista como uma operação de filtragem, na qual uma fonte de som excita o tracto vocal e/ou o tracto nasal. Os resultados desta operação são observáveis nos espectrogramas. Diga o entende por espectrograma e refira qual a razão pela qual as consoantes não possuem formantes. (1,5)

Tempo: 5 m

- ✓ 3) Quais as propriedades das vogais que as distinguem das glides e das consoantes? (1,0)

Tempo: 10 m

- 4) As unidades linguísticas abstractas, como os fonemas e as palavras, não correspondem exactamente a segmentos de um sinal de voz, o que indica a existência do fenómeno da Variabilidade. Caracterize este fenómeno, indicando e caracterizando os factores em que este tem origem. Dê um exemplo da existência da variabilidade observável num espectrograma. (1,5)

Tempo: 10 m

- ✓ 5) A predição linear é uma das técnicas de processamento de sinal, e um dos métodos mais poderosos de análise e caracterização de sinais de fala. Um dos populares métodos de estimação de formantes é o cálculo das raízes LPC. Identifique no quadro seguinte, quais os valores onde se prevê existirem formantes, justificando a sua resposta.

|   | $F_k$ (Hz) | $B_k$ (Hz) | $r_k$ |
|---|------------|------------|-------|
| → | 200        | 77,56      | 0,97  |
|   | 225        | 150,2      | 0,65  |
| → | 600        | 130,6      | 0,99  |
|   | 1600       | 568,4      | 0,7   |
| → | 2600       | 120        | 0,979 |
|   | 2700       | 199,1      | 0,98  |
|   | 3400       | 908,2      | 0,99  |
|   | 3600       | 130        | 0,85  |

(2,0)

Tempo: 5 m

- ✓ 6) Faça a transcrição Sampa da seguinte frase: "Quer a bebida com ou sem gelo?". (1,0)

↓ ↓ ↓  
h e b e d i d a c o m o u s e m g e l o ?

✓ Tempo: 10 m

- 7) Em certas aplicações é necessário detectar a presença de actividade de voz (VAD). Explique quais técnicas que conhece para a detecção de voz, identificando os factores que dificultam esta tarefa e quais as principais aplicações de VAD. (1,5)

Tempo: 10 m

- ✓ 8) Tendo presente a seguinte frase: "o zé rasgou a camisola!"

Indique a palavra ou palavras que contém:

- a) Uma consoante bilabial; b p m  
b) Uma consoante líquida; l  
c) Uma consoante fricativa; z  
d) Uma consoante oclusiva; (1,5) t, k, g, p, t, k → rasgou, causou, zé

Tempo: 10 m

- 9) Um dos métodos de estimação da F0 e decisão de vozeamento são os métodos no domínio da frequência. Enuncie as formas em que se baseia esta técnica para estimar a F0 e decidir se uma trama é ou não vozeada. (2,0)

Tempo: 10 m

- 10) Em que é que consiste a síntese por concatenação. Faça o esquema da sua arquitectura básica de um módulo de síntese por concatenação. (2,0)

Tempo: 10 m

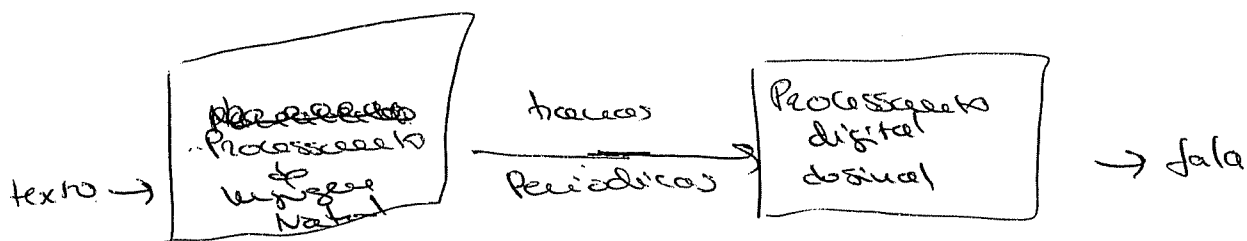
- ✓ 11) Uma das várias técnicas de síntese de fala é a Síntese sinusoidal. Caracterize esta técnica de síntese e indique quais os parâmetros de entrada. (1,5)

Tempo: 10 m

- ✗ 12) Indique e descreva as principais fases de desenho de um típico reconhecedor de fala. (1,5)

Tempo: 10 m

- 13) Escreva uma gramática em JSGF que permita que uma aplicação desenvolvida com a JSAPI possa reconhecer os números em português de 0 a 100. (2,0)



Frequência 22/Maio/06

- c.4 1) Os sinais de fala são compostos por uma sequência de sons ou segmentos fonéticos, regulados pelas regras da língua e pelas características do orador.

Para entender, sintetizar, reconhecer ou processar os sinais de fala é necessário perceber o mecanismo da sua produção.

O aparelho fonador humano é o primeiro bloco na cadeia da comunicação da fala. Este tem a função de produzir sons e é constituído por 2 aparelhos, o aparelho digestivo e o aparelho respiratório.

O aparelho auditivo constituído pelo ouvido, que se trata do órgão sensor do aparelho fonador, tem por função captar o som, localizá-lo e conduzir esse som por forma a que ele seja interpretado.

- c.4 2) Um espectrograma é um conjunto de espectros numa fila, visto de cima para baixo e onde os picos do espectro estão representados como pontos escuros no espectrograma. Este possui os valores de frequência no eixo vertical enquanto o tempo é representado no eixo horizontal. A amplitude é representada pelos pontos escuros.

As consoantes são caracterizadas pela constricção do tracto vocal.

~~MC~~

As consoantes têm tendência a desaparecer.

- c.4 3) As vogais produzem sons normalmente vozeados e produzidos com o tracto vocal numa forma fixa.

Estas têm normalmente uma duração maior que as glides e consoantes e uma melhor definição em frequência.

- c.3 4) A variabilidade tem origem em factores como a co-articulação, a taxa de voz, diferenças no tracto vocal e sotaques. Este fenómeno é caracterizado pelas variações da voz dependendo do contexto.

Ocorre a variabilidade num espectrograma no caso da consoante /K/, este varia a sua frequência e amplitude dependendo do contexto em que está a ser utilizado, por exemplo /ki/ e /ku/.

c.3.5) Prevê-se que existam formantes em:

$$F_k(H_z) = 200;$$

$$F_k(H_z) = 600;$$

$$F_k(H_z) = 2600;$$

visto que estes apresentam uma largura de banda (Bk) suficientemente estreita e o valor  $r_k$  apresentar um valor próximo de 1.

c.4.6) "Ver a bebida com ou sem gelo"

Transição SAMPA: KER 6 bebida com ou sem "Zelu"

7) A detecção de Actividade de voz é razoavelmente simples de detectar caso se esteja num ambiente silencioso, mas em zonas com uma baixa relação sinal-ruído ou na presença de música de fundo esta tarefa fica dificultada.

As aplicações da VAD são:

→ Redução do débito binário na codificação com transmissão de pacotes

→ Reconhecimento de palavras isoladas ou fala contínua

→ Cancelamento de ruído e cancelamento acústico de eco

Para a detecção de voz existe a técnica de detecção de palavras isoladas proposta por Lori Lamel. Esta baseia-se na detecção de pulsos de energia, estimação do ruído de fundo em zonas de ausência de actividade vocal e palavras constituídas por um ou mais pulsos de energia, distanciados de menos de 75 ms.

c.4.8)

a) consoantes bilabiais: (b, p, m)

camisola

b) consoantes líquidas: (l, L, r, R)

rasgou, camisola

c) consoantes fricativas: (v, f, s, z, ʃ, ʒ)

zé, camisola, rasgou

d) consoantes oclusivas: (t, d, g, p, k)

rasgou

c.5 9) os métodos no domínio da frequência baseiam-se numa medida de alinhamento entre harmónicas.

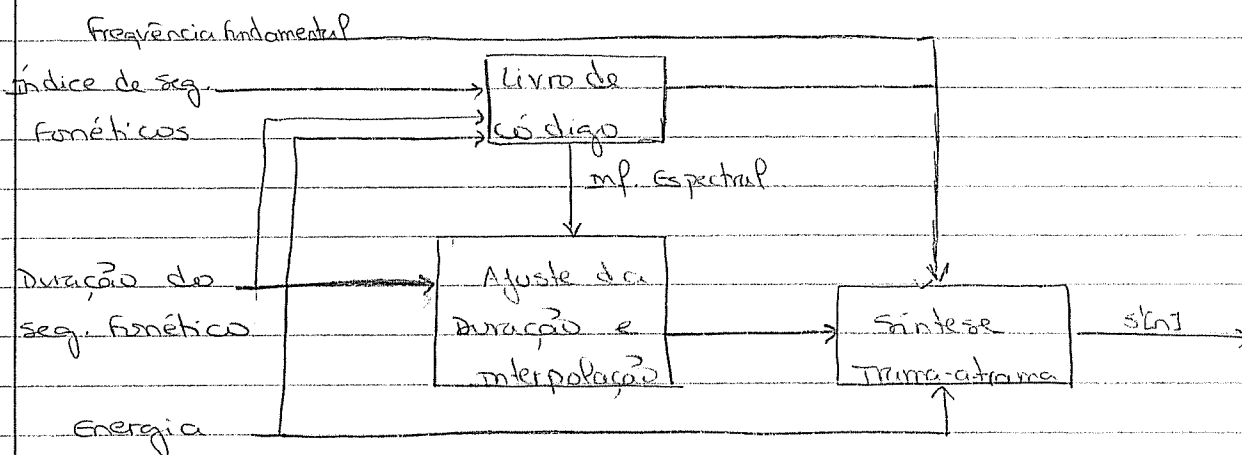
A decisão de vozeamento de uma trama depende da intensidade espectral dada pelo máximo do módulo da Transformada de Fourier Localizada (TFL) no intervalo 200 a 800 Hz.

A estimação de  $F_0$  é efectuada após a decisão do tramo como vozeado e é feita por tentativas, verificando qual dos candidatos tem o melhor alinhamento das suas harmónicas.

c.6 10) A síntese por concatenação consiste na concatenação de segmentos fonéticos pré-gravados num modelo.

uma das abordagens no desenvolvimento destes sistemas é o armazenamento da forma de onda de cada segmento fonético.

outra abordagem utiliza um modelo paramétrico de síntese trama-atrma



c.6 11) os sons vozeados são produzidos quando as cordas vocais vibram produzindo um sinal quase periódico.

Esta quase periodicidade tem implicações no espectro, tornando-o discreto se o sinal for absolutamente periódico. Por outro lado, o sinal torna-se não harmónico quando da diminuição da periodicidade ou nas zonas não vozeadas.

uma trama do sinal pode no entanto ser sempre descrita por uma soma de sinusóides.

sendo conhecidas as frequências e as respectivas amplitudes e fases de cada uma das sinusóides, é possível sintetizar o sinal.

O objectivo da síntese sinusoidal é modelar o sinal por sobreposição de sinusóides.

Tem como parâmetros de entrada frequências, amplitudes e fases nas fronteiras de tramas.

O sinal é modelado por sobreposição de sinusóides, facilitando a interpolação dos parâmetros amostra-a-amostra intra-tramas.

c. 7.12) o reconhecimento de fala é o processo de conversão de fala para texto.

As principais fases de desenho de um típico reconhecedor de fala são:

1) Desenho de gramática

↳ As gramáticas definem as palavras que podem ser faladas por um orador e os padrões pelos quais estas podem ser faladas.

2) Processamento de sinal

↳ Análise das características dos espectros do sinal de áudio de entrada

3) Reconhecimento de fonemas

↳ comparação dos padrões do espectro com os padrões dos fonemas da linguagem a ser reconhecida

4) Reconhecimento de palavras

↳ comparação das seqüências que se parecem com palavras, com padrões de palavras especificadas nas gramáticas ativas

5) Geração de Resultados

↳ extração da informação acerca das palavras reconhecidas.

c. 7.13) # ysgf v.1.0;

grammar grammars numeros;

< unidades > = um | dois | três | quatro | cinco | seis | sete | oito | nove;

< dezenas > = dez | vinte | trinta | quarenta | cinquenta | sessenta | setenta | oitenta | noventa;

< deza > = onze | doze | treze | catorze | quinze | dezasseis | dezassete | dezoito | dezanove;

< centenas > = cem;

public <numeros> = <unidades> | <dezenas> | <deza> | <centenas> |  
( <dezenas> [and] <unidades> );



## Processamento da Voz Tecnologias da Fala

Exame

Época Recurso

11/ 07/ 2005

LEI / LEITI

Duração: 2:00 h

Tempo: 8 m

- 1) Os Sistemas de Processamento Digital da Fala (SPDF) podem ser classificados ou divididos de acordo o tipo de aplicação. Enuncie e caracterize as classificações de SPDF que conheça. (1.0)
- codificação da fala, síntese de fala, reconhecimento de fala, verificação e identificação do orador.*

Tempo: 8 m

- 2) O processo de audição e percepção é um processo complexo, se se considerar a natureza dos sinais de voz, e outros factores associados, tais como a *continuidade* dos sinais de fala, *variabilidade*, *linearidade* e *ruido*. Caracterize cada um destes factores e diga qual a sua contribuição para a complexidade da percepção da fala. (1.5)

Tempo: 8 m

- 3) Quais as propriedades das vogais que as distinguem das *glides* e das consoantes? (1,5)

Tempo: 8 m

- 4) A complexidade da fala leva a que seja necessária uma normalização das diferentes produções de voz. Explique porquê. (1.5)

Tempo: 6 m

- 5) A predição linear é uma das técnicas de processamento de sinal, e um dos métodos mais poderosos de análise e caracterização de sinais de fala. Na a análise de fala por predição linear, quantos formantes se obtêm se utilizar uma ordem  $p=5$ ? Justifique a sua resposta. (1.5)

*2*

*5 = 2 v. 3 formantes*

Tempo: 8 m

- 6) Faça a transcrição Sampa da seguinte frase: "Processamento de voz e tecnologias da fala..". (1.5) ?

Tempo: 8 m

- 7) Imagine que está a desenvolver um quiosque multimédia cujo objectivo é disponibilizar notícias e informações sobre a instituição onde este estiver instalado (Hospitais, Escolas, etc...). De modo a permitir a acessibilidade a deficientes visuais a informação deve disponibilizada também com recurso a voz. Indique qual o sistema de síntese de voz acha que seria o mais adequado. (1.5)

*caracterização*

Tempo: 6 m

- 8) Tendo presente a seguinte frase: "Processamento de voz e tecnologias da fala !" Indique a palavra ou palavras que contém:

Indique a palavra ou palavras que contém:

- a) Uma consoante líquida;
- b) Uma consoante fricativa;
- c) Uma consoante nasal;
- d) Uma consoante oclusiva; (1.5)

Codificação da fala - Representação eficiente do sinal  
com vista à sua transmissão e armazenamento,  
mantendo a qualidade acima da exigida pela aplicação.

Síntese da fala - parte de onde é gerado o sinal  
acústico da fala a partir de um texto.

Vantagens - Diminuição da quantidade de memória usada no  
armazenamento dos sinais respectivos.

- Possibilidade de minimizar o espaço de adaptação do  
sistema a novos textos, não sendo necessário re-gerar novas  
fases

→ 2 Problemas

→ conversão de texto em representações linguísticas (segmentos fonéticos)  
respeitando duração, conteúdo de energia e da frequência de vibração  
dos cordos vocais.

→ Síntese da onda acústica a partir da representação linguística.

Reconhecimento da fala -

Desempenham o papel do ouvido, convertendo uma onda acústica  
numa msg escrita

Tempo: 8 m

- 9) As frequências dos formantes e respectivas larguras de banda são alguns dos parâmetros espectrais que se podem extrair do sinal de fala. Enuncie e caracterize os métodos de estimação de formantes que conheça (1.5)

Coeficientes de espectro  
Raízes do filtro de predição linear  
Coeficiente LSF

Tempo: 8 m

- 10) Enuncie e caracterize os métodos de estimação da F0 e decisão de vozeamento, que conheça. (2.0)

Tempo: 8 m

- 11) Em certas aplicações é necessário detectar a presença de actividade de voz (VAD). Explique o funcionamento do método proposto por *Lori Lamel* para a detecção de palavras isoladas, identificando os factores que dificultam a VAD, e as principais suas principais aplicações. (2.0)

Tempo: 8 m

- 12) A maioria dos modelos de síntese utiliza ainda hoje a estrutura fonte-filtro na modelação dos sinais de fala (utilizada já nas primeiras experiências de sintetizadores de voz), em que o conjunto de filtros é excitado por ruído nas zonas não vozeadas e por pulsos periódicos nas zonas vozeadas modelando a abertura e fecho da glote. Explique qual razão da utilização de diferentes tipos de excitação para os filtros. (1.5)

Tempo: 8 m

- 13) O teste é visto como uma das áreas chave na investigação e desenvolvimento de síntese de fala, uma vez que é necessário para a avaliação e comparação de novos desenvolvimentos, e preferências dos utilizadores. Descreva os principais desafios relativos ao teste de sintetizadores de fala, e caracterize os principais tipos de teste que podem ser efectuados. (1.5)

- testes sequenciais

q2 não existe uma norma definida

tipos

Standard

Free

Procedural

## Exame Recurso 11/Julho/2005

c.2 1) Os sistemas de Processamento digital da voz encontram-se divididos em:

- Codificação de fala
- Síntese de fala
- Reconhecimento de fala
- Verificação e Identificação do orador

Codificação da fala  $\Rightarrow$  Aplicação antiga do processamento digital da fala, consiste na representação eficiente do sinal com vista à sua transmissão ou armazenamento, mantendo a qualidade acima da exigida pela aplicação.

Síntese de fala  $\Rightarrow$  consiste na síntese de fala a partir de texto, onde é gerado um sinal acústico de fala a partir de um texto. Tem como principais vantagens a diminuição da quantidade de memória utilizada no armazenamento dos sinais respectivos e a possibilidade de minimizar o esforço de adaptação do sistema a novas tarefas, não sendo necessário pré-gravar novas frases. Estes sistemas são implementados dividindo o problema em dois:

- $\rightarrow$  conversão de texto em representação linguística
- $\rightarrow$  síntese da onda acústica a partir da representação linguística.

Reconhecimento de fala  $\Rightarrow$  Desempenham o papel de ouvinte, convertendo uma onda acústica numa mensagem escrita ou equivalente, ou identificando um comando falado.

Verificação e Identificação do orador  $\Rightarrow$  um sistema de verificação do orador deve decidir se um orador é quem ele ma ser. Deve decidir qual dos oradores, dentro de um conjunto de oradores, produziu determinada frase.

c.3 2) A continuidade dos sinais de fala confronta o ouvinte com um problema de segmentação visto que no domínio contínuo os segmentos

entre as palavras são difíceis de determinar.

A variabilidade ocorre devido às unidades linguísticas abstractas, como os fonemas e as palavras, não corresponderem exactamente a segmentos de um sinal de voz. Este fenómeno tem origem na co-articulação, taxa de voz, diferenças do tracto vocal e sotaques.

A linearidade ocorre devido à natureza do processo de produção de voz, não é possível um mapeamento um-para-um entre os segmentos linguísticos e as propriedades do sinal acústico. Os oradores não produzem os segmentos fonéticos em série, ao contrário do que sucede no texto escrito. Os segmentos produzidos podem sobrepor-se no tempo.

Em relação ao ruído o sinal de voz é transmitido, frequentemente, com pouca qualidade. Ele pode aparecer distorcido, com ruído ou sobreposto por outras vozes. Contudo, os ouvintes são capazes de distinguir a informação.

c.4 3) As vogais produzem sons normalmente vozeados e produzidos com o tracto vocal numa forma fixa.

Estas têm normalmente uma duração maior que as glides e consoantes e uma melhor definição em frequência.

c.3 4) As pessoas têm dificuldade que compreender a fala de vários e diferentes oradores. Estes podem variar em idade, sexo, dialecto, pronúncia, sotaque, etc.

contudo, os ouvintes têm capacidade de sintonizar as diferentes formas acústicas e de as normalizar, o que nos leva ao Problema de Normalização.

Se observarmos o espectrograma de um orador masculino e outro de um orador feminino, verificamos que estes são bastante distintos. Isto leva a que seja necessária uma normalização das diferentes produções de voz dos dois oradores.

c.5 5)  $p = 5$

$$p = 2 \times \underbrace{n^{\circ} \text{ formantes}}_{\text{for}} \leq 1 \quad \bar{5} = 2 \text{ for } \leq 1 \quad \text{for} = \bar{5} \leq 1 \quad \text{for} = 2,5 \quad \text{for} = 2$$

cada formante é modelado com um circuito ressonante de 2ª ordem, ou seja, com preditores de 2ª ordem.

Até à banda telefônica (4 kHz) são produzidos no máximo 4 formantes.

A ordem mínima deve ser 2 vezes o número de formantes. Ordens mínimas superiores à prevista têm um ganho constante.

Dai que para termos uma ordem mínima igual a 5 serem necessários 2 formantes, como calculado anteriormente.

c. 1. 6) "Processamento de voz e tecnologias da fala.

Transcrição SANPA: processamento de "voz e "tecnologias da  
"fala

c. 5. 7) sistema de síntese por concatenação.

c. 4. 8)

a) consoante líquida: (l, L, r, R)

Processamento, tecnologias, fala

b) consoante fricativa: (f, v, s, z, ʃ, ʒ)

voz, tecnologias, fala, processamento

c) consoante nasal: (m, n, ɲ)

Processamento, tecnologias

d) consoante oclusiva: (p, b, t, d, k, g)

processamento, de, tecnologias, da

c. 5. 9) os métodos de estimação dos formantes que congeço são:

→ Máximos do espectro

→ Raízes do filtro de predição linear

→ coeficientes LSF

### → Método dos Máximos do Espectro

As frequências dos formantes podem ser estimadas através dos máximos locais do espectro de curta duração, calculado através da transformada de Fourier localizada (TFL) ou pela envolvente espectral produzida por predição linear.

Este método tem por desvantagens o facto de ser um método pouco preciso, pois a energia do espectro está concentrada em múltiplos da frequência fundamental. Para vozes femininas e crianças, este problema agrava-se se por acaso a frequência da formante se situa perto de um múltiplo de  $F_0$  e quando a largura de banda do formante é menor que  $F_0$ , esta pode não incluir nenhuma harmónica.

### → Método das Raízes dos Filtros LPC

Este envolve a análise das raízes do denominador do filtro de predição linear.

Uma raiz ( $f_k$ ) pode ser identificada como formante se apresentar uma largura de banda ( $B_k$ ) suficientemente estreita, e se  $R_k$  apresentar um valor próximo de 1.

### → Métodos dos Coeficientes LSF

os coeficientes LSF são atraídos pelos formantes, coeficientes LSF próximos podem corresponder a formantes.

c. 10) os métodos de estimação de  $F_0$  e decisão de vozeamento que conheço são:

- Método de autocorrelação
- Método de correlação cruzada normalizada
- Método no domínio da frequência
- Método do Pitch fraccionário

### → Método de Autocorrelação

Neste método uma trama é considerada vozeada se o primeiro valor máximo da função de autocorrelação, normalizada pela energia do

do sinal e procurado dentro dos limites possíveis de vibração das cordas vocais, apresentar um valor razoável.

A frequência fundamental é estimada com o valor correspondente desse valor máximo.

A decisão de vozeamento pode tomar em consideração outras características, tais como o número de passagens por zero, se este for superior a 2 passagens/ms a trama é não vozeada, e a relação entre as energias de alta e baixa frequência.

Este método possui alguns inconvenientes nomeadamente o facto de as janelas terem dimensão pelo menos duas vezes o máximo período e a vibração das cordas vocais estar abaixo dos 500 Hz.

→ Método de correlação cruzada normalizada

Este método ultrapassa as dificuldades do método anterior. Esta correlação pode ser calculada sobre o sinal original.

→ Métodos no domínio da frequência

Estes métodos baseiam-se numa medida de alinhamento entre harmónicas.

A decisão de vozeamento depende da intensidade espectral dada pelo máximo do módulo da Transformada de Fourier Localizada (TFL) no intervalo 200 a 800 Hz.

A estimação de  $F_0$  é feita por tentativas, após decidida a trama como vozeada, verificando qual o candidato que tem o melhor alinhamento das suas harmónicas.

c.5 11) o método proposto por Lori Lamel para a detecção de palavras isoladas baseia-se na detecção de pulsos de energia, na estimação do ruído de fundo em zonas de ausência de actividade de voz. Palavras constituídas por um ou mais pulsos de energia, distanciados de menos de 75 ms.

A detecção de Actividade de voz (VAD) é razoavelmente simples caso se esteja num ambiente silencioso, mas em zonas de mais baixa relação sinal-ruído ou com música de fundo esta

tarifa é dificultada.

A VAD aplica-se na redução de débito binário na codificação com transmissão de pacotes, no reconhecimento de palavras isoladas ou fala contínua e no cancelamento de ruído e cancelamento acústico do eco.

12) Não Demos!!

13) Não existem métodos normalizados, para além dos poucos testes segmentais, os testes mais desenvolvidos referem-se ao nível segmental, ou a medir apenas a inteligibilidade.

A inteligibilidade não fornece muita informação acerca da aceitabilidade e naturalidade.

Os métodos de teste segmentais tratam-se de testes de inteligibilidade de um simples fonema/segmento. tipicamente testa-se apenas a inteligibilidade das consoantes com palavras.

É económico visto ser fácil de desenhar e de administrar. existem 3 tipos de testes que podem ser efectuados:

→ Teste segmental standard

- Listas de sequências sem sentido

- Testar todas as consoantes com /a/, /i/, /u/

→ Teste de frases

- Um conjunto de 100 frases para testar a inteligibilidade de palavras no contexto de uma frase.

- Precisão, com base no contexto

→ Teste de Prosódia

- Não existem métodos standard

- Geralmente fazem-se experiências de classificação

A avaliação da qualidade geral é feita com base na pontuação obtida por opinião e na taxa de subjectividade.

A avaliação final depende do ouvinte.

um espectrográfico é um conjunto de espectros numa fila, visto de cima para baixo e onde os eixos ~~do~~ do espectro estão representados com pontos curvos no espectro. O espectrográfico é visto para se ver a transição tempo, de freq. e amplitude a 3 dimensões. ~~A amplitude~~ O espectrográfico é visto diagonalmente com o tempo e a freq. nos eixos de tempo. A amplitude é vertical para cada espectro e o formato pode se ver como pontos de frequência alta e um espectro

As it can be seen, the process of perception is a complex process involving the analysis of the input and the synthesis of the output. The main components of the process are: the input, the analysis, the synthesis, the output, and the feedback.

if A, gl. to semi-vowels & poken like *isto*, como vogais mas com maior centralização e menor divergência que as vogais abertas. A principal diferença de consonantes é ~~comp~~ de ~~sem~~ ~~vogais~~ semi-vogais. As vogais têm um longo morado, e as gl. & consoantes com o mesmo ~~tempo~~ ~~de~~ ~~freq.~~ de freq.

1) A observabilidade é dada as unidades linguísticas abstratas, como as frases e as palavras em correspondência com os segmentos de um sinal de voz. Este fenômeno tem origem

- localizável - diferenças no tempo ou no espaço e na ~~estrutura~~ que pode variar dependendo do contexto
- transitiva

As freq das formantes e respectivas larguras de banda são alguns parâmetros importantes para poder extrair do sinal de fala. Indique e caracterize os métodos de estimação de formantes que conhece.

Os métodos que se conhece são: método das raízes do espectro - onde as freq podem ser estimadas através das raízes dos polos do espectro nos eixos complexos

Raízes do polinômio de predição linear no qual envolve uma análise do numerador do denominador do filtro. Uma raiz pode ser interpretada como uma formante se estiver no eixo real e ~~for~~ banda suficientemente estreita.

Coefficients LSF - os coeficientes <sup>LSF</sup> ~~LSF~~ são usados para estimar as formantes  
os <sup>LSF</sup> ~~LSF~~ podem corrigir as formantes

0) Os métodos ~~que~~ ~~em~~ ~~que~~ se baseiam no domínio de freq ~~para~~ ~~baseiam-se~~ ~~para~~ ~~na~~ ~~medida~~ ~~de~~ ~~distinção~~ entre harmônicos para a análise de voz e estimação de F0

1) As técnicas de análise para a detecção de voz são: Detecção de Ritmo de energia, Estimação do nível de fundo em zonas de ausência de atividade de voz e como a fala é caracterizada por um ou mais ~~pequenos~~ pulsos de energia distribuídos de menos 75ms. Os fatores são difíceis de obter se a energia for muito baixa. Uma técnica difusa é a zona de baixa energia sinal-voz. Os métodos de fundo. As principais aplicações são: Redução do ruído de fundo na codificação de TTX de pontos. Reconhecimento de palavras isoladas e fala contínua e cancelamento de ruído e cancelamento de eco, para o qual é necessária uma boa estimativa do ruído.

12)

a)

Ordem A - mulher

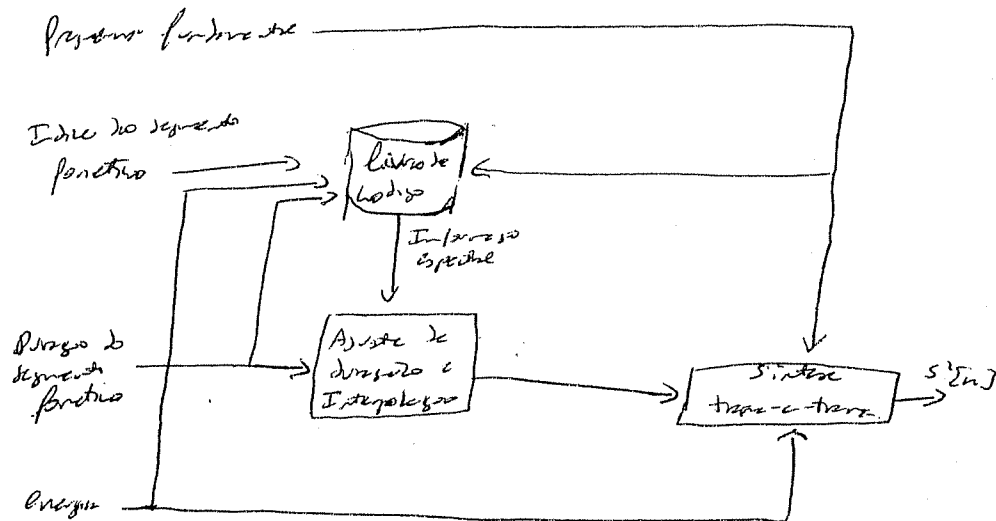
Ordem B - homem ou mulher

Ordem C - criança

b) e amplitude

13) A síntese de fala tem respeito à geração de fala por computador. Este tipo de sistema pode ser de 2 categorias: os que reproduzem fala previamente gravada e os que usam uma base de regras, sintetizam o sinal de fala a partir do texto. O modelo LPC utiliza um modelo simplificado da produção de fala. Este ~~modelo~~ modelo obtém o sinal através de um filtro de predição linear excitado se tiver muito poucos polos sem ruído. Se houver um grande número de polos com a presença de ruído como as vibrações das cordas vocais por ser vozeadas.

A síntese por contagem condizente de condutores segmentos por-gradus  
 em modelo. Concatenado



11/07/05

ordenar linha

5)  $2 \times n^2$  de fontes

ordenar  $p = 5$

Passagem de  
7/5  
inicial

25 = 2 x uc fontes

$\frac{25}{2}$  = uc fontes  
 $\approx 3$  fontes

P - Prolego

Acabou trabalhar 8 e 12, onde há  
 pouco trabalho.

## Exame 14/junho/2005

- 1- Diga o que entende por espectograma e quais as vantagens e desvantagens da utilização do espectograma waterfall para a extração das características de um sinal de voz?
- 2- Os modelos de percepção da fala são teorias sobre os diferentes aspectos relativos aos mecanismos de percepção da fala. Estas teorias podem ser classificadas como activas e passivas. Caracterize cada uma destas teorias dando exemplo de técnicas em cada uma destas classificações.

### Respostas

- c.1 1) um espectograma é um conjunto de espectros numa fila, visto de cima para baixo e onde os picos do espectro estão representados com pontos escuros no espectograma.
- O espectograma waterfall é outra forma de se ver a trama do tempo, da frequência e amplitude a três dimensões.
- O espectograma é visto diagonalmente com o tempo e a frequência nos eixos de baixo. A amplitude é visível para cada espectro.
- Os formantes podem ser vistos como pontos sequenciais altos em cada espectro.
- c.2 2) As Teorias Activas postulam que o processo de percepção da fala envolve aspectos da produção de fala. Os sons são captados e analisados pelas suas propriedades fonéticas com base no modo em que são produzidos. O reconhecimento é baseado nas correspondências entre as unidades captadas e processadas e a sua representação interna.
- As Teorias Passivas enfatizam os mecanismos de captação e filtragem e o mapeamento directo da mensagem com as características acústicas e fonéticas da linguagem. Relegam o papel da produção da fala para segundo plano, sendo apenas utilizado em situações mais complicadas.

como exemplo de técnicas activas temos "the Motor Theory," que se baseia nos gestos articulatórios, "Análise por síntese" que se baseia nas características acústicas, e "Hyper and Hypo speech".  
como exemplo de técnicas passivas temos "Fant's auditory model" que se baseia nas características acústicas, e "the quantal nature of speech", que se baseia na quantidade de fala.