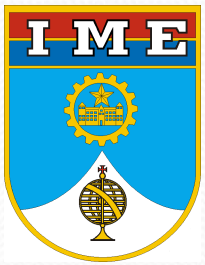


RECONHECIMENTO AUTOMÁTICO DE LOCUTOR E SUA APLICAÇÃO EM FONÉTICA FORENSE

José Antonio Apolinário Jr.
Dirceu Gonzaga da Silva

X Congresso Nacional de Fonética e Fonologia
IV Congresso Internacional de Fonética e Fonologia

Sumário

- Introdução
  - Laboratório de Voz do IME
  - InnoVox – Processamento de Áudio e Voz
- Técnicas Atuais
  - Características e Classificadores
  - Compensação de Canal
  - Avaliação
- Combinando Reconhecimento de Voz e de Locutor
- Conclusão

Laboratório de Voz do IME

- O IME tem uma longa tradição na pesquisa de processamento de voz (primeira tese em 1977)
- Temos hoje um total de 22 teses na área de voz (16 anteriores a 2001 e 6 com os professores atuais)
- Desde 1985 o reconhecimento de locutor vem sendo estudado; mas somente após 1996 focamos no caso de sinal independente do texto

Laboratório de Voz do IME

- Desde 2001 estamos particularmente interessados no problema de verificação robusta de locutor
- Nossa pesquisa tem contemplado a solução de problemas típicos de verificação de locutor independente do texto (compensação de canal, tratamento do ruído aditivo, etc.)
- Desde 2006 a pesquisa tem sido orientada para o problema da perícia fonética – já tivemos 2 dissertações de mestrado abordando o tema.

Laboratório de Voz do IME

- Grupo de trabalho atual:
  - Prof. José Antonio Apolinário Jr. , DSc
  - Dirceu Gonzaga da Silva (doutorando da PUC-Rio)
- Colaboradores:
  - Prof. Roberto Miscow , M.C.
  - Prof Edson Cataldo, Dr. (UFF)
  - Simone Aiex - Fonoaudióloga
- Alunos atuais
  - Daniel Nicolalde – Mestrado → (Edição de Áudio)
  - 1 aluno UFF – Mestrado → (envelhecimento da voz)
  - 1 aluno Iniciação Científica → Scrambler

Laboratório de Voz do IME

- Outros temas de pesquisa desenvolvida pelo grupo no IME
  - Reconhecimento de voz
  - Criptofonia (temporal e frequencial)
  - Criptoanálise (temporal com 2 dissertações e 1 frequencial)
  - Filtragem adaptativa e suas aplicações
  - Processamento de sinais em geral
  - Processamento de sinais em arranjos de sensores

- Parcerias:
  - InnoVox – Processamento de Áudio e Voz

InnoVox - Processamento de Áudio e Voz

- *Surgiu em maio de 2008*

  A partir de um grupo de pesquisa do Laboratório de Processamento de Voz da Seção de Engenharia Elétrica do Instituto Militar de Engenharia.

- *Quem Somos*

  A Innovox se propõe a desenvolver sistemas que envolvam processamento de sinais de voz e áudio, tais como reconhecimento de locutor e voz, criptofonia, programação em DSPs para sistemas embarcados, sistemas de transmissão digital e reconhecimento de áudio em geral.

Interesses comuns com a Innovox

- Afinidade de interesses (ex-alunos)
- Intercâmbio científico
  (co-orientações, publicações científicas conjuntas)
- Desenvolvimento de projetos conjuntos
- Somos clientes e consultores
- Exemplos:
  - SisVAL
  - Sistema de apoio à perícia fonética
  - Sistema de detecção e determinação de direção de tiro de arma de fogo

Técnicas Atuais de RAL

Informação da Voz

Reconhecimento de Voz

Reconhecimento de Idioma

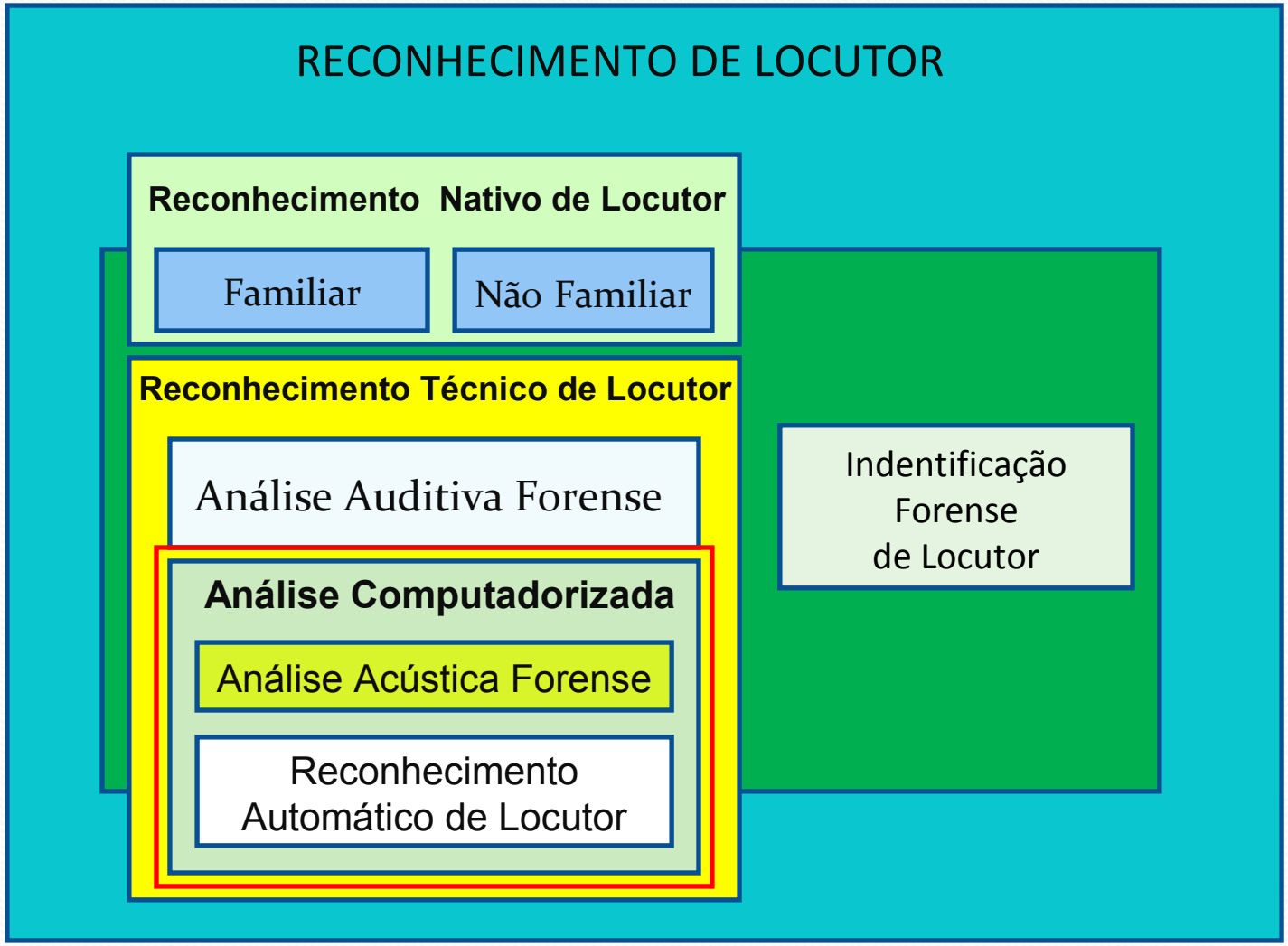
Reconhecimento de Locutor

Palavras
"Como vai você"

Idioma
"Português"

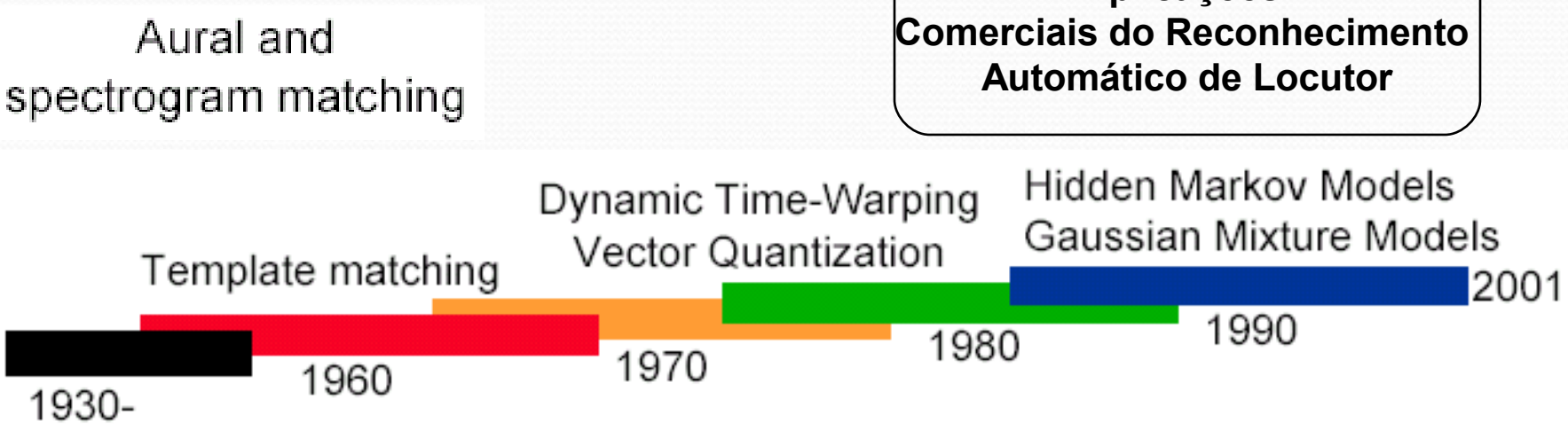
Locutor
"João"

Formas de Reconhecimento de Locutor



Evolução do RAL

Aplicações Comerciais do Reconhecimento Automático de Locutor



Banco de dados pequeno, sinal limpo, voz controlada → **Banco de dados grande, sinal em ambiente não controlado.**

Principais Aplicações

Controle de Acesso

Autenticação para transações

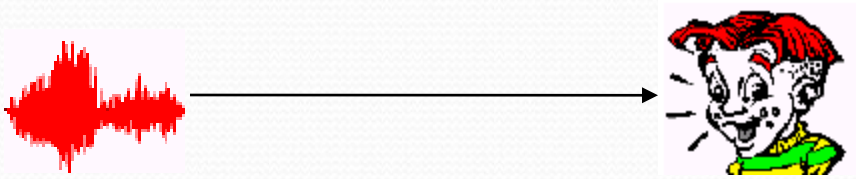
Aplicações Forenses

Principais Tarefas

Identificação

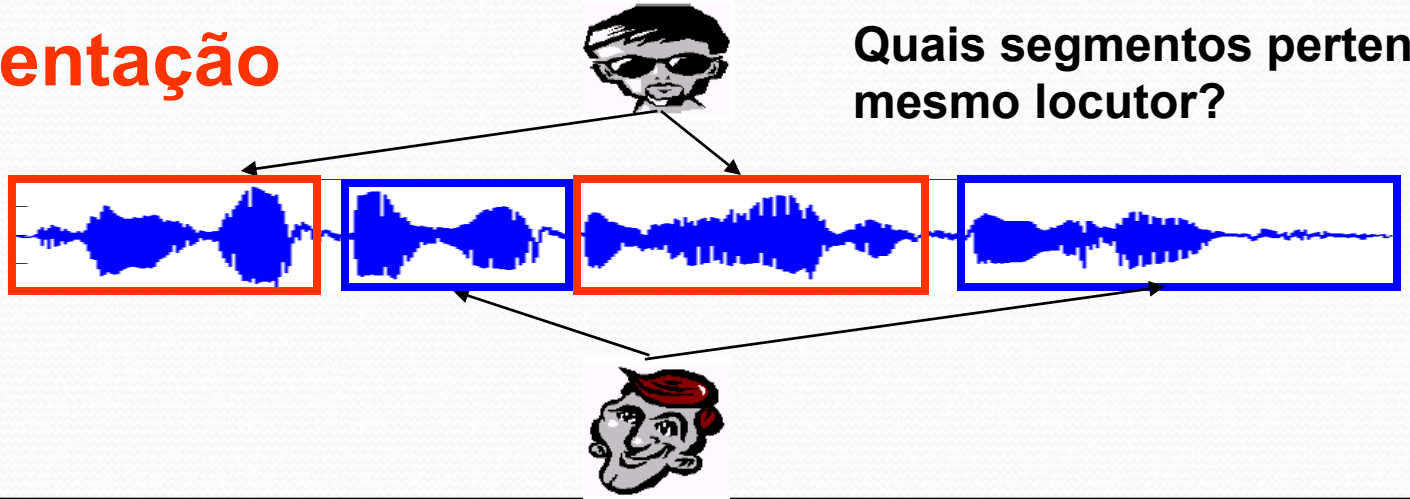
De quem é esta voz?

Verificação



Esta voz é de João?

Segmentação



Quais segmentos pertencem ao mesmo locutor?

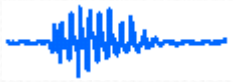
Modalidades

- Reconhecimento Dependente do Texto
  - Ex: frases fixas, reconhecimento por dígitos.
  - Utilizado em aplicações onde se tem controle da entrada
- Reconhecimento Independente do Texto
  - Ex: conversa livre, frases aleatórias
  - usado em aplicações onde não se tem controle das entradas
  - Sistema mais flexível porém envolve um problema mais difícil

Fases de um Sistema de RAL

Fase de Treinamento

João



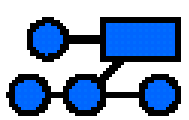
Pedro

Extração de Características

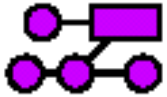
Treinamento do Modelo

Modelo para cada locutor

João



Pedro



Fase de Verificação

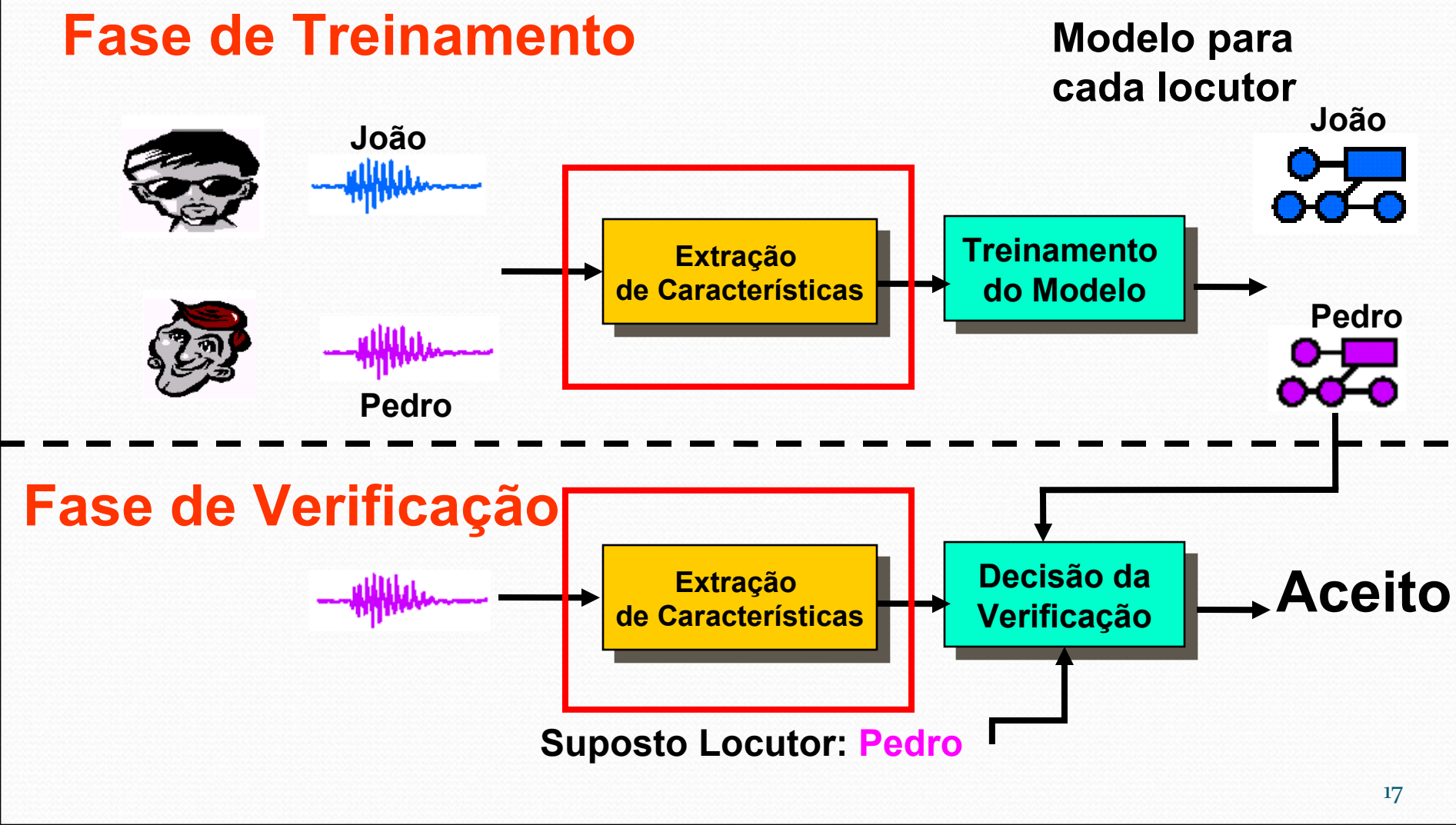
Extração de Características

Decisão da Verificação

Aceito

Suposto Locutor: Pedro

Fases de um Sistema de RAL



Hierarquia das Características para RAL

Atributos Desejáveis das Características

Prática

- Ocorre naturalmente e de forma freqüente
- Fácil de se medir

Robusta

- Não mude com o tempo e não sujeita a condições de saúde dos locutores;
- Não ser afetada por ruído ambiente nem seja dependente de um canal de transmissão

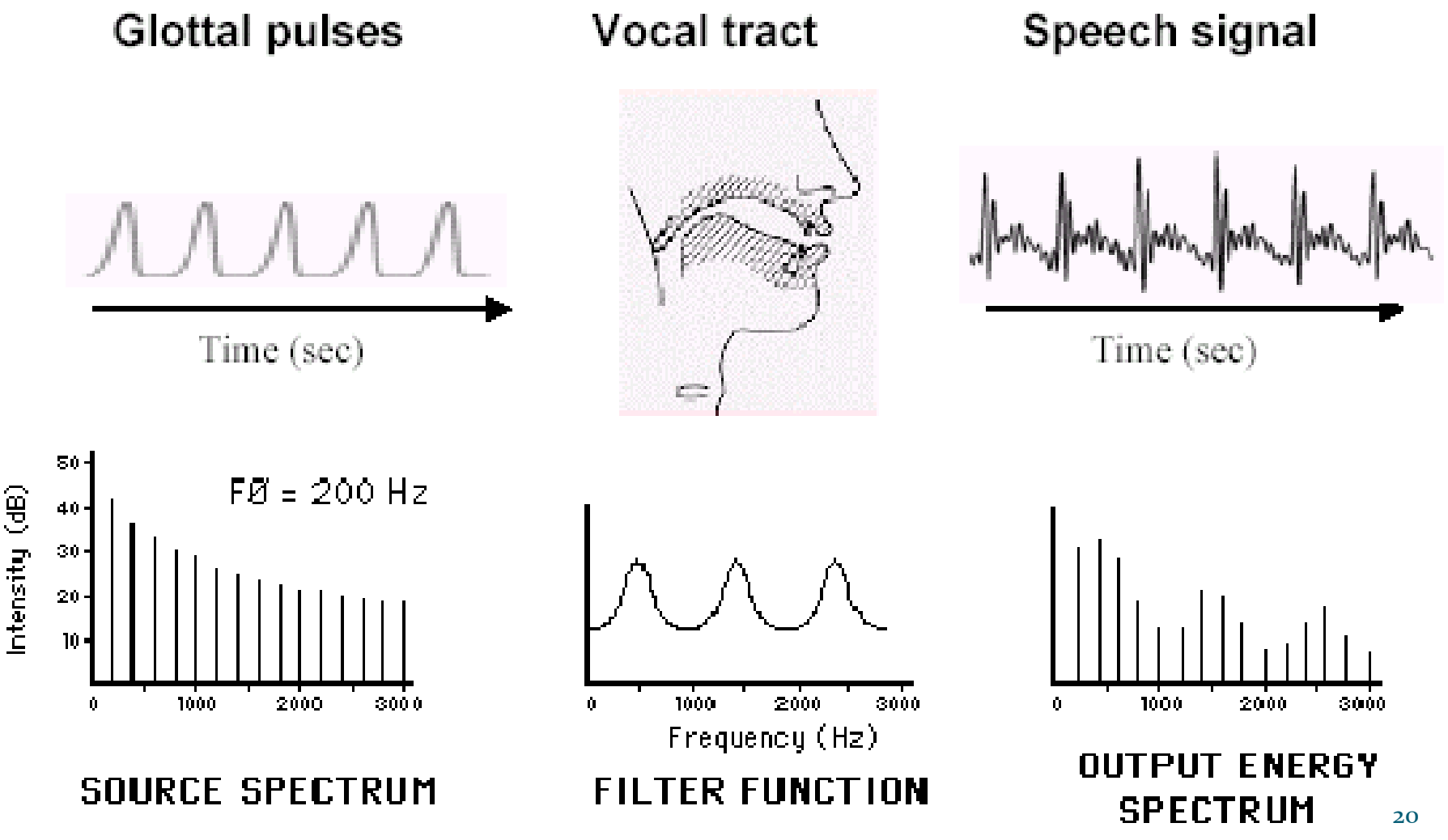
Segura

- Não ser sujeita a mímico

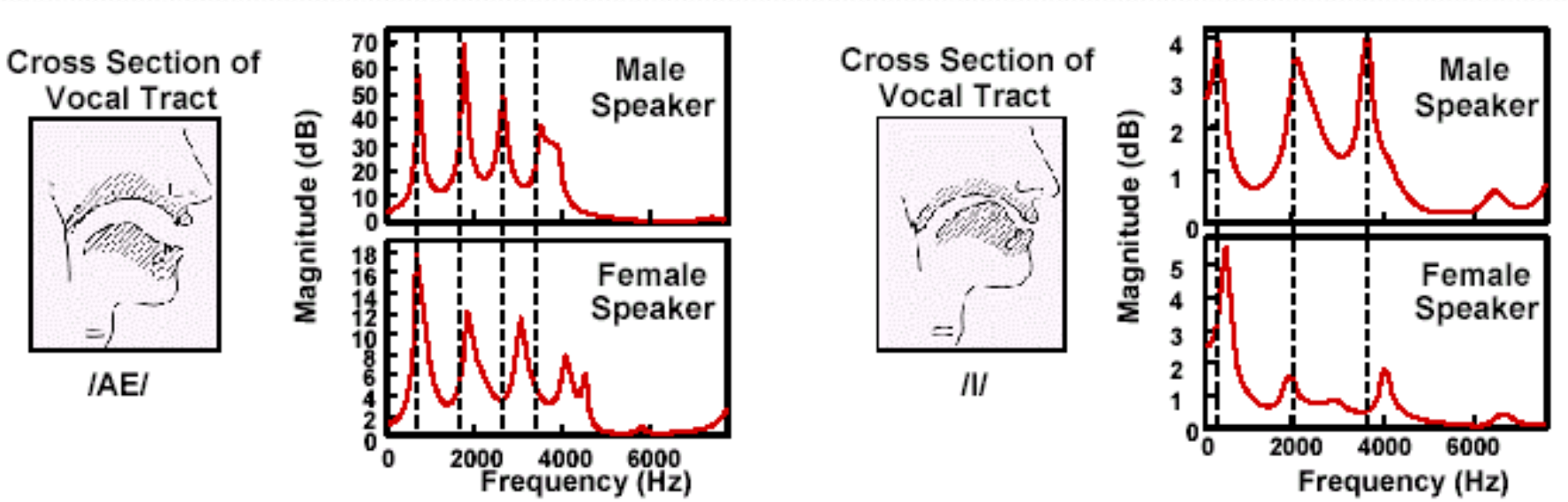
Nenhuma característica possui todos estes atributos

Características Acústicas tem obtido melhores resultados

Modelo de Produção da Voz



Características para RAL

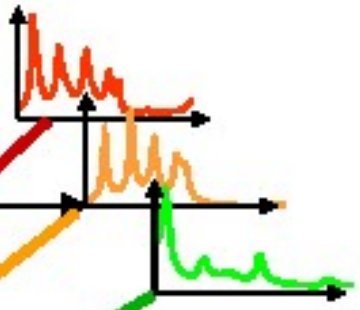


- **Diferentes locutores terão espectros diferentes para sons similares**
- **As diferenças estão na localização e no módulo dos picos do espectro**
  - Os picos são conhecidos com formantes e representam as freqüências ressonantes do trato vocal

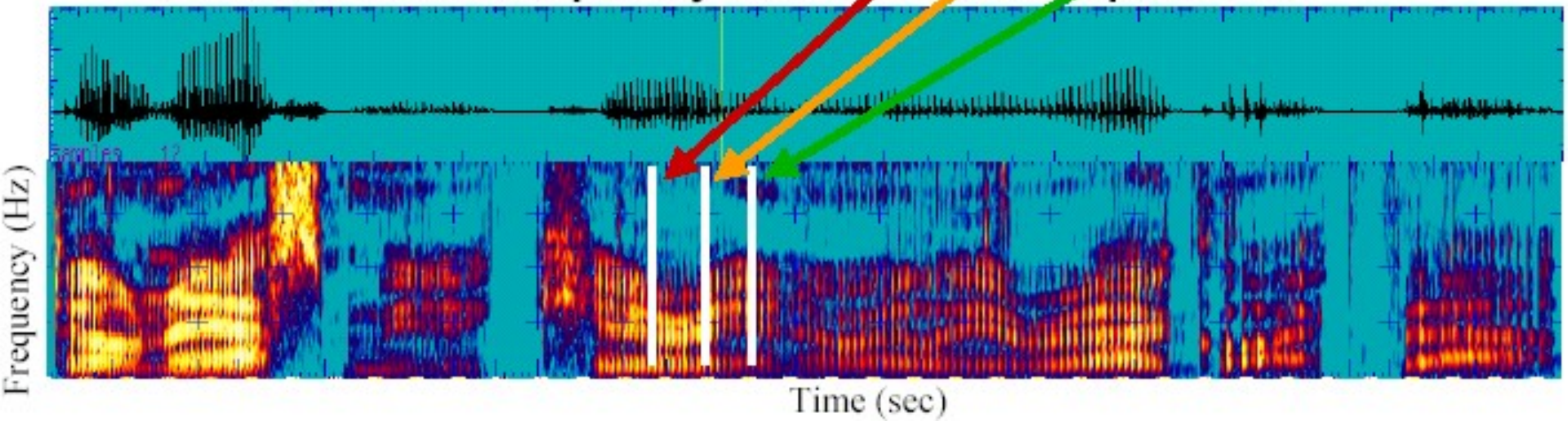
Características para RAL

Fourier Transform

Magnitude

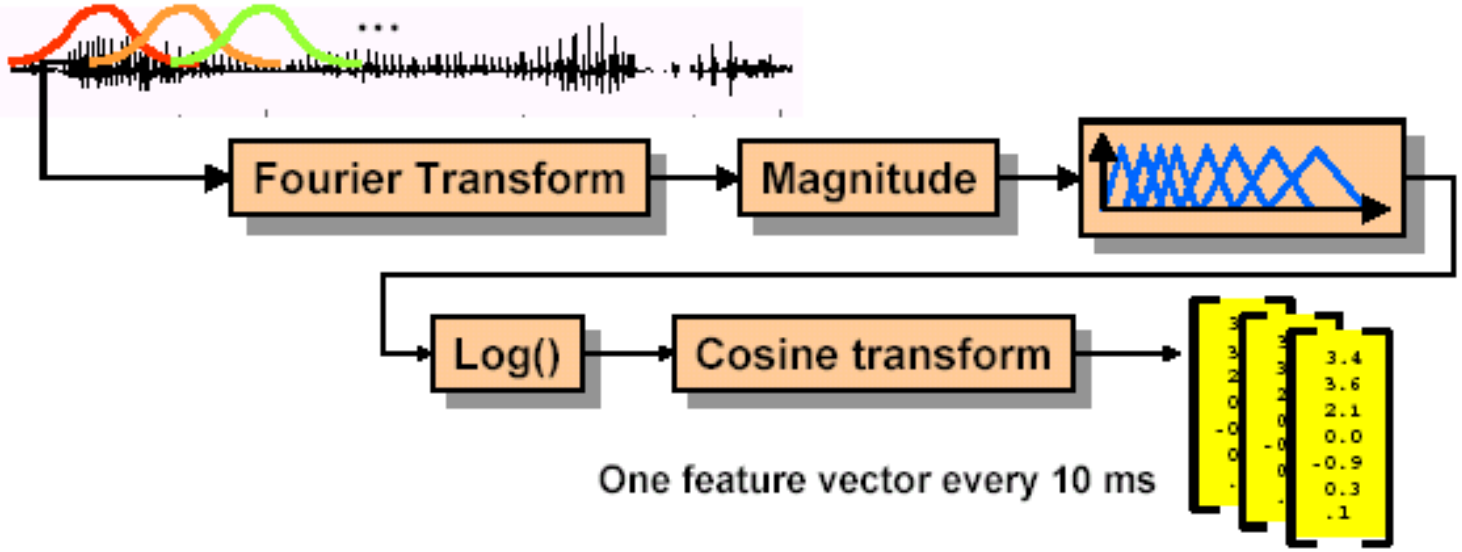


- Produces time-frequency evolution of the spectrum



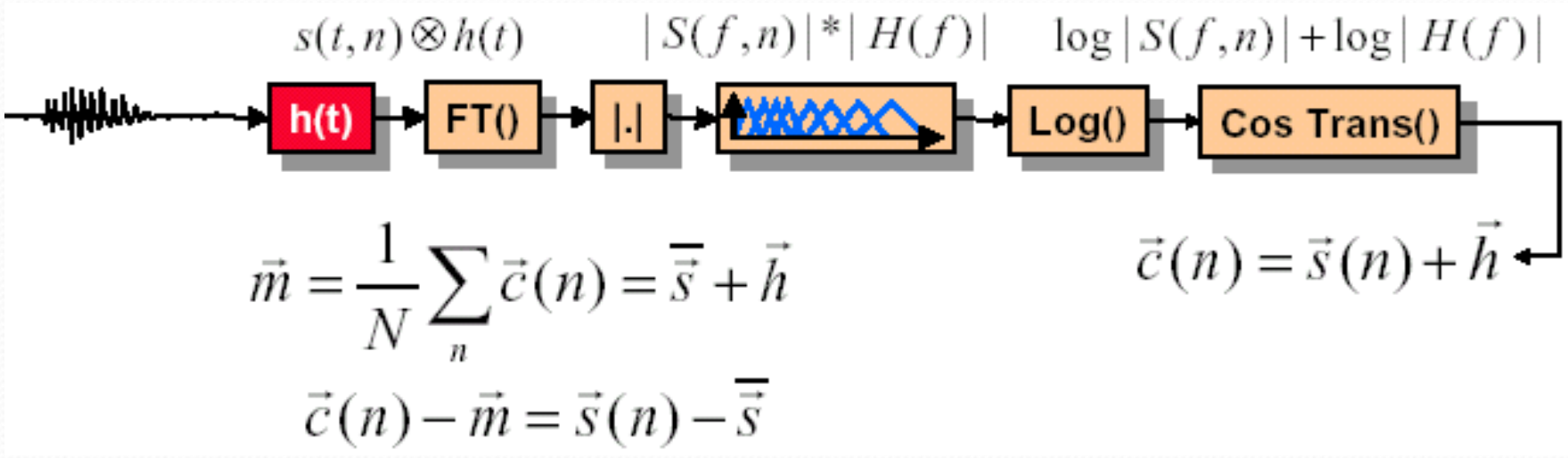
Características para RAL - Extração do Cepstro

- **Uma das principais características utilizada no RAL é o cepstro, ou o Mel-cepstro quando se utiliza um banco de filtros perceptual com escala mel ou bark;**
- **A função Log() transforma a convolução devida ao canal em uma adição → mais fácil a remoção dos efeitos do canal;**
- **A transformada cosseno ajuda na descorrelação dos elementos do vetor de características.**



Características para RAL - Deconvolução Cega

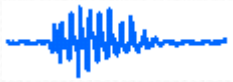
- **Para reduzir os efeitos do canal, utiliza-se a subtração da média cepstral (CMS) ou a filtragem RASTA aplicada aos vetores de características cepstro;**
- **Alguma informação do locutor é perdida porém a literatura mostra que o CMS melhora o desempenho dos sistemas;**
- **A filtragem RASTA é como um CMS variante no tempo.**



Fases de um Sistema de RAL

Fase de Treinamento

João

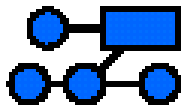


Pedro

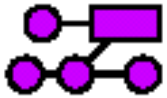
Extração de Características

Treinamento do Modelo

Modelo para cada locutor



João



Pedro

Fase de Verificação

Extração de Características

Decisão da Verificação

Aceito

Suposto Locutor: Pedro

Modelos de Locutores

- **Os modelos representam as informações de um locutor contidas nos vetores de características;**
- **Atributos desejados no modelo**
  - **Significado teórico**
  - **Generalização para novos dados**
  - **Viabilidade da Representação (tamanho e custo computacional compatíveis)**
- **Muitos técnicas tem sido aplicadas no RAL**

Modelos de Locutores - Tecnologias Utilizadas

- **Template Matching**
  - **“Dynamic time warping” para alinhar as seqüências de características dos dados de treinamento e teste**
  - **Utilizado principalmente para aplicações dependentes do texto.**
- **Vizinho mais Próximo**
  - **Retém todas as características durante o treinamento**
  - **Para cada vetor de característica das locuções de teste, acha a distância para os vetores mais próximos doas dados de treinamento;**
  - **Memória utilizada para armazenar o modelo pode ser elevada**

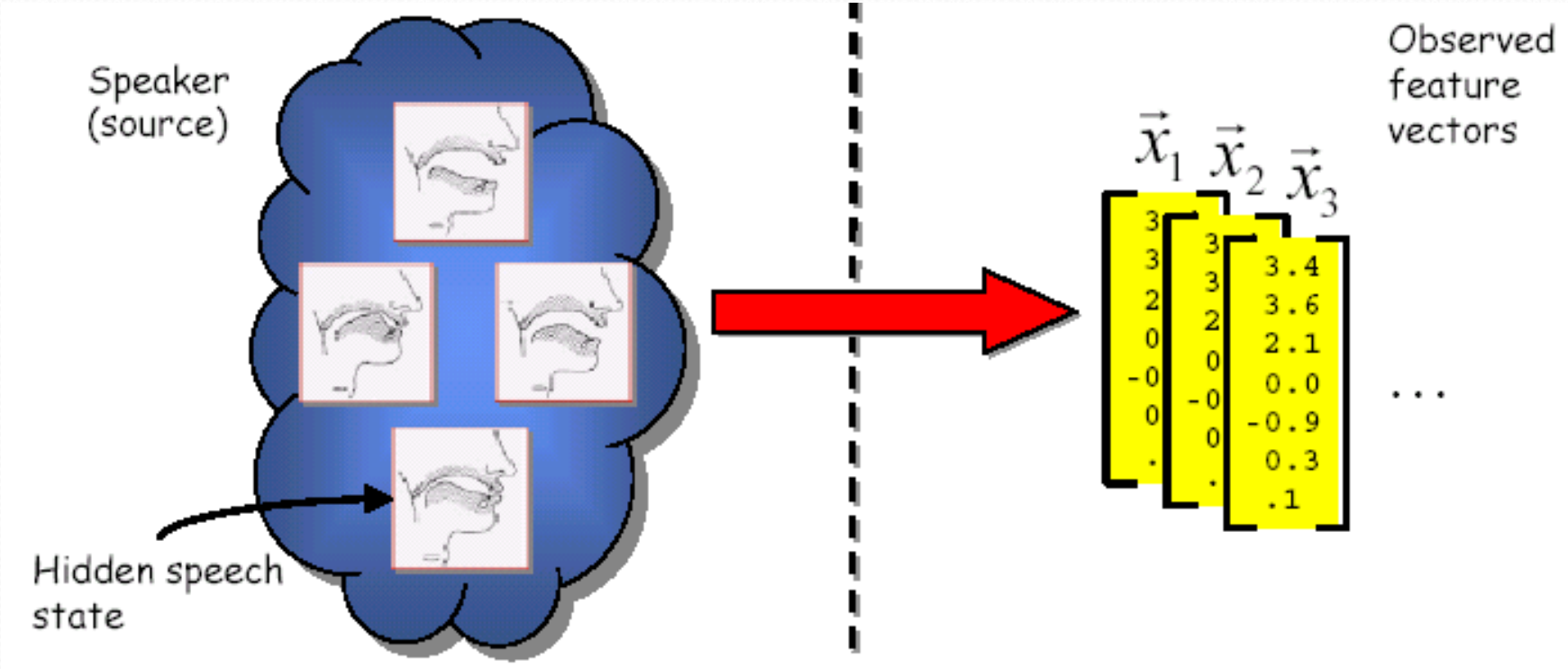
Modelos de Locutores - Tecnologias Utilizadas

- **Neural Networks**
  - **Muitas formas: multi-layer perceptrons, radial basis functions, neural-tree networks.**
  - **Explicitamente treinada para discriminar um locutor de outros**
  - **O treinamento pode ser computacionalmente custoso e algumas vezes pode não generalizar.**

- **Hidden Markov Models**
  - **Representação estatística de como um locutor produz um som;**
  - **Sólida base teórica;**
  - **Principal modelo utilizado em modernos sistemas de RAL**

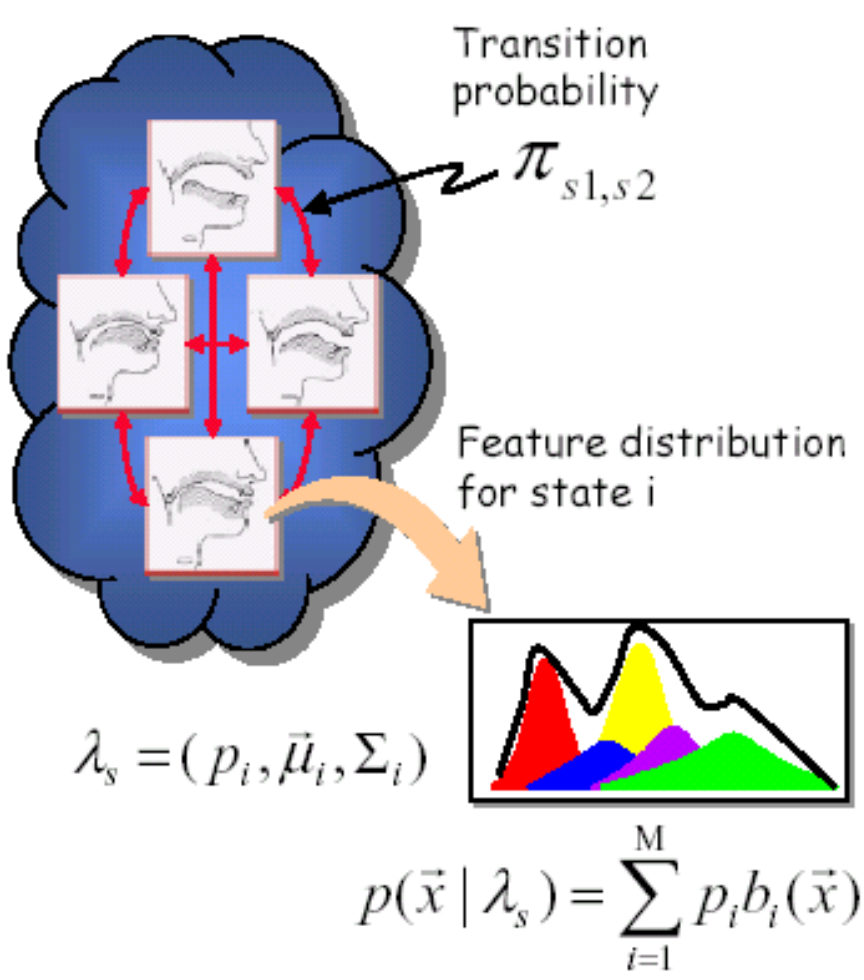
Modelos de Locutores - HMM

- **Trata os locutores como fontes aleatórias escondidas gerando vetores de características**
  - **Fontes tem estados correspondendo a diferentes sons**



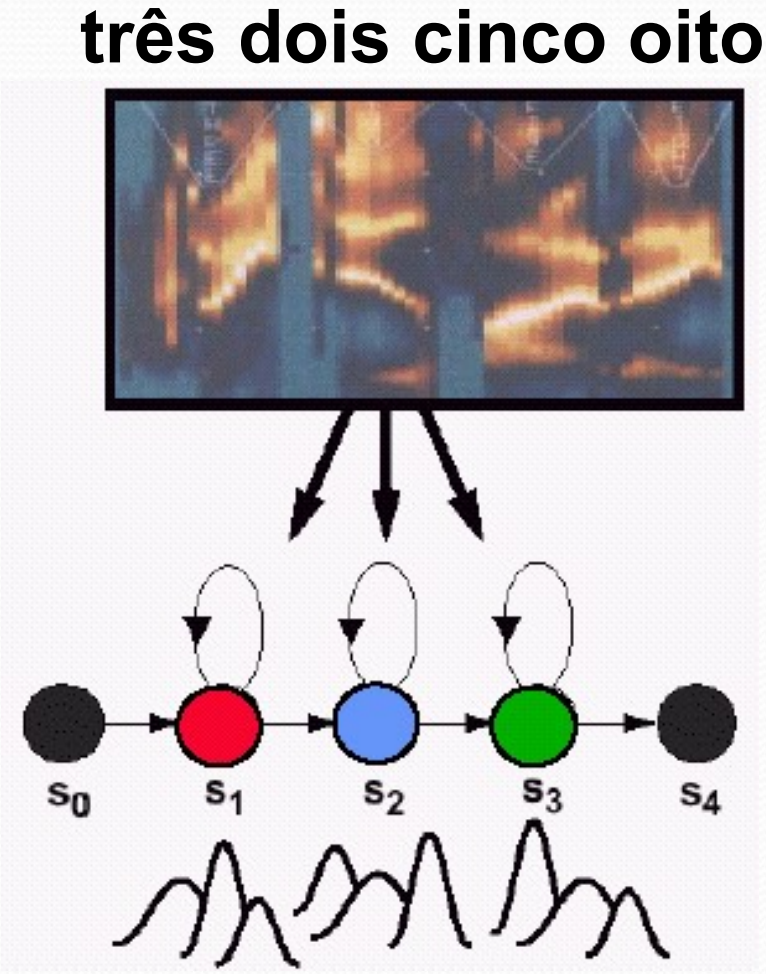
Modelos de Locutores - HMM

- **Vetor de Características gerado em cada estado obedece a um distribuição de misturas gaussianas;**
- **Transição entre estados é baseada na modalidade da voz**
  - **Caso dependente do texto obedece o modelo left-rigth**
  - **Caso independente do texto segue o modelo ergótico**
- **Parâmetros do Modelo**
  - **probabilidade das transições**
  - **parâmetros das misturas dos estados**
- **Treinamento é feito através do algoritmo EM**



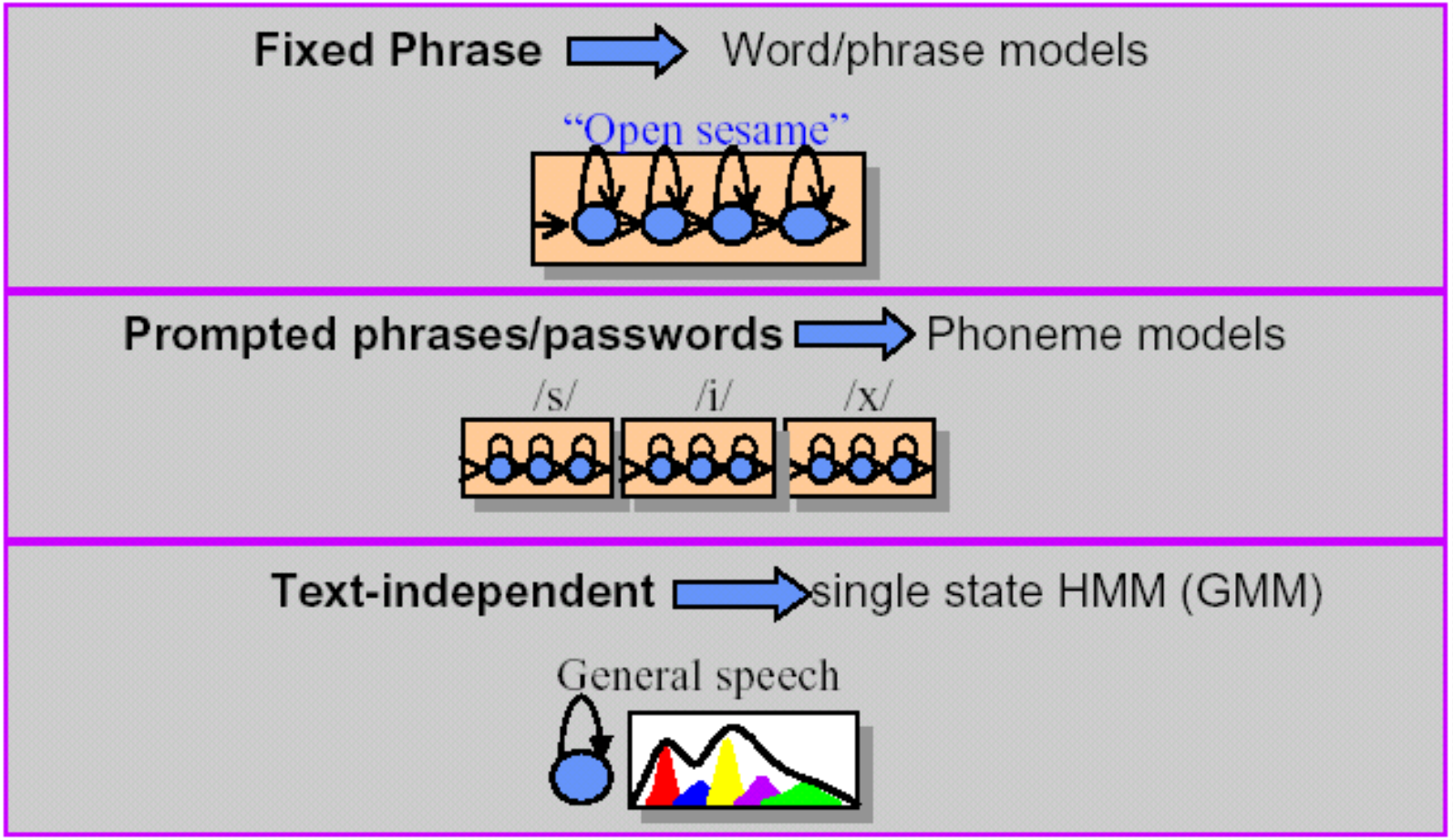
Modelos de Locutores - HMM

- **HMMs codificam a evolução temporal das características**
- **HMMs representam as variações estatísticas no estado da voz (ex: fonemas) e mudanças temporais da voz entre os estados.**
- **Ele provê um modelo estatístico de como um locutor produz os sons da fala**
- **O projetista deve configurar**
  - **A Topologia (# estados e tipos de transições)**
  - **Número de Gaussianas/Estado**



Modelos de Locutores - HMM

O tipo de HMM depende da aplicação

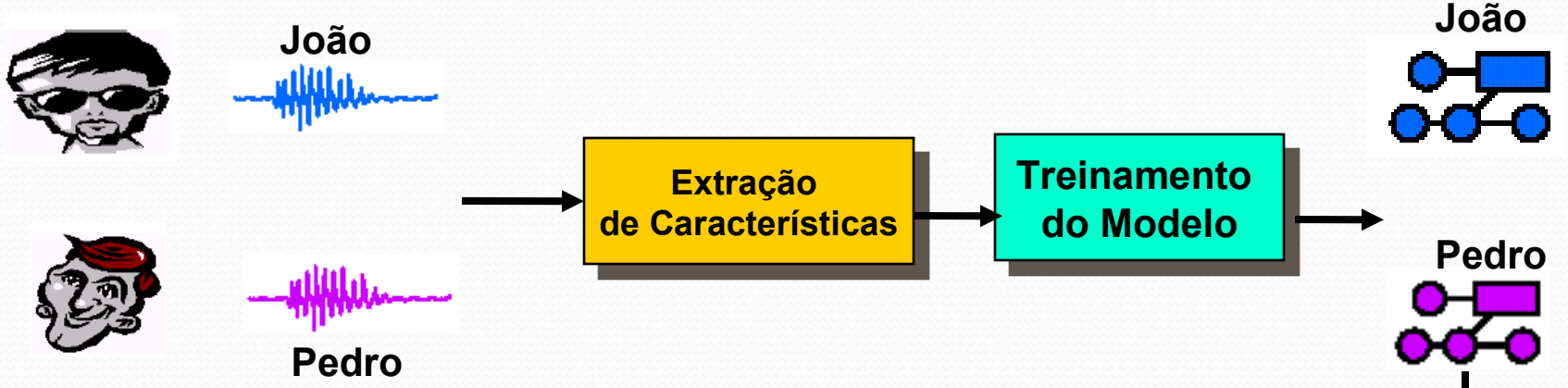


Modelos de Locutores - HMM

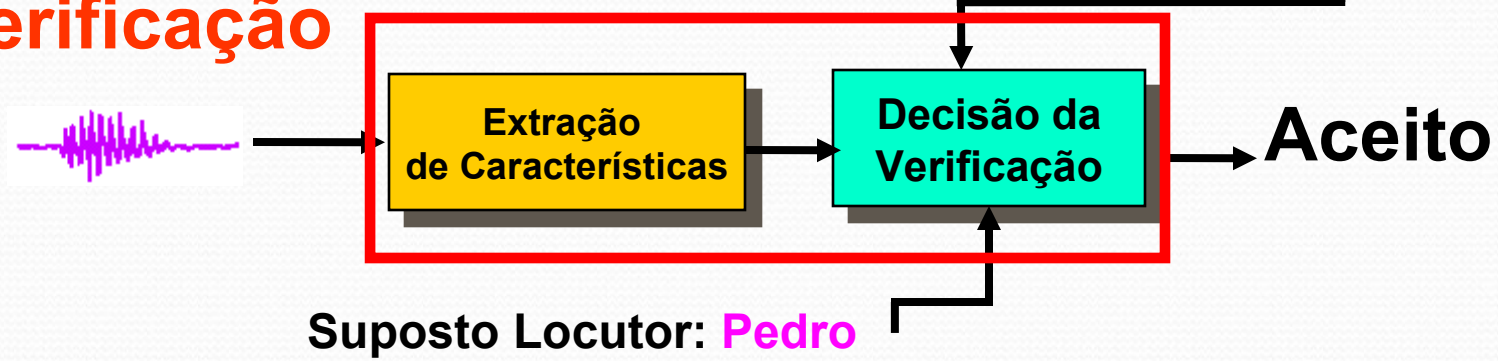
- **O fator dominante no modelo para o desempenho em sistemas de reconhecimento de locutor é o número de gaussianas utilizadas.**
- **A seleção do Nr de Gaussianas depende dos seguintes fatores**
  - **Topologia do HMM**
  - **Quantidade dos dados de treinamento**
  - **Tamanho do modelo**
- **Não há um técnica teórica para o nr de Gaussianas**
  - **Normalmente a escolha é empírica**

Fases de um Sistema de RAL

Fase de Treinamento



Fase de Verificação



Fase de Verificação

- **A verificação é fundamentalmente um teste de hipótese com duas classes**
  - **Ho:** A voz S é de um impostor
  - **H1:** A voz S é de locutor correto
- **Selecionamos a hipótese mais verossímil (Teste de Bayes para erro mínimo)**

$$\Pr(H1 \mid S) \underset{<}{\overset{>}{}} \Pr(H0 \mid S)$$

$$\frac{p(S \mid H1)\Pr(H1)}{p(S)} \underset{<}{\overset{>}{}} \frac{p(S \mid H0)\Pr(H0)}{p(S)}$$

$$\frac{p(S \mid H1)}{p(S \mid H0)} \underset{<}{\overset{>}{}} \frac{\Pr(H0)}{\Pr(H1)}$$

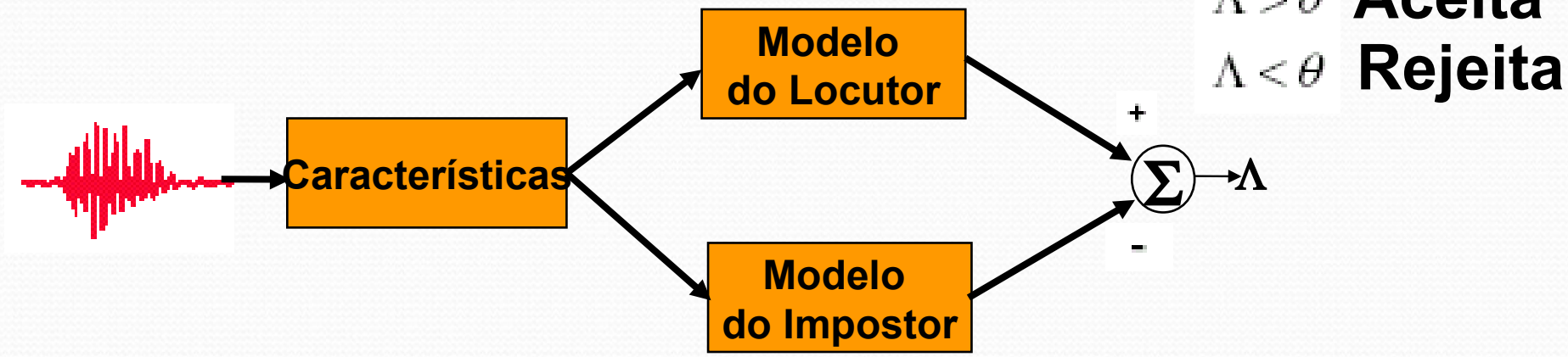
- **Isto é conhecido como Teste de Razão de Verossimilhança (likelihood ratio test)**

$$LR = \frac{p(S \mid H1)}{p(S \mid H0)} \qquad \begin{array}{ll} LR > \theta & \text{Accept } H1 \\ LR < \theta & \text{Accept } H0 \end{array}$$

Fase de Verificação

Normalmente é utilizado o *log-likelihood ratio*

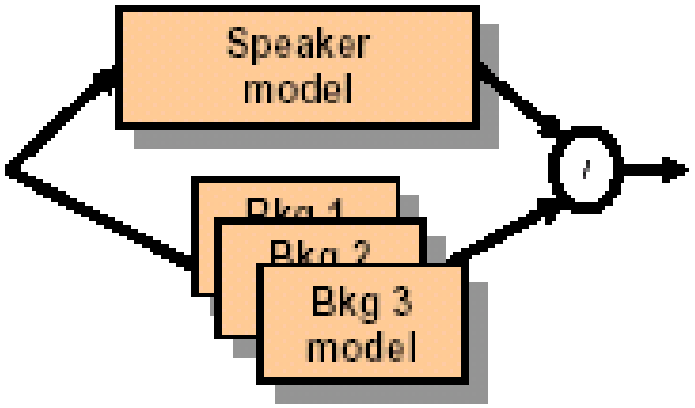
$$LLR = \Lambda = \log p(S \mid H1) - \log p(S \mid H0)$$



- A verossimilhança H1 é calculada utilizando o suposto locutor verdadeiro
- Requer um modelo alternativo ou impostor para a verossimilhança de Ho

Modelo de Background

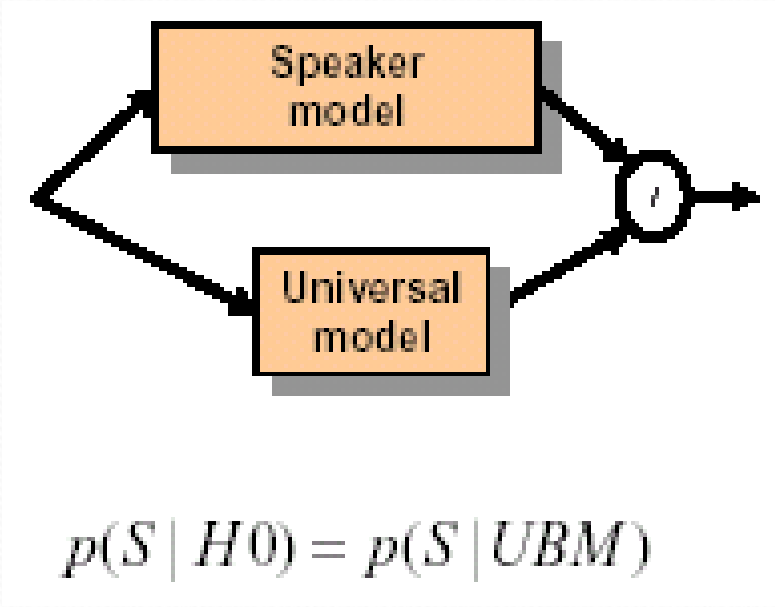
- **Há duas técnicas normalmente utilizadas para criar o modelo alternativo ou do impostor para o teste de razão de varossimilhança**
- **Cohorts/Likelihood Sets/Background Sets**
  - **Utiliza uma coleção de modelos de outros locutores**
    - **A verossimilhança do modelo alternativo é alguma função, tal como a média, das verossimilhanças dos modelos individuais dos impostores**



$$p(S \mid H0) = f(p(S \mid Bkg(b),\ b = 1,\ldots, B)$$

Modelo de Background

- **General/World/Universal Background Model**
  - **Utiliza um único modelo independente do locutor**
  - **Treinado com sinais de um grande número de locutores para representar um padrão geral de voz**



$$p(S \mid H0) = p(S \mid UBM)$$

Modelo de Background

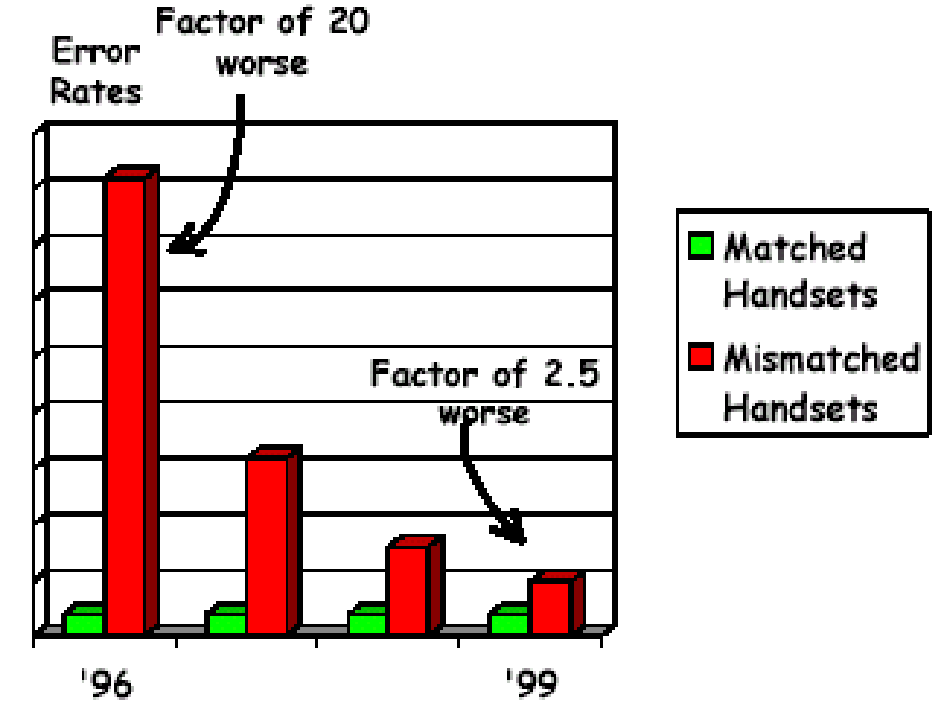
- **O modelo do *background* é primordial para um bom desempenho do sistema**
  - **Atua como uma normalização para ajudar a minimizar a variabilidade devida a informações que não dependentes do locutor na decisão**
- **Utilizando apenas as verossimilhanças devido ao modelo do suposto locutor não se obtém um bom desempenho**
  - **Muito instável para ajustar um limiar de decisão**
  - **Influenciado por muitos fatores não dependentes do locutor**
- **O modelo de background deveria ser treinado utilizando sinais de voz de possíveis impostores**
  - **Mesmo tipo de voz (linguagem, canal)**
  - **Tipos possíveis de microfones**

Problemas no Reconhecimento

- **Variabilidade refere-se a diferenças nos sinais de treinamento e de teste produzidas por variações no canal ou no locutor**
- **Efeitos de Canal**
  - **Os microfones: Carvão, eletreto, etc**
  - **O meio acústico: Escritório, carro, aeroporto.**
  - **O canal de transmissão: linha telefônica fixa, celular, VoIP**
- **Tudo que possa afetar o espectro pode causar problemas**
  - **Efeitos devido aos locutores e as mudanças de canal estão presentes no espectro e podem alterar as características de voz**

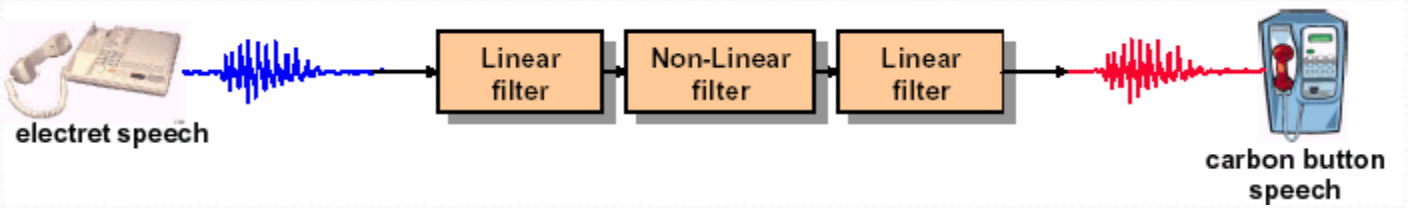
Compensação de Canal

- **Há basicamente três áreas no emprego de algoritmos de compensação**
- **Baseado em Características**
  - **CMS e RASTA**
  - **Mapeamento não linear**
- **Badeados no Modelo**
  - **Modelos de background dependentes do Handset**
- **Baseado em Score**
  - **Hnorm, Tnorm**

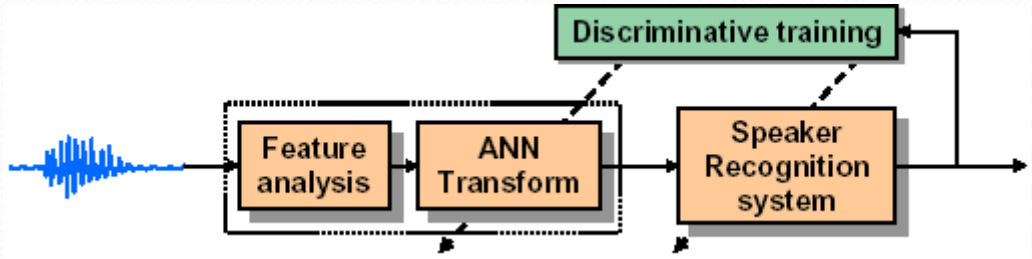


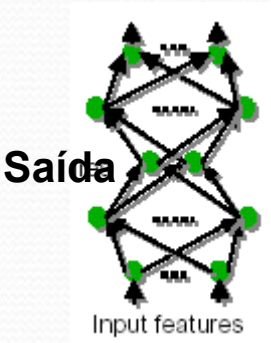
Compensação de Canal - Baseada em Características

- **CMS e RASTA capturam unicamente efeitos lineares nas características**
- **Há algumas técnicas que mapeiam efeitos não lineares**
  - **Mapeamento Não-linear (Quatieri, TrSAP 2000)**
    - **Uso de séries de Volterra para mapear os sinais de voz entre diferentes tipos de Handset**



  - **Transformação de características (Heck, SpeechCom 2000)**
    - **Uso de Redes Neurais para achar características que discriminem os locutores**

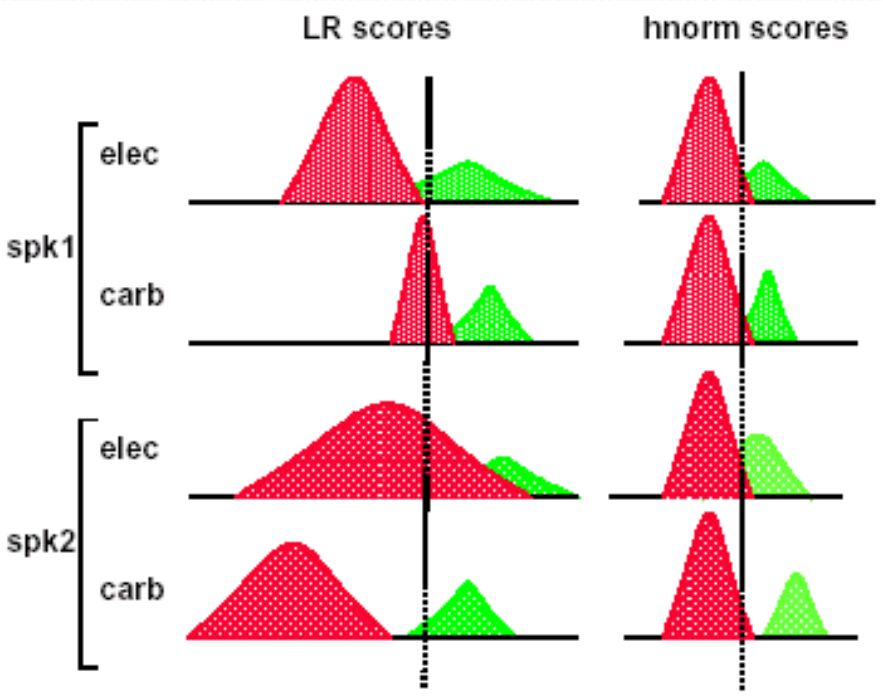




Compensação de Canal - Baseado em Score

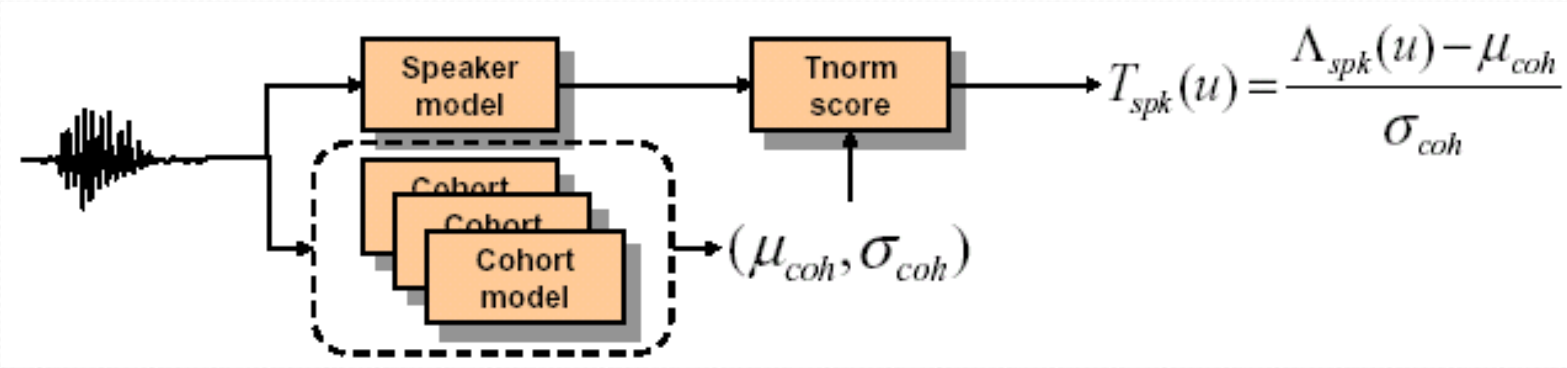
- **LR tem polarizações distintas para elocuções de diferentes tipos de microfones**
- **Hnorm procura remover esta polarização dos LR (Reynolds, NIST eval96)**
  - Estimar a média e desvio padrão do impostor, mesmo sexo e diferentes microfones
  - Durante a verificação normalizar o LR baseado no tipo de microfone utilizado

$$H_{spk}(u^{carb}) = \frac{\Lambda_{spk}(u^{carb}) - \mu_{spk}^{carb}}{\sigma_{spk}^{carb}}$$



Compensação de Canal - Baseado em Score

- **Tnorm/HTnorm - Estima a polarização e escala dos parâmetros utilizando "cohort" (Auckenthaler, DSP Journal 2000)**
  - **Normalização feita durante o teste**
  - **Normalizes o LR do locutor alvo em relação ao modelo do impostor**



- **Cohorts usados do mesmo sexo e canal do locutor alvo**
- **Pode ser usado em conjunto com o Hnorm**

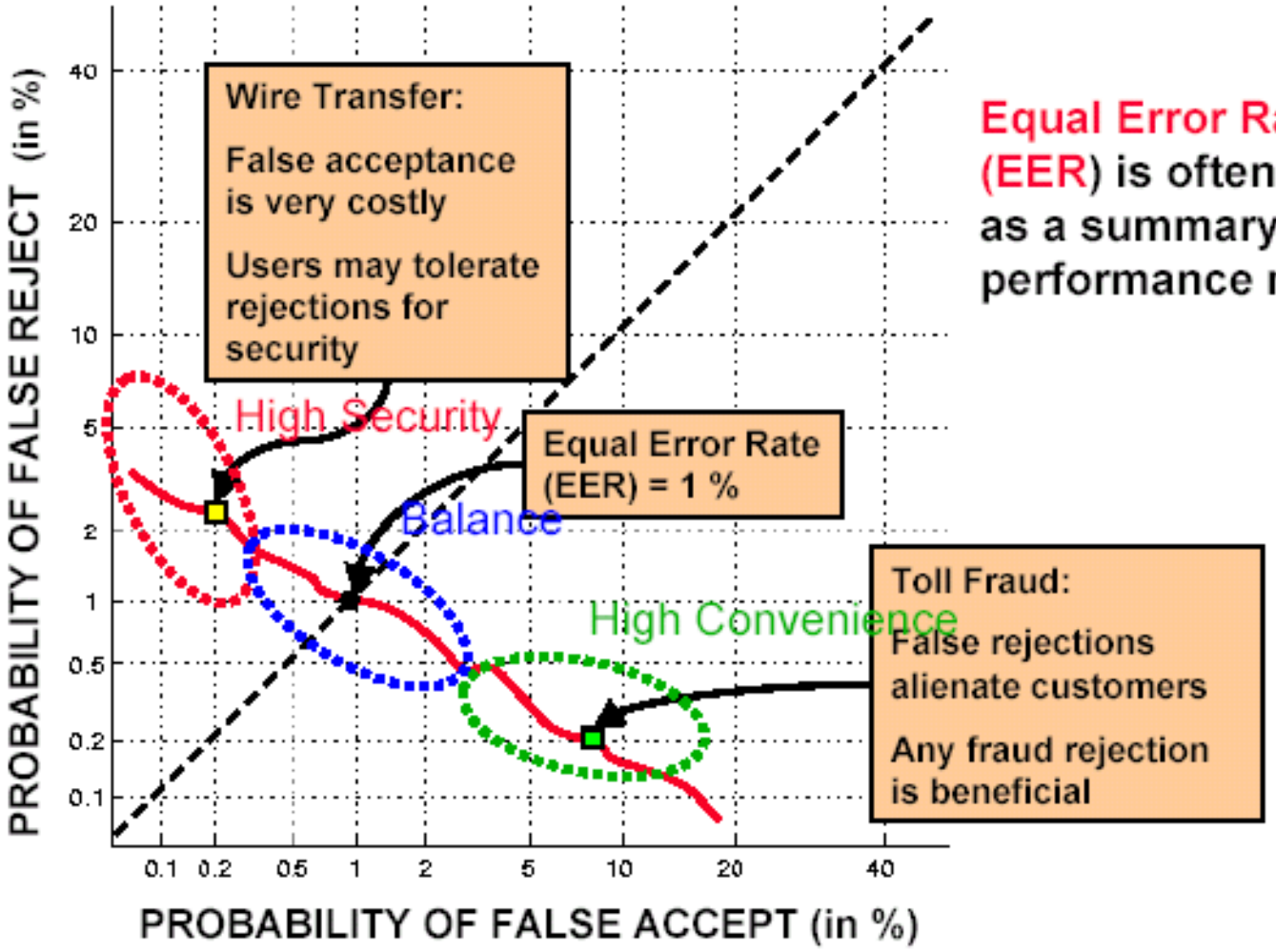
Métrica de Avaliação

- **Em verificação de Locutor podem ocorrer dois tipos de erros**
  - **Falsa Rejeição: rejeitar incorretamente um locutor. Também conhecido como erro Tipo I**
  - **Falsa aceitação: aceita incorretamente um impostor. Também conhecido como erro Tipo II.**
- **O desempenho de sistemas de verificação é medido levando-se em conta estes dois tipos de erros**
  - **O ajuste do sistema é controlado pelo limar de decisão adotado.**
- **Em uma avaliação, calculam-se as probabilidades de falsa aceitação e falsa rejeição para diferentes limiares utilizando-se N_{true} sinais verdadeiros e N_{false} sinais falsos (voz de um impostor).**

Fatores para Avaliação

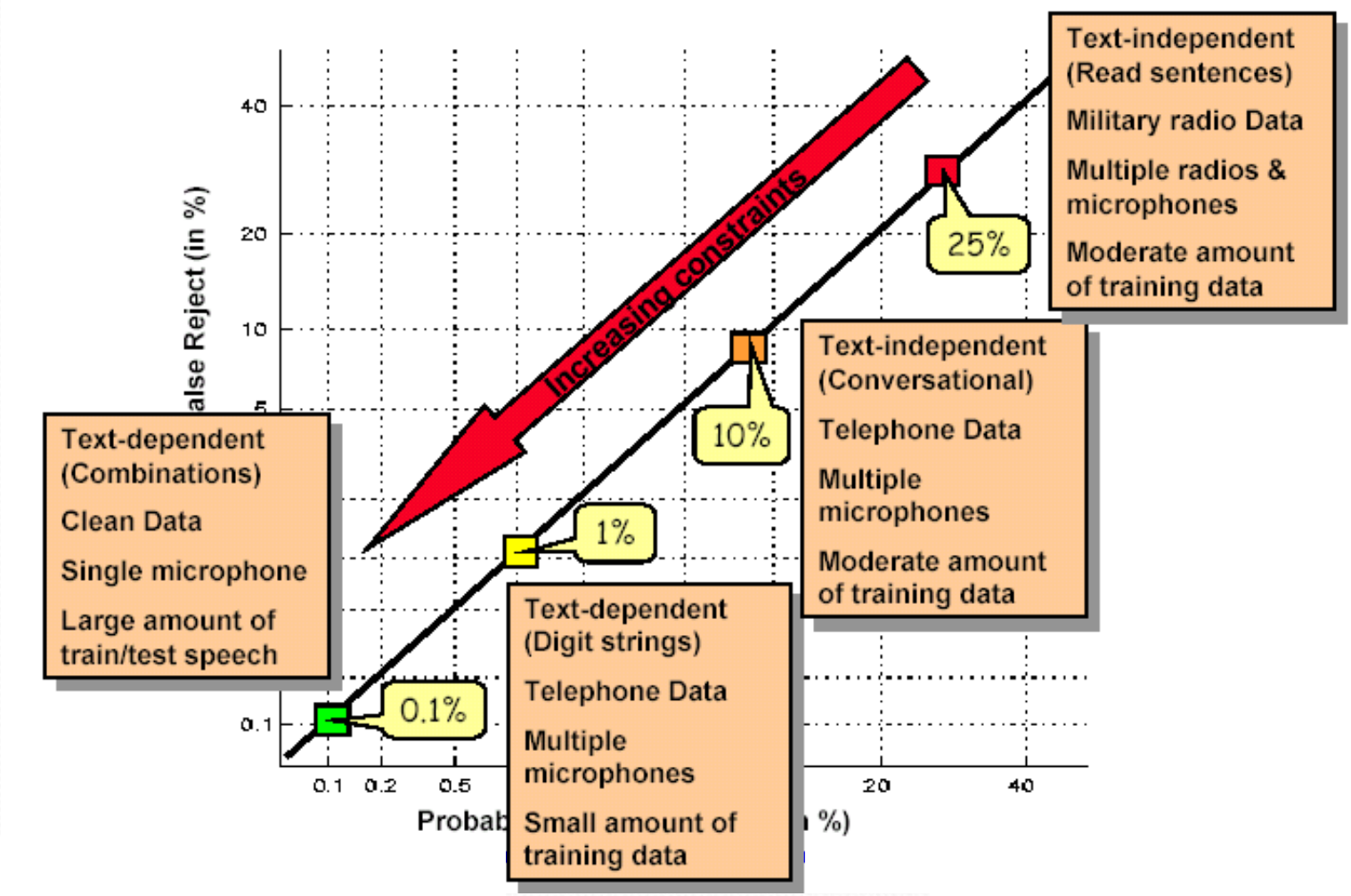
Modalidade de Voz	**Texto fixo /frases selecionadas/ texto livre**
Duração da Voz	**Duração e número de seções de treinamento e de verificação**
Qualidade da Voz	**Características do canal e do microfone** **Variação entre as locuções de treinamento e de teste** **Nível de ruído ambiente**
População de locutores	**Tamanho e composição** **Experiência**

Curva DET

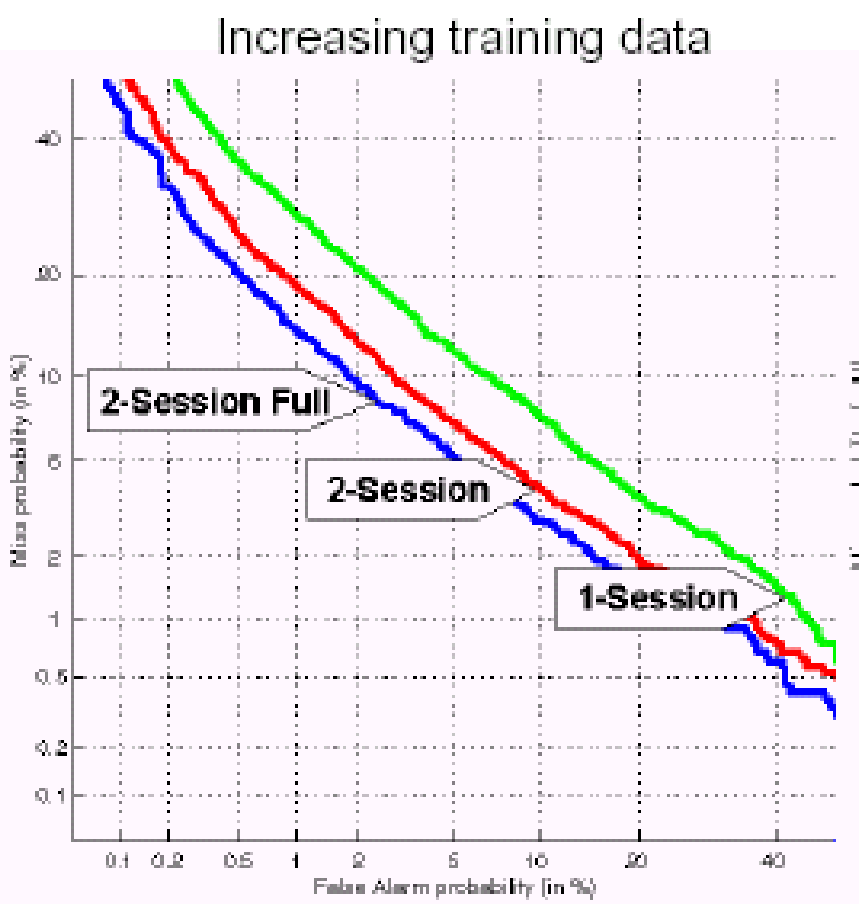


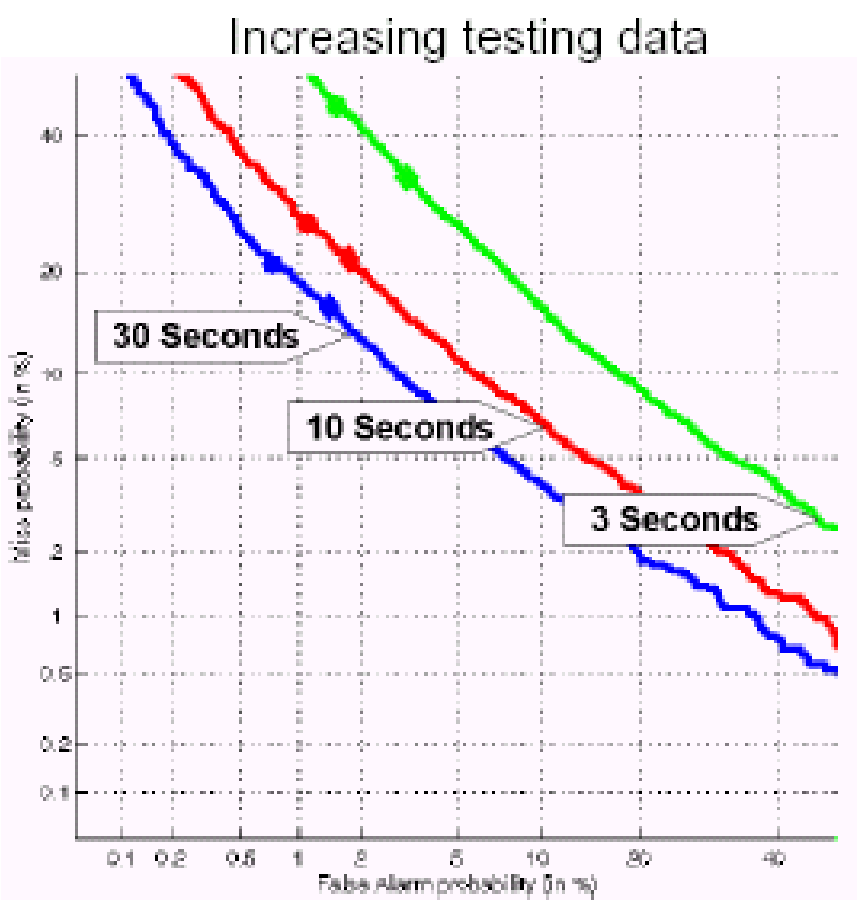
Equal Error Rate (EER) is often quoted as a summary performance measure

Desempenho

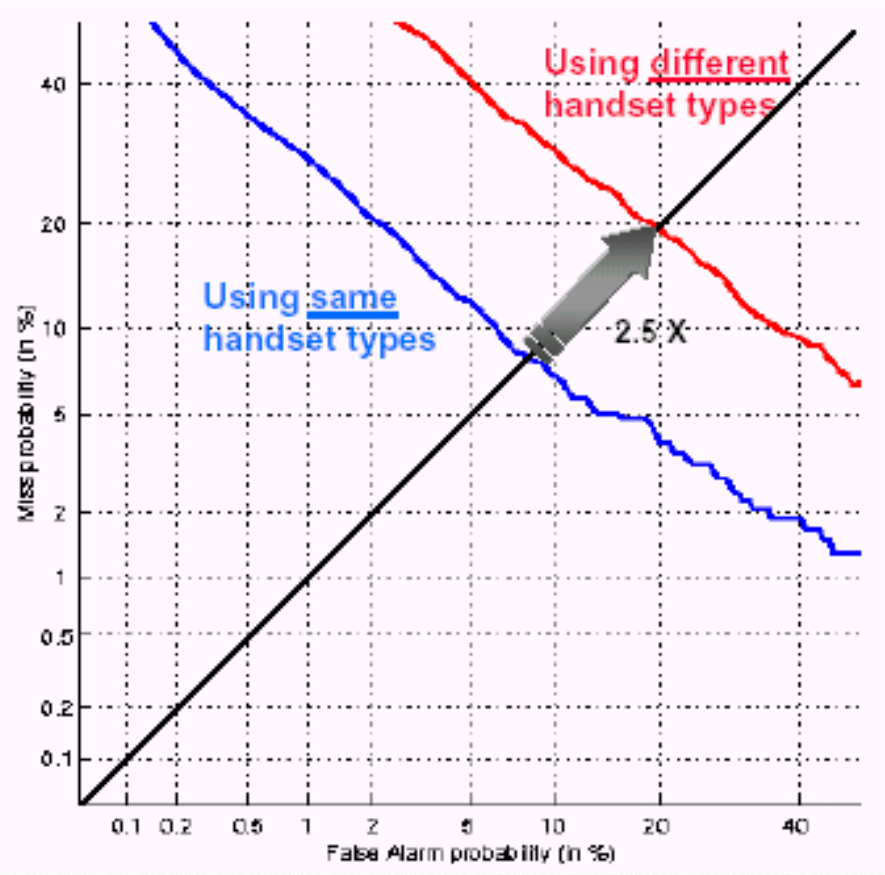

Increasing constraints
Text-independent (Read sentences)
Military radio Data
Multiple radios & microphones
Moderate amount of training data
25%
Text-independent (Conversational)
Telephone Data
Multiple microphones
Moderate amount of training data
10%
Text-dependent (Digit strings)
Telephone Data
Multiple microphones
Small amount of training data
1%
Text-dependent (Combinations)
Clean Data
Single microphone
Large amount of train/test speech
0.1%
40
20
10
0.1
0.2
0.5
20
40

Efeitos da Duração dos Sinais de Treinamento e de Teste





Efeitos Causados por Descasamento de Microfones

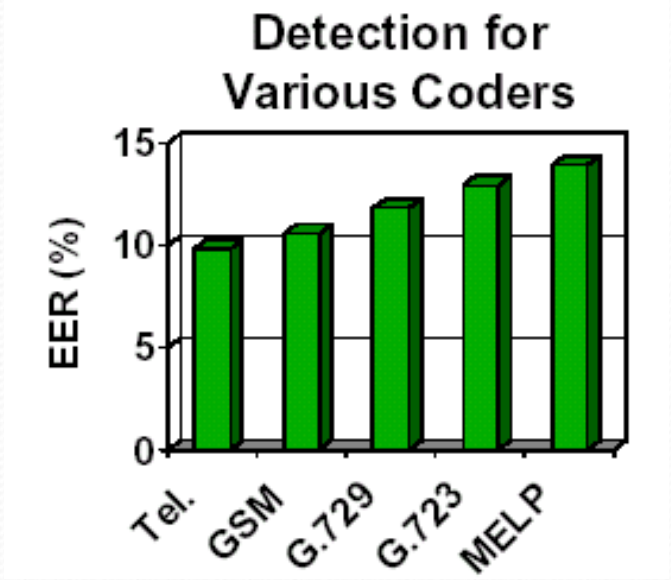


- **No plano de avaliação do NIST, o desempenho foi medido utilizando-se o mesmo microfone e diferentes tipos de microfone em telefones fixos (carvão x eletreto)**

- **Com descasamento de microfones o EER aumenta por um fator**

Efeitos com Codificadores de Voz

- **Reconhecimento da voz reconstruída**
- **Erro aumenta quando a taxa de bits diminui**
  - **GSM tem um desempenho comparável a um sinal não codificado**



Coder Rates:
T1 - 64.0 kb/s
GSM - 12.2 kb/s
G.729 - 8.0 kb/s
G.723 - 5.3 kb/s
MELP - 2.4 kb/s

- **Reconhecimento utilizando os parâmetros do codificador**
- **Pequeno aumento na EER com aumento da eficiência computacional**

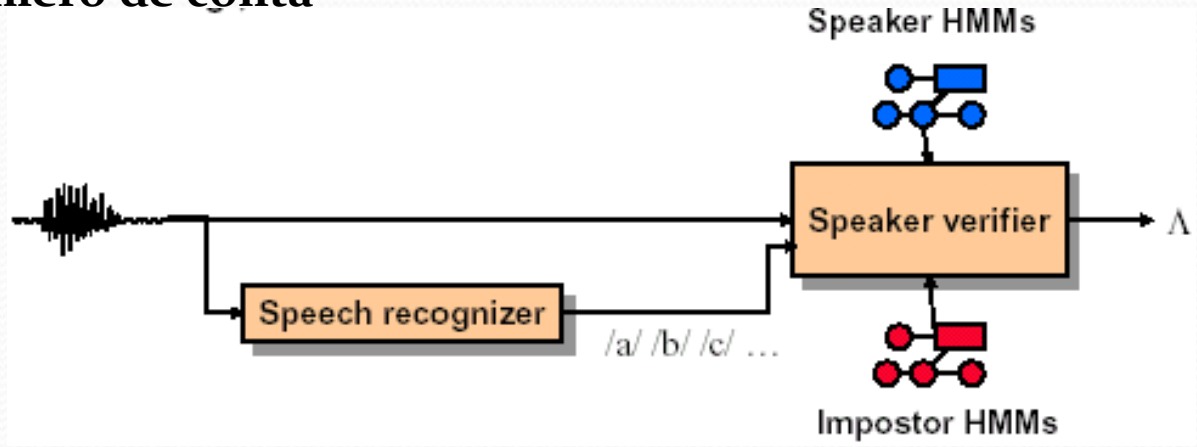
Combinando Reconhecimento de Voz e de Locutor

Explorando Informações de Alto Nível

- **Objetivo: extrair e aplicar todos os níveis de informações da voz para o reconhecimento de locutor**
- **Níveis de Informação**
  - **Acústico: Utiliza características espectrais do trato vocal**
  - **Prosódia: Utiliza características derivadas da prosódia (pitch, evolução da energia) para caracterizar padrões de prosódia dos locutores**
  - **Fonético: Utiliza seqüência de fones para caracterizar a pronúncia dos locutores e seus padrões de fala**
  - **Idioleto: Utiliza seqüência de palavras para caracterizar padrões de palavras usados pelos locutores**
- **Combinação dos níveis de informação**
  - **Modelos acústicos dependentes do texto**
  - **Fusão de informações dos diferentes escores**

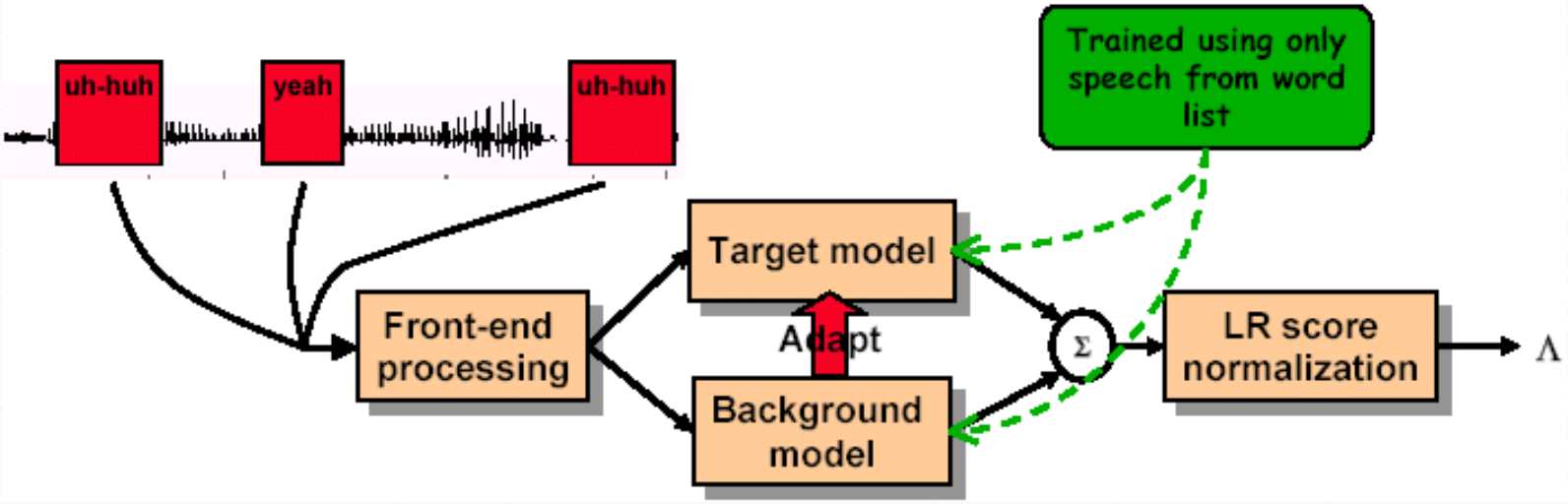
Reconhecedores de Voz e de Locutor

- Reconhecedores de voz utilizados para segmentar o sinal de voz para treinamento e verificação
- Dependência da tarefa, diferentes unidades linguísticas são reconhecidas
  - palavras, fones
- A frase reconhecida pode também ser utilizada para verificar o locutor
  - – Ex: Número de conta



GMM-UBM Utilizando Dependência de Texto

- **Objetivo**: Converter o reconhecimento independente do texto em um sistema dependente do texto.
- **Técnica**: Unidades acústicas baseadas em palavras
  - Seleção de um conjunto de palavras baseadas em algum critério
  - Treinar o UBM e o modelo alvo utilizando somente sinais das palavras selecionadas
  - Calcule os scores utilizando somente as palavras selecionadas



Conclusão

- Reconhecimento de Locutor é um tecnologia viável para aplicações diversas.

Conclusão

- Reconhecimento de Locutor é um tecnologia viável para aplicações diversas.

- Pode-se esperar um melhor desempenho dos sistemas com a utilização de características de alto nível.

Conclusão

- Reconhecimento de Locutor é um tecnologia viável para aplicações diversas.
- Pode-se esperar um melhor desempenho dos sistemas com a utilização de características de alto nível.
- Para o desenvolvimento do RAL, é importante a participação de engenheiros, lingüistas, foneticistas e fonoaudiólogos.

Obrigado.

apolin@ime.eb.br
dirceu@ime.com.br