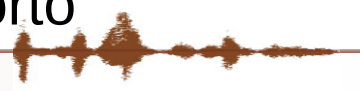
# Áudio Computacional

# Reconhecimento automático da Fala (introdução, reconhecimento de palavras isoladas)

Mestrado Integrado em Engenharia Electrotécnica e Computadores

Faculdade de Engenharia da Universidade do Porto



# Reconhecimento Automático da Fala

# Conteúdo

- Introdução
  - Noção de reconhecimento automático da fala (RAF, *ASR*, *Speech to text converter – STT)*
  - Complexidade do RAF
  - Aplicações do RAF
  - Sistemas de reconhecimento automático da fala (SRAF)
    - Estrutura base
    - Módulo de Análise
    - Módulo de Classificação
  - Ciclo de desenvolvimento dos SRAF

- Reconhecimento de fala baseado no classificador Bayesiano
  - Critério de classificação
  - Modelo acústico (MA)
  - Modelo linguístico (ML)

# Conteúdo

- Reconhecimento de palavras-isoladas (IWR)
  - Introdução
  - Técnica *template matching* e algoritmo DTW (*dynamic time warpping*)
  - Modelos Escondidos de Markov
    - Formalismo estatístico
    - Modelos acústicos baseados em D-, C- e SC-HMMs
  - Redes Neuronais Artificiais (ANN)
    - Perceptrão de camadas múltiplas (MLP, *multi-layer perceptron*)
    - Classificador baseado num MLP

- Reconhecimento de fala-contínua (CSR, *continuous speech recognition*)
  - Introdução
  - Sistema *standard*
  - Modelação acústica
  - Modelação linguística
  - Algoritmos de descodificação
  - *Estado da arte*

# Introdução

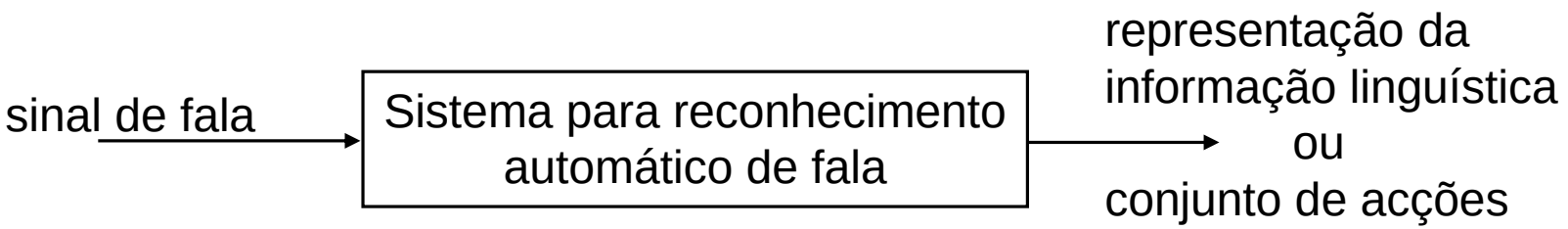


# Reconhecimento Automático da Fala (RAF)

- O Reconhecimento Automático da fala (RAF) baseia-se na extracção, de maneira automática, do conteúdo linguístico associado ao sinal de fala

  (outros tipos de reconhecimento baseados no sinal de fala: reconhecimento do falante; reconhecimento do idioma, ou dialecto; etc.)

- Os sistemas para RAF efectuam o reconhecimento automático de padrões associados ao sinal de fala:
  - classificando-os em determinadas representações simbólicas
  - ou conduzindo à realização de determinado conjunto de acções



# Restrições ao estudo do RAF em ASRF

- (ao nível do sinal de fala)

  considera-se apenas o conjunto de sequências de padrões associados à onda acústica (exclui-se, por exemplo, o estudo do reconhecimento audio-visual, baseado também em padrões de natureza visual)

- (ao nível da abordagem ao problema geral do reconhecimento de padrões)

  são apenas tratados os sistemas baseados em modelos estatísticos de representação do conhecimento e de suporte à decisão (são excluídas as abordagens sintácticas, técnicas de I.A., etc.)

- (ao nível da saída dos sistemas)

  são apenas tratados os sistemas que geram uma determinada representação simbólica da informação linguística, ao nível lexical ou sintáctico, veiculada pelo sinal; por exemplo os “editores de texto por voz” que transcrevem a ortografia associada ao sinal de fala (são excluídos os sistemas que utilizam, ou têm por objectivo extrair, informação linguística de nível semântico ou pragmático)

# Factores de dificuldade do RAF

- Complexidade da língua
  - vocabulários extensos (>$10^5$ palavras) – ambiguidade (‘noz’ ou ‘nós’?) e confusão (‘noz’ ou ‘nos’?) entre palavras é frequente
  - estruturas linguísticas complexas – níveis: lexical; sintáctico; semântico; pragmático
  - espontaneidade da comunicação oral - afastando-se dos cânones da língua (frequentemente o discurso apresenta inflecções, é incompleto, etc.)
  - diluição da informação linguística juntamente com informação para-linguística (prosódia) e extra-linguística (estados emocionais, idade, sexo, etc.)

- Natureza contínua do sinal da fala
  - indefinição das fronteiras entre as unidades linguísticas
  - acentuados efeitos de co-articulação

# Factores de dificuldade do RAF *(cont)*

- Variabilidade das trajectórias acústicas e da sua evolução temporal
  - depende fortemente do falante (em geral; acentuado pelos acentos locais ou regionais)
  - variando também (embora em menor grau, em geral) para cada falante, em função da celeridade do discurso, do estado físico ou psíquico, etc.

- Durante a transmissão (canal acústico, eléctrico, etc.) o sinal da fala é em geral afectado por diversas perturbações
  - ruído e interferência (por exemplo, outros falantes)
  - distorção

- O problema do RAF é fortemente multi-disciplinar (psico-acústica, acústica, linguística, matemática, electrónica, informática, neurologia, etc.) - é uma tarefa vasta e complexa o desenvolvimento de técnicas capazes de integrar eficazmente esse conhecimento (estão desfeitas as ilusórias expectativas iniciais, ~1950, sobre a facilidade do RAF)

# Restrições nas aplicações de RAF - exemplos

- Imposição de restrições na aplicação (tarefa de reconhecimento), ao nível linguístico
  - vocabulários pequenos e selecção, quando possível, de léxicos menos ambíguos e com maior contraste acústico
  - modelos linguísticos rígidos (em geral são demasiado artificiais)
  - reconhecimento de palavras-isoladas, ou fala contínua com “boa” locução (estilo leitura, e não fala espontânea)

- Restrições ao nível acústico
  - utilizador único e conhecido, ou conjunto de utilizadores conhecidos
  - ambiente acústico favorável (SNR>30dB)

# Soluções nos sistemas de RAF - exemplos

- Soluções adoptadas pelos sistemas (reconhecedores), ao nível linguístico
  - processamento integrado dos diversos níveis linguísticos
  - heurísticas eficazes específicas para a aplicação

- Soluções ao nível acústico
  - técnicas de extracção de características robustas (a diversos tipos de contaminação acústica)
  - embora de maneira ainda “insipiente”, são já utilizadas técnicas que tentam emular determinadas características do ouvido interno humano (os chamados “modelos auditoriais”)
  - técnicas de modelação sofisticadas, por exemplo considerando a informação acústica de contexto, adaptação prosódica, etc.
  - em sistemas para múltiplos utilizadores, treino com material acústico relativo a múltiplos falantes, capacidade de adaptação ao falante, etc.

# Exemplos de Aplicações do RAF

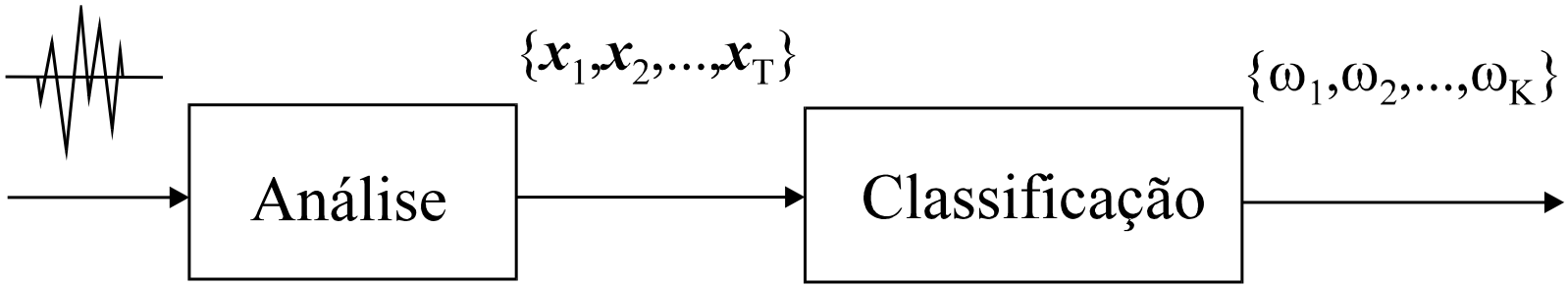
- Telecomunicações (sistemas de acesso a dados e serviços por via telefónica; etc.)
- Ambiente de “escritório” (edição de documentos; agendas; controlo de sistemas informáticos; etc.)
- Sistemas de apoio a actividades profissionais específicas (produção de relatórios médicos; *memos*; tradução automática, etc.)
- Ambiente fabril (comandos accionados por voz; registo de informação; etc.)
- Sistemas de apoio a deficientes
- Controlo de utensílios e aparelhos de utilização doméstica
- Actividades de laser, etc.

# Resultados de Aplicações do RAF

- Resultados típicos (1998) de reconhecimento considerando diferentes restrições na aplicação:
  - modo do discurso
  - dimensão do vocabulário e perplexidade
  - (in)dependência do falante

| Fala | Vocabulário | Gramática (perplexidade) | Dependência do falante | Taxa de erro (%) |
|---|---|---|---|---|
| isolada | 10 dígitos | não (10) | sim | 0.0 |
| | | | não | 0.1 |
| | 39 $\alpha$-dígitos | não (39) | sim | 4.5 |
| | | | não | 7.0 |
| contínua | 991 palavras | sim (60) | não | 3.0 |
| | 1.800 palavras | sim (25) | | 3.0 |
| | 20.000 palavras | sim (145) | | 12.0 |

# Estrutura base dos Sistemas de RAF




- O **módulo de Análise** converte o sinal de entrada numa representação mais adequada ao processo de classificação

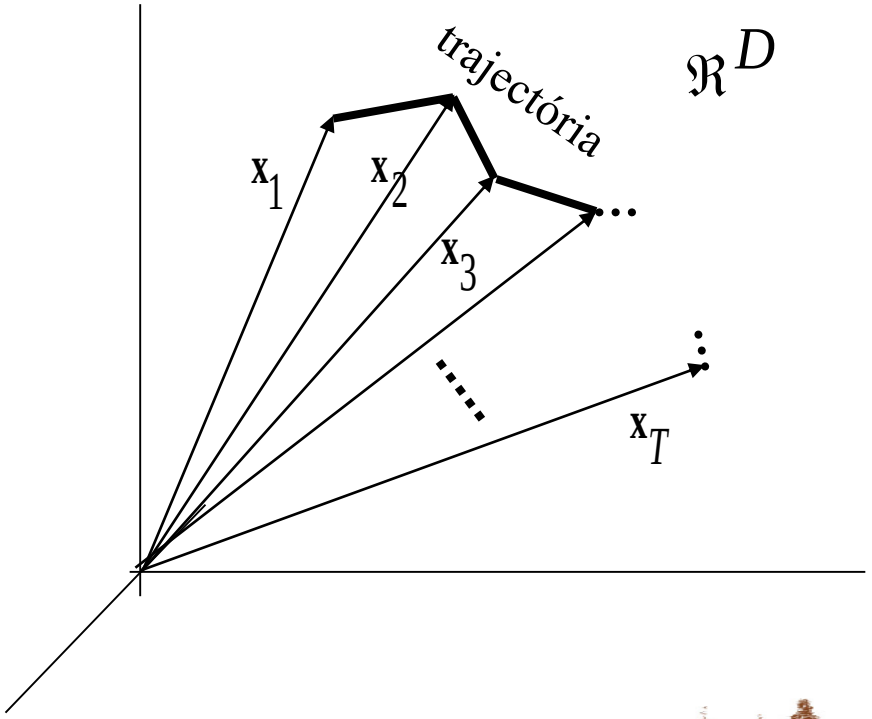
- O **módulo de Classificação** transforma essa representação numa sucessão de símbolos (por exemplo, palavras) pertencentes ao vocabulário e relacionados com as classes padrão

# Análise do sinal – *stream* acústico

- A informação extraída do sinal é representada (geralmente) sob a forma de uma sucessão de vectores de características - ***stream* acústico** - definidos num espaço real de dimensão D

$$s(t),\ t_1 \le t \le t_2 \ \rightarrow\ \mathbf{X}_1^T = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\},\ \ \mathbf{x}_i \in \Re^D$$

Qualquer realização acústica pode então ser representada através de uma trajectória no chamado "espaço acústico"



# Análise ideal do sinal

- Idealmente, o módulo de Análise deveria:
  - **extrair do sinal apenas a informação discriminativa**, ou seja, relevante para o processo de classificação ...

    (favorecendo a separação entre as classes, facilitando assim a tarefa do módulo de Classificação e o desempenho de todo o sistema)

  - **conduzir a uma representação compacta do** sinal

    (permitindo usar técnicas de classificação mais poderosas; problemas da estacionaridade do sinal ...)

  - **colmatar deficiências conhecidas do módulo de Classificação**

    (por exemplo, se o Classificador tem “dificuldades de memória”, o módulo de análise pode atenuar o problema introduzindo termos dinâmicos no vector de características)

# Análise típica do sinal

- Tipicamente, em cada *frame com* duração entre 10ms e 25ms e onde se regista a quase estacionaridade do sinal, a extracção de características é baseada
  - na análise espectral
  - e em modelos de produção da fala (exemplo do cepstrum - MFCC é o actual *standard* - ou dos LPCs)

- Exemplos de alternativas que têm vindo a ser investigadas (pouco usadas)
  - modelos auditoriais (emular propriedades importantes do sistema auditorial humano)
  - utilização de janelas de análise com duração muito longa (~200ms); experiências em psico-acústica apontam para características capazes de representar dependências estatísticas de longo alcance
  - extracção de características orientadas para a modelação de “eventos acústicos” que ocorrem na transição entre os períodos de relativa estacionaridade (partes mais ricas e robustas do sinal ...)

# Classificação

- A representação acústica disponível à saída do módulo de Análise é (nos sistemas que aqui interessa estudar) transformada, pelo módulo de Classificação, numa sucessão de símbolos linguísticos pertencentes a um determinado vocabulário $\Gamma_\omega$ .

$$\mathbf{X}_1^T \rightarrow W_1^K = \{\omega_1, \omega_2, ..., \omega_K\}, \qquad \omega_i \in \Gamma_\omega$$

(em grande parte dos sistemas os símbolos linguísticos correspondem a palavras)

- O cerne do Classificador é um modelo matemático baseado num conjunto de funções discriminantes que definem as fronteiras entre as diferentes classes - a aplicação dessas funções exige algoritmos capazes de suportar a dimensão temporal associada à operação de classificação (contrariamente aos classificadores “estáticos”, o que torna o problema bastante complexo)

# Classificação – abordagem Estatística/Simbólica

- **Modelos Estatísticos**
  - abordagem mais adequada se o problema de reconhecimento evidencia uma natureza estatística essencial, permitindo o processamento da informação facilmente quantificável com base em modelos estatísticas
  - exemplos: modelos acústicos representando a distribuição estatística dos vectores de características; gramáticas estatísticas (N-grams)

- **Símbolos e Regras**
  - abordagem mais adequada se a informação fundamental para o reconhecimento é veiculada através de relações estruturais - dificilmente quantificáveis mas passíveis de descrição através de estruturas de informação adequadas - entre unidades elementares de informação a que se associam símbolos
  - exemplos: regras acústico-fonéticas; alguns léxicos; interpretadores semânticos e analisadores pragmáticos

- Actualmente os Modelos Estatísticos têm-se revelado indispensáveis (em grande medida a sua vantagem está associada à existência de poderosos algoritmos de treino automático)

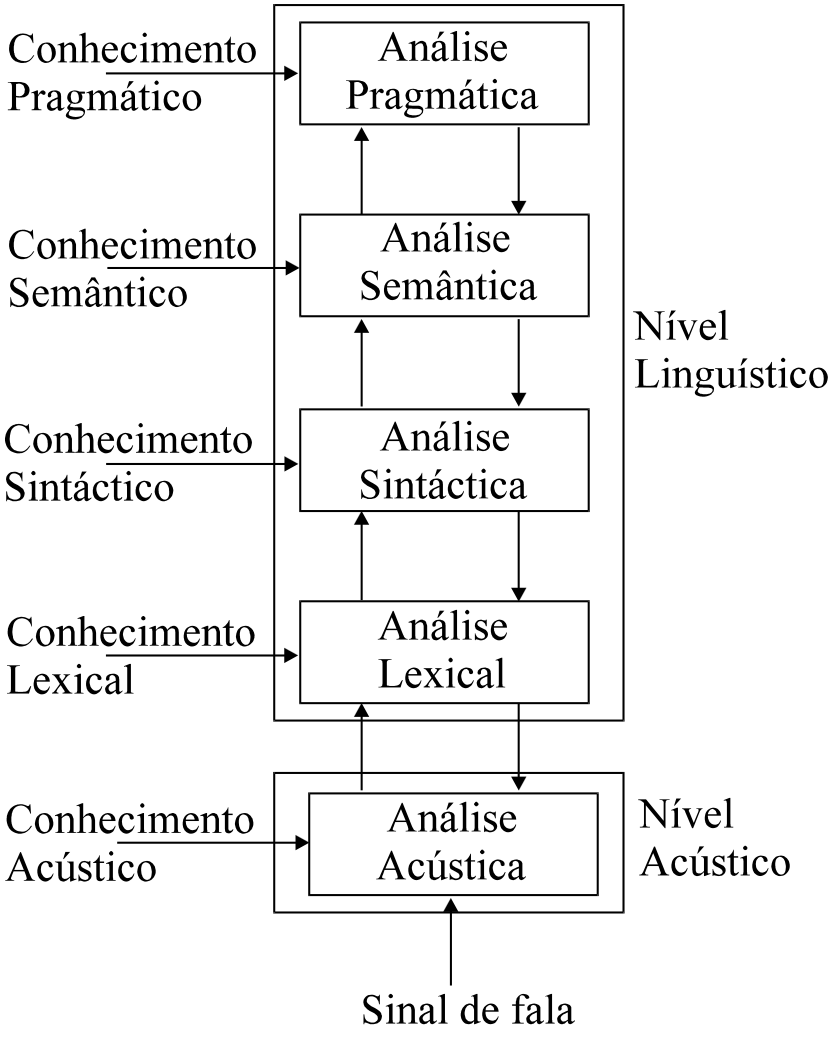
# Classificação - Níveis de processamento

- Necessidade de processar eficientemente a informação que se encontra associada ao sinal da fala, aos diferentes (sub-)níveis, para obter elevado desempenho em aplicações exigentes

- **Nível linguístico**
  - o reconhecedor processa informação associada a diferentes patamares de abstracção do sinal da fala
  - podem co-existir vários sub-níveis: lexical, sintáctico, semântico e pragmático

- **Nível acústico**
  - o reconhecedor relaciona directamente segmentos do sinal da fala com os modelos que representam internamente esses segmentos
  - essas correspondências são geralmente determinadas apenas com base em distâncias entre as características físicas do sinal da fala e os modelos acústicos

# Classificação - representação do conhecimento




Modelo Peircian

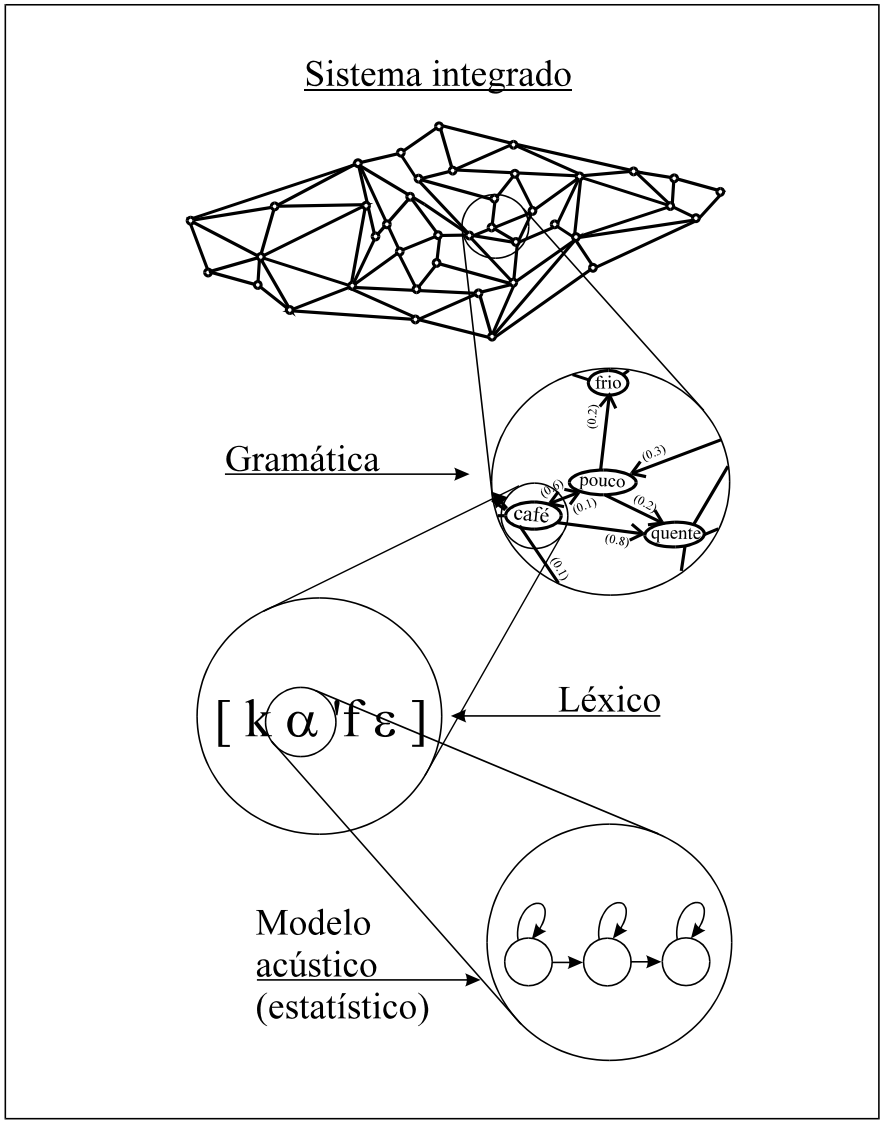
- Exemplo das “estruturas hierárquicas de conhecimento” (por exemplo, o modelo Peircian) com fluxo de informação no sentido
  - *bottom-up*
  - *top-down*
  - híbrido
- Estas estruturas hierárquicas, assim como diversas outras (algumas delas muito complexas) têm conduzido a resultados inferiores às expectativas

  ... têm-se imposto as chamadas “Estruturas Integradas”

# Classificação – Estruturas integradas

Estruturas Integradas para representação do conhecimento

- os diferentes níveis de conhecimento são compilados numa única “rede”
- permitem a utilização de algoritmos de reconhecimento simples e eficazes (*estado da arte*)



# Desenvolvimento dos SRAF – Aplicação

## Especificação da aplicação

- dependente ou independente do falante?
- fala isolada ou contínua?
- dimensão do vocabulário?
- ambiente silencioso ou não controlado?
- operação em tempo real ou não?
- etc.

# Desenvolvimento dos SRAF – Concepção

## Concepção do sistema de reconhecimento

- conhecimento para tomar decisões fundamentais relativamente às técnicas mais indicadas
  - no módulo de Análise: que método de representação? inclusão ou não de derivadas de 1ª e 2ª ordem? utilização de escala perceptiva Mel? etc.
  - no módulo de Classificação: tecnologia dos HMMs, MLPs, ou híbridos? que restrições estruturais (limitações às dependências estatísticas) nos modelos acústico e linguístico? etc.
- testes experimentais preliminares para ajuda à tomada de decisões e afinação dos métodos

# Desenvolvimento dos SRAF - Dados

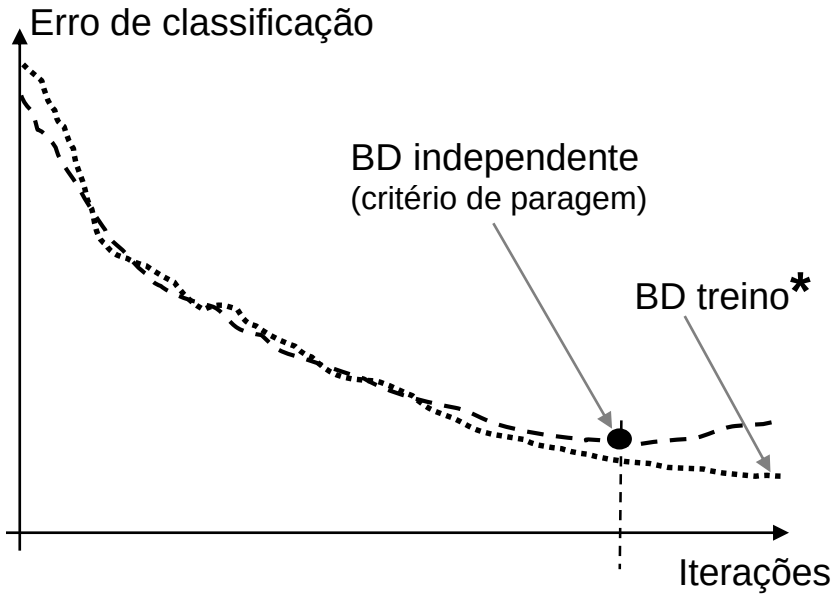
**Criação de base(s) de dados acústica (e de texto)**

- planeamento (volume de dados necessários? exigências de equilíbrio no conteúdo? ...)
- tipicamente divide-se a base de dados em 3 partes (técnicas alternativas ...)
  - a maior (geralmente ) é directamente utilizada na adaptação (treino) dos parâmetros livres
  - outra é utilizada como auxiliar no algoritmo de treino e para testes de desenvolvimento
  - a outra é utilizada para avaliação do sistema
- gravação (equipamento, instalação, ...)
- pós-processamento (segmentação e anotação, ...)

# Desenvolvimento dos SRAF - Treino

## Treino supervisionado dos parâmetros acústicos e linguísticos

- É o processo de determinação do conjunto dos parâmetros livres do sistema, ajustando as fronteiras de decisão, com base na informação contida num conjunto de exemplos de acordo com determinado critério (conhecimento apriorístico)




Não existindo um método analítico para a determinação dos valores dos parâmetros livres, o processo de treino é iterativo

* exemplos utilizados directamente pelo algoritmo de adaptação

# Desenvolvimento dos SRAF - Avaliação

## Avaliação do desempenho do Reconhecedor

- A avaliação do desempenho deve ser feita utilizando a parte da base de dados ainda não utilizada no treino do Reconhecedor (exemplo da técnica “jacknife” de validação cruzada), medindo, portanto, a sua capacidade de generalização

- Os indicadores de desempenho geralmente utilizados são as taxas de
  - palavras erradas

$$WER = \frac{p.Substituidas. + p.Apagadas + p.Inseridas}{p.Total} \times 100$$

  - frases correctas

$$SA = \frac{f.Correctas}{f.Total} \times 100$$

# Reconhecimento de Fala baseado no Classificador Bayesiano



# Critério de Classificação

- **Resultados conhecidos** (dados):

  ***X*** sucessão de vectores de características, que se pretende classificar, extraída do sinal da fala

  $\boldsymbol{W_c}$ , **c=1,…,N** classes, ou hipóteses de reconhecimento; para cada classe (**c** fixo), tem-se a sucessão de símbolos linguísticos correspondentes (sucessão de fonemas, ou palavras, etc.)

  $P(\boldsymbol{W_c}|X)$, **c=1,...,N** probabilidade *a posteriori*, dado ***X***, correspondente a cada uma das **N** classes

- **Questão**:

  Conhecido ***X*** , pertencente a uma das classes definidas $\boldsymbol{W_c}$ (**c** desconhecido), qual é a Regra que conduz à probabilidade mínima de errar a classificação? Ou seja, qual é o **critério óptimo de classificação** (qualquer outro classificador apresenta, em média, maior erro de classificação)?

- **Resposta**:

  A sucessão de vectores deve ser atribuída à classe que apresenta, de entre todas, a **maior probabilidade *a posteriori***

$$P\left(W_c \mid \mathbf{X}\right) > P\left(W_j \mid \mathbf{X}\right), \quad \forall j=1,...,C, \quad j \neq c \quad \Rightarrow \quad \mathbf{X} \rightarrow W_c$$

# Critério de Classificação

- Dada a dificuldade em modelar directamente as probabilidades *a posteriori* usa-se a regra de Bayes

$$P(W_j|\mathbf{X}) = \frac{P(\mathbf{X}|W_j)P(W_j)}{P(\mathbf{X})}, \quad \forall j=1,\ldots,C$$

- … e sendo $P(\mathbf{X})$ constante durante o reconhecimento
- … o **critério óptimo de classificação** fica assim definido:

$$P(W_c)P(\mathbf{X}|W_c) > P(W_j)P(\mathbf{X}|W_j), \quad \forall j=1,\ldots,C, \quad j \neq c \quad \Rightarrow \quad \mathbf{X} \rightarrow W_c$$

- A função discriminante associada a cada classe passa a ser definida pelo produto de dois modelos estatísticos:
  - $P(W)$ , associado ao modelo Linguístico
  - $P(X|W)$ , associado ao modelo Acústico

# Modelo Linguístico

- Determina a **probabilidade *a priori*** de observar a classe, isto é, a probabilidade, não condicionada ao conhecimento do sinal da fala, da "frase" ocorrer

$$P(W_c), \quad c = 1,2,...,C$$

- Exemplos
  - Reconhecimento de palavras isoladas, com modelos-de-palavra equi-prováveis (na prática, o modelo é dispensável)

$$P(W) = \frac{1}{num\_palavras}$$

  - Reconhecimento de fala contínua usando uma gramática do tipo 3-Gram

$$P(W) = P(\omega_0)P(\omega_1|\omega_0)\prod_{t=2}^{N} P(\omega_t|\omega_{t-1},\omega_{t-2})$$

# Modelo Acústico

- Determina a probabilidade* de observar a sucessão de vectores de características condicionada ao conhecimento da classe, isto é, a **verosimilhança** do sinal da fala sendo conhecida a "frase"

$$p(\mathbf{X}|W_c),\ c=1,2,...,C$$

- Modelos acústicos elementares - partilha de parâmetros e de recursos de treino - ao nível fonético, silábico, da palavra, etc.
- Tecnologia utilizada na modelação acústica
  - Modelos Escondidos de Markov (HMM)
  - Redes Neuronais Artificiais (ANN)
  - Técnicas de Programação Dinâmica (algoritmo Viterbi)

* ou, se $\mathbf{X}$ é uma variável aleatória contínua, o valor da função densidade de probabilidade

# Reconhecimento de palavras-isoladas

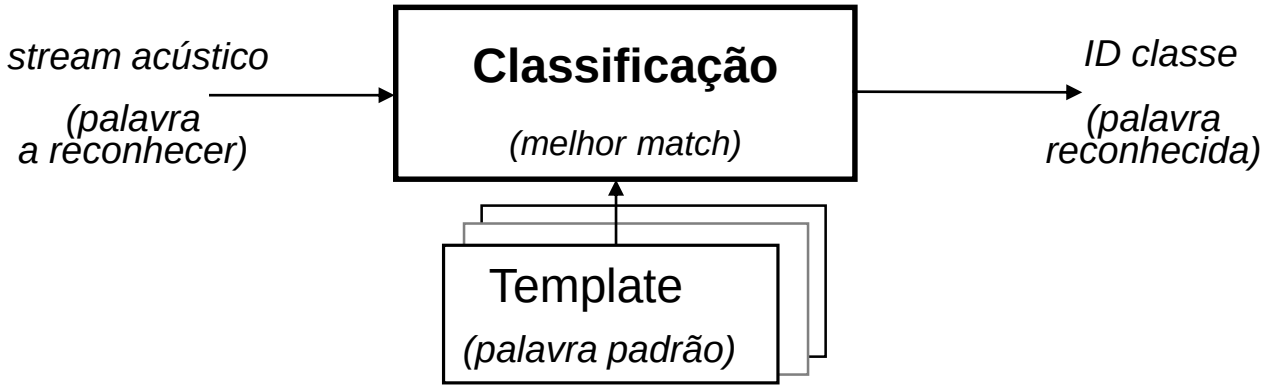


# Introdução

- Pronunciar as palavras isoladamente não é o modo natural de falar
- … mas continuam a existir muitas aplicações interessantes (ex: comandos accionados por voz)

- As dificuldades deste problema de RAF são, comparativamente ao reconhecimento de fala contínua, muito inferiores
- Outras simplificações aqui consideradas facilitando (perspectiva didáctica) a introdução à tecnologia da modelação acústica (DTW, HMM, ANN, ...)
  - **palavras já segmentadas** (*word end-point detection* já efectuado)
  - **modelos-de-palavra** (possível se o vocabulário é pequeno)
  - **modelo linguístico dispensável** (palavras equi-prováveis)

# Técnica de Classificação *Template Matching*



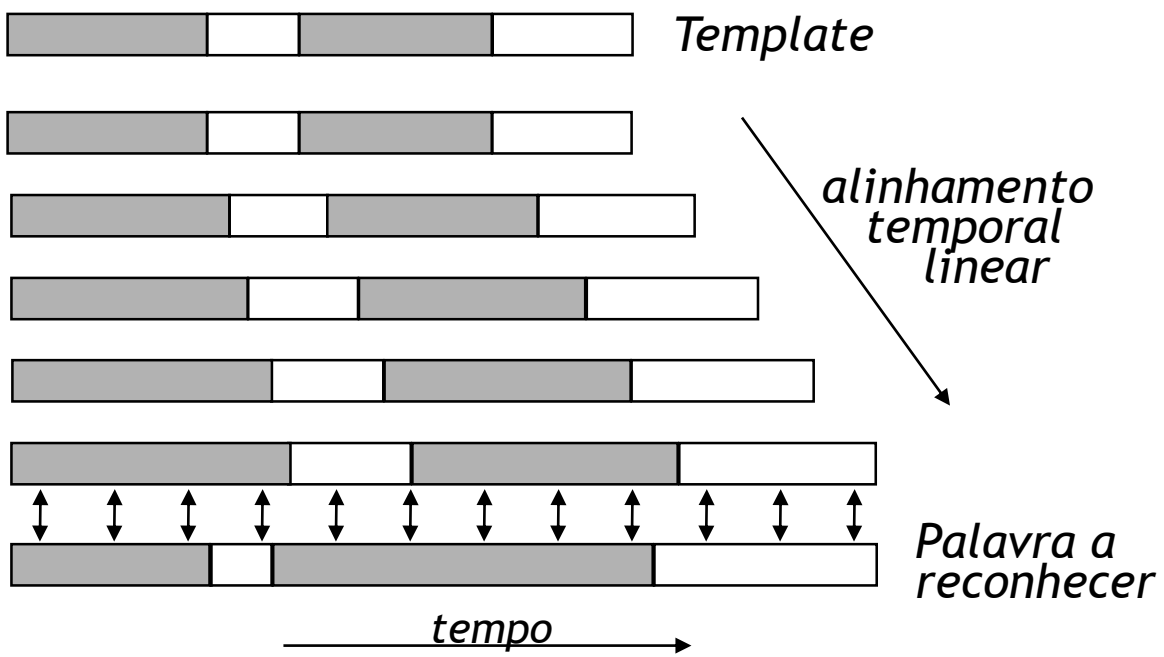

- Classificação é baseada na comparação das **distâncias** entre o *stream* a reconhecer e os templates correspondentes às hipóteses de reconhecimento
- Critérios típicos de decisão sobre a palavra a reconhecer
  - aquela que corresponde ao *template* mais próximo (*nearest neighbour, NN*)
  - aquela cuja média de distância aos K *templates* mais próximos é menor (*K-NN*)
- *Template Matching* pode apresentar resultados satisfatórios quando
  - o nº de *templates* (problema da sua determinação ...) é pequeno, assim como nº de classes (aliviando o processo de comparação)
  - a variabilidade acústica entre classes é grande comparada à que se verifica no interior de cada classe (o problema de classificação é relativamente fácil)

# Problema da Deformação Temporal do Sinal

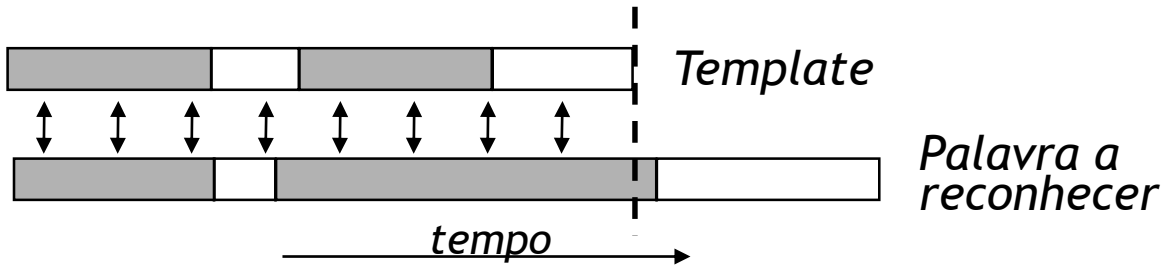
- Mesmo admitindo (hipótese bastante favorável) que a variabilidade das trajectórias acústicas é pequena ...
- … diferentes realizações acústicas de uma mesma palavra apresentam geralmente uma substancial distorção temporal (exemplo dos espectrogramas da palavra "*helicopter*")

# Técnicas de Alinhamento Temporal

- Linear

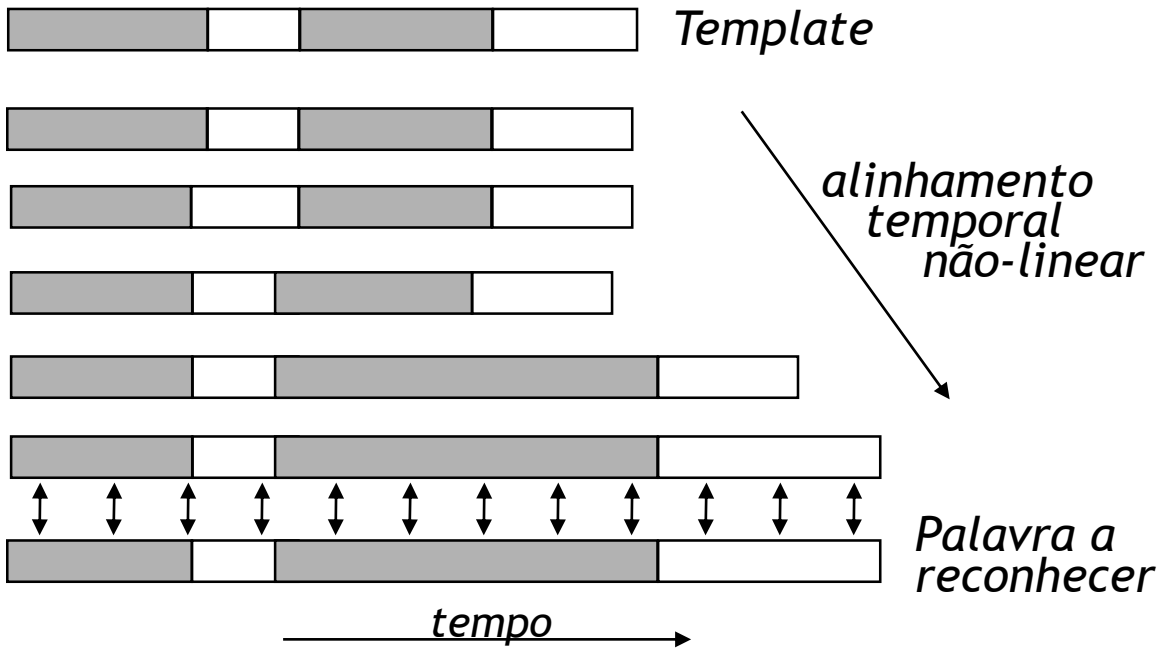



- Truncagem



# Técnicas de Alinhamento Temporal

- **Não Linear** (usando técnicas de Programação Dinâmica)



# Algoritmo DTW

Dados

- palavra a reconhecer – *stream* acústico

$$\mathbf{X}_1^T = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}, \quad \mathbf{x}_i \in \Re^D$$

- *template* da hipótese de reconhecimento (palavra) – *stream* acústico

$$\mathbf{R}_1^N = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N\}, \quad \mathbf{r}_i \in \Re^D$$

- medida de distância entre quaisquer vectores **x** e **r**

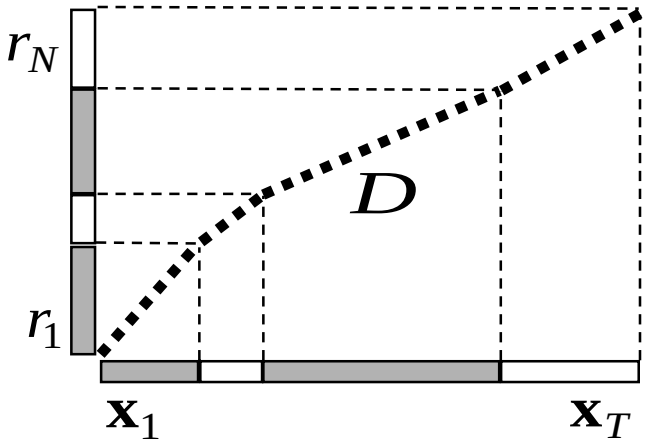
$$d_{i,j} = dist(\mathbf{x}_i, \mathbf{r}_j), \quad i = 1,2,...,T, \quad j = 1,2,...,N$$

Algoritmo DTW

- encontra o melhor alinhamento temporal (caminho) entre as duas sucessões
- calcula a menor "distância" entre as duas sucessões (pelo melhor caminho)

$$D = \min\{Dist(\mathbf{X}, \mathbf{R})\}$$

$$= \min_{caminho} \{d_{1,1} + ... + d_{i,j} + ... + d_{T,N}\}$$



# Algoritmo DTW

*algoritmo DTW*

$D_{1,1} = d_{1,1}$ — *(inicia distância no nó (1,1))*

*para* i=2,…,T — *(ciclo de cálculo recursivo das distâncias em todos os nós possíveis)*

  *para* j=2,…,N

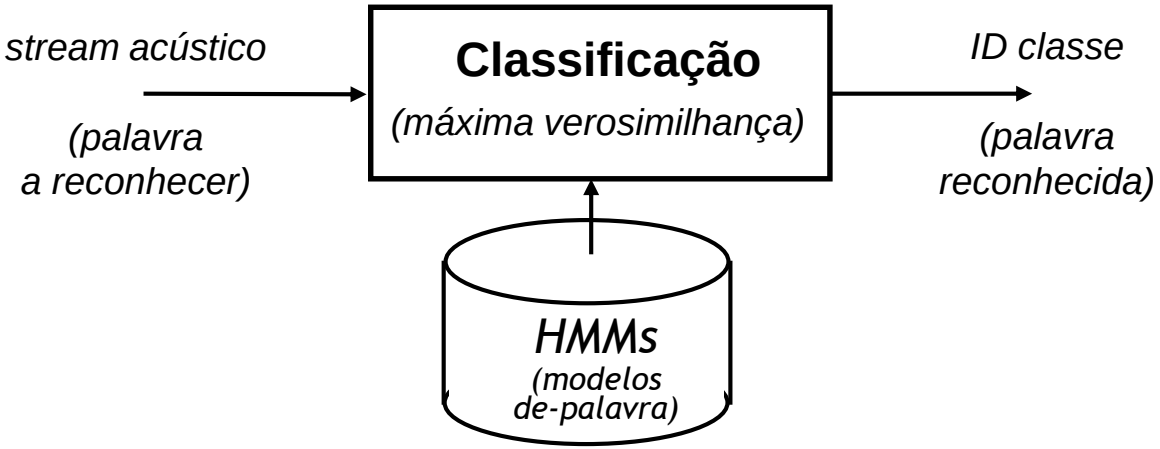
$$D_{i,j} = \min_{(u,v) \in AL_{i,j}} \{D_{u,v}\} + d_{i,j}$$

*(a menor distância parcelar em (i,j) é igual ao mínimo das menores distâncias nos nós (u,v), antecessores legais a (i,j), adicionada à distância entre os vectores $\mathbf{x}_i$ e $\mathbf{r}_j$)*

$D = D_{T,N}$ — *(resultado: menor distância entre* **X** *e* **R***)*

- O conjunto $AL_{i,j}$ dos nós antecessores legais ao nó *(i,j)* resulta de restrições locais e/ou globais (ex: o desvio do caminho relativamente à diagonal que liga os nós *(1,1)* e *(T,N)* não deve ser superior a determinado valor)
- A distância $d_{i,j}$ entre quaisquer vectores $\mathbf{x}_i$ e $\mathbf{r}_j$ é geralmente calculada através da norma euclidiana depois um processo de descorrelação e normalização das componentes dos vectores de características (ex. transformação linear Karhunen-Loève)

# Reconhecimento baseado em HMMs




- Sistema:

  - existe um HMM para cada classe (palavra do vocabulário)

- Reconhecimento:

  - para cada HMM é calculada a verosimilhança do *stream* acústico – como? ...

  - a palavra é classificada como pertencendo à classe (HMM) correspondente à **maior verosimilhança**

$$p(\mathbf{X}|\omega_c) > p(\mathbf{X}|\omega_j), \quad \forall_{j=1,\ldots,C,\ j\neq c} \quad \Rightarrow \quad \mathbf{X} \rightarrow \omega_c$$

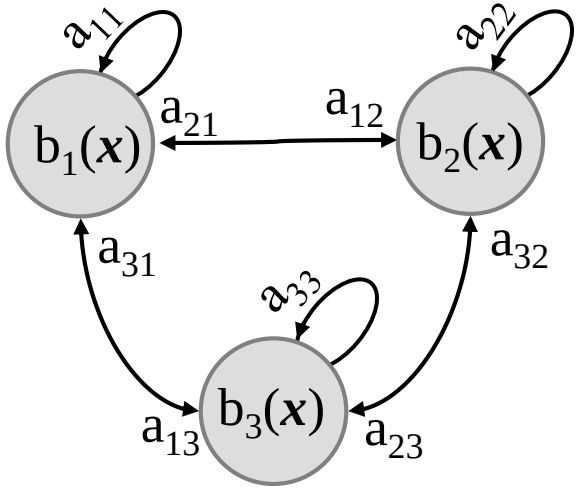
# Modelos Escondidos de Markov (HMM)

- Um HMM pode ser entendido como uma máquina de estados finita à qual estão associados dois processos estocásticos concorrentes
  - a **cadeia de Markov**
  - a **densidade de probabilidade de observação em cada estado**

- Exemplo: HMM com 3 estados
  - a "cadeia" baseia-se numa matriz com as probabilidades de transição

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

  - a "densidade" baseia-se na função $b(\boldsymbol{x})$ definida em cada estado



# HMM: Cadeia de Markov

- Este processo é representado pela variável aleatória $\boldsymbol{q}$, definida para valores de $\boldsymbol{t}$ discretos e identificando um dos $\boldsymbol{S}$ estados do modelo

$$q_t \in \Gamma_q = \{1,2,\ldots,S\}, \qquad t = 1,2,\ldots,T$$

diz-se que no instante *t* o processo estocástico associado à cadeia de Markov se encontra no estado $\boldsymbol{q_t}$ .

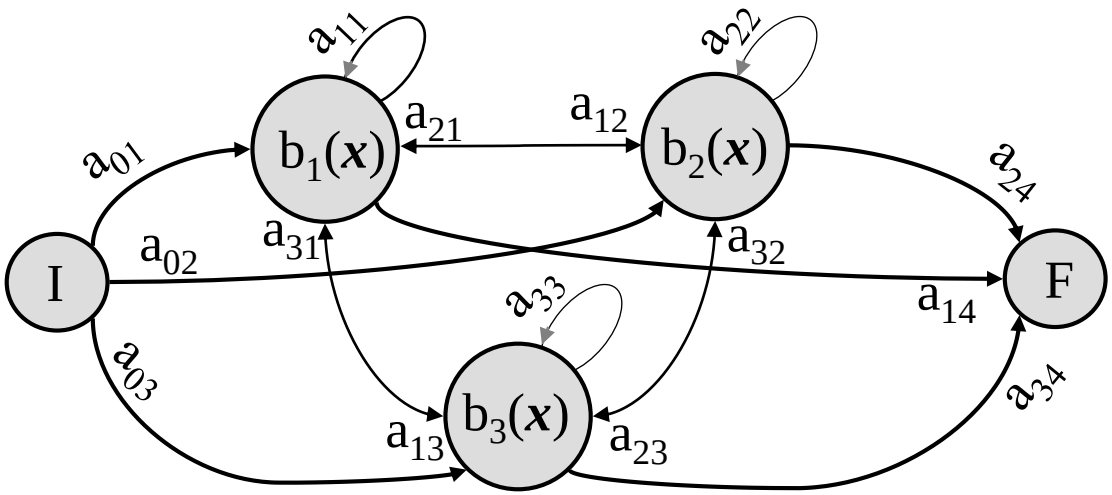
- A probabilidade de uma cadeia de Markov de primeira ordem se encontrar, num qualquer instante $\boldsymbol{t}$, no estado genérico $\boldsymbol{j}$, depende apenas do estado em que se encontra no instante imediatamente anterior;

ficam assim definidas as **probabilidades de transição entre estados**:

$$a_{i,j} = P\left(q_t = j \middle| q_{t-1} = i\right)$$

# HMM: extensão da cadeia de Markov

- Por conveniência, são arbitrados mais dois estados permitindo definir facilmente o início e o fim do processo:
  - imediatamente antes de iniciar o processo, $\boldsymbol{q}_0=0$, logo, a probabilidade do modelo ocupar o estado $\boldsymbol{k}$ no instante inicial é $\mathbf{a}_{0,k}=\boldsymbol{P}(\boldsymbol{q}_1=k)$.
  - imediatamente depois de terminar o processo, $\boldsymbol{q}_{T+1}=S+1$, pelo que apenas os estados com probabilidade de transição $\mathbf{a}_{k,S+1}$ não nula podem ocupar o estado $\boldsymbol{k}$ no instante final.



# HMM: matriz das probabilidades de transição

- As probabilidades definidas obedecem à seguinte equação

$$\sum_{j=1}^{S+1} a_{i,j} = 1, \quad i = 0,\ldots,S$$

- Sendo as probabilidades de transição entre os estados estacionárias, a cadeia de Markov é invariante no tempo

- As probabilidades de transição são geralmente organizadas na forma matricial

$$A = \begin{bmatrix} a_{0,1} & a_{0,2} & \cdots & a_{0,S+1} \\ a_{1,1} & a_{1,2} & \cdots & a_{1,S+1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{S,1} & a_{S,2} & \cdots & a_{S,S+1} \end{bmatrix}$$

Os elementos nulos da matriz, caso existam, estabelecem a topologia da cadeia de Markov

# HMM: probabilidade de um caminho

- Associado à cadeia de Markov surge o conceito de sucessão de estados, ou **caminho**

$$Q^* = \{q_1^*, q_2^*, \ldots, q_T^*\}, \quad q_t^* \in \Gamma_q$$

- A **probabilidade de um caminho particular ocorrer** (note-se que se trata apenas do processo "cadeia de Markov", não sendo ainda conhecido o sinal acústico) pode ser assim calculada

$$P(Q^*) = \prod_{t=1}^{T+1} a_{q_{t-1}^*, q_t^*} \qquad \textit{(EQ 1)}$$

- A *EQ1* evidencia uma característica da cadeia de Markov de 1ª ordem, com implicações negativas importantes na modelação da estrutura temporal do sinal da fala: a forma funcional do modelo de permanência num estado é exponencial

$$P(q_t = q_{t+1} = \ldots = q_{t+\tau} = i) = a_{i,i}^{\tau}$$

# HMM: densidade de probabilidade de observação

- Este processo estocástico modela a verosimilhança das “observações locais”, ao nível do *frame*, que o HMM realiza sobre o sinal, $\mathbf{X}_1^T$, que adere à cadeia de Markov

- Para cada estado da cadeia de Markov está definida uma **função densidade de probabilidade**, $\boldsymbol{b}(\mathbf{x})$, em que se baseia o cálculo da verosimilhança de um vector de características ser observado por esse estado

$$b_j(\mathbf{x}_t) = p(\mathbf{x}_t | q_t = j), \qquad j = 1, \ldots, S$$

- Geralmente, em qualquer estado a função densidade de probabilidade é definida com base em combinações lineares de funções de Gauss (misturas Gausseanas)

$$b_j(\mathbf{x}_t) = \sum_m c_{m,j} N(\mathbf{x}_t ; \mu_{m,j}, \Sigma_{m,j})$$

# HMM: verosimilhança de um caminho

- Nos HMM *standard* as densidades de prob. de observação são condicionalmente independentes das observações anteriores

$$p\left(\mathbf{x}_t \middle| q_t = j, X_1^{t-1}\right) = p\left(\mathbf{x}_t \middle| q_t = j\right)$$

esta assunção de independência **limita a memória** dos modelos e origina uma deficiente modelação da correlação existente entre os sucessivos vectores de características;

por outro lado, esta assunção de independência reduz muito significativamente a complexidade dos modelos;

- A verosimilhança da sucessão de observações associada a um caminho conhecido, $Q^*$ , é dada por

$$p\left(\mathbf{X}_1^T \middle| Q*\right) = \prod_{t=1}^{T} b_{q_t^*}\left(\mathbf{x}_t\right) \qquad \textit{(EQ 2)}$$

# HMM: verosimilhança total

- A densidade de probabilidade conjunta do modelo de Markov observar $\mathbf{X}_1^T$ e utilizar um caminho particular, $Q^*$, com comprimento $T$, nessa observação pode ser obtida a partir das equações *(EQ 1)* e *(EQ 2)*, resultando

$$p\left(\mathbf{X}_1^T, Q^*\right) = \prod_{t=1}^{T} \left[a_{q_{t-1}^*, q_t^*} b_{q_t^*}\left(\mathbf{x}_t\right)\right] a_{q_T^*, S+1}$$

- A verosimilhança total ou densidade probabilística do HMM observar a sucessão é obtida somando este resultado para o conjunto, $Q^P$, de todos os caminhos possíveis

$$p\left(\mathbf{X}_1^T\right) = \sum_{Q^* \in Q^P} p\left(\mathbf{X}_1^T, Q^*\right)$$

a utilização directa desta equação é impraticável porque o número de caminhos possíveis é muito grande (cresce exponencialmente com $T$)

# HMM: algoritmo "*forward*"

- Permite **calcular eficientemente** $p\left(\mathbf{X}_1^T\right)$, baseando-se no cálculo recursivo da chamada probabilidade *forward*

$$\alpha_{j,t} = p\left(\mathbf{X}_1^t, q_t = j\right)$$

é a probabilidade conjunta do modelo observar a sucessão parcial dos vectores de características $\mathbf{X}_1^t = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t\}$ e de se encontrar no estado $j$ no instante $t$

*algoritmo forward*

*para* $j$=1,…,$S$ *(inicia algoritmo)*

$$\alpha_{j,1} = b_j(\mathbf{x}_1) a_{0,j}$$

*para* $t$=2,…,$T$ *(cálculo recursivo da prob. forward)*

*para* $j$=1,…,$S$

$$\alpha_{j,t} = b_j(\mathbf{x}_t) \sum_{i=1}^{S} \alpha_{i,t-1} a_{i,j}$$

$$p\left(\mathbf{X}_1^T\right) = \alpha_{S+1,T+1} = \sum_{i=1}^{S} \alpha_{i,T} a_{i,S+1}$$ *(resultado)*

# HMM: algoritmo "*Viterbi*"

- Aproxima o resultado obtido pelo algoritmo "*forward*", considerando apenas o "melhor caminho" de comprimento $T$

$$p_{Q^*}\left(\mathbf{X}_1^T\right) = \max_{Q^* \in Q_T^P} \left\{ p\left(\mathbf{X}_1^T, Q^*\right) \right\} \approx p\left(\mathbf{X}_1^T\right)$$

- É o algoritmo *standard* de em reconhecimento de fala contínua

*algoritmo Viterbi*

*para* $j=1,\ldots,S$ *(inicia algoritmo)*

$$\alpha_{j,1} = \log\left[ b_j\left(\mathbf{x}_1\right) a_{0,j} \right]$$

Log dispensa re-escalar as probabilidades (problemas numéricos)

*para* $t=2,\ldots,T$ *(cálculo recursivo da prob. forward)*

*para* $j=1,\ldots,S$

$$\alpha_{j,t} = \log\left[ b_j\left(\mathbf{x}_t\right) \right] + \max_i \left\{ \alpha_{i,t-1} + \log\left(a_{i,j}\right) \right\}$$

$$p\left(\mathbf{X}_1^T\right) = \alpha_{S+1,T+1} = \max_i \left\{ \alpha_{i,T} + \log\left(a_{i,S+1}\right) \right\} \quad \textit{(resultado)}$$

# HMM: Treino

- É o processo de determinação do conjunto dos parâmetros livres, $\theta$
  - a matriz das probabilidades de transição entre estados
  - os parâmetros que em cada estado definem a função densidade de probabilidade para o cálculo da verosimilhança local

- É utilizada a informação contida num conjunto de exemplos

$$U = \left\{ \left( \mathbf{X}^n, \omega_c^n \right) n = 1,..,N_T \right\}$$

cada exemplo é constituído por uma sucessão de vectores de características e a identidade da palavra correspondente (treino supervisionado)

- Destacam-se dois critérios de treino:
  - critério da máxima probabilidade a posteriori (MAP)
  - critério da máxima verosimilhança (ML)

# HMM: Treino pelo critério MAP

- **Critério da máxima probabilidade *a posteriori (MAP)***
  - os parâmetros são determinados de maneira a maximizar a probabilidade *a posteriori* estendida ao conjunto de treino

$$\theta = \arg\max_{\theta} \prod_{n=1}^{N_T} P_\theta\left(\omega_c^n | \mathbf{X}^n\right)$$

  - é discriminativo, treinando os modelos para representar bem a sua classe e mal as restantes

$$\theta = \arg\max_{\theta} \prod_{n=1}^{N_T} \frac{p_\theta\left(\mathbf{X}^n | \omega_c^n\right)}{\sum_k p_\theta\left(\mathbf{X}^n | \omega_k^n\right)}$$

(pela regra de Bayes; e admitindo palavras equiprováveis)

  - exige enorme esforço computacional ...

# HMMs – Treino pelo critério ML

- **Critério da máxima verosimilhança (ML)**
  - permite uma solução aproximada ao critério MAP, reduzindo muito a complexidade do treino, embora à custa de capacidade discriminativa
  - os parâmetros passam a ser determinados de acordo com o critério de maximização da verosimilhança estendida ao conjunto de treino

$$\theta = \arg\max_{\theta} \prod_{n=1}^{N_T} p_\theta\left(\mathbf{X}^n | \omega_c^n\right)$$

  - cada modelo é apenas treinado com os exemplos da sua classe (não precisa "conhecer" os exemplos das classes restantes, para representá-las mal), existindo dois algoritmos básicos (não serão aqui detalhados)
    - ***Baum-Welsch***: em cada exemplo são considerados todos os alinhamentos possíveis entre a sucessão de vectores e os estados
    - ***Segmental k-means (alg. Viterbi)***: em cada exemplo é apenas considerado o "melhor caminho"

# HMMs – Fórmulas de treino (caso SCHMM)

- Iterativamente:

$$a_{i,j} = \frac{\sum_{n=1}^{N_\tau} \sum_{t=1}^{T^n-1} \delta(q_t = i, q_{t+1} = j)}{\sum_{n=1}^{N_\tau} \sum_{t=1}^{T^n-1} \delta(q_t = i)}$$

$$\gamma_{j,m}(\boldsymbol{x}) = \frac{c_{j,m} N_{\alpha_{j,m}}(\boldsymbol{x})}{b_j(\boldsymbol{x})}$$

$$c_{j,m} = \frac{\sum_{n=1}^{N_\tau} \sum_{t=1}^{T^n} \delta(q_t = j) \gamma_{j,m}(\boldsymbol{x}_t)}{\sum_{n=1}^{N_\tau} \sum_{t=1}^{T^n} \delta(q_t = j)}$$

$$\boldsymbol{\mu}_m = \frac{\sum_{n=1}^{N_\tau} \sum_{t=1}^{T^n} \sum_{j} \delta(q_t = j) \gamma_{j,m}(\boldsymbol{x}_t) \boldsymbol{x}_t}{\sum_{n=1}^{N_\tau} \sum_{t=1}^{T^n} \sum_{j} \delta(q_t = j) \gamma_{j,m}(\boldsymbol{x}_t)}$$

$$\boldsymbol{\sigma}_m^2 = \frac{\sum_{n=1}^{N_\tau} \sum_{t=1}^{T^n} \sum_{j} \delta(q_t = j) \gamma_{j,m}(\boldsymbol{x}_t) (\boldsymbol{x}_t - \boldsymbol{\mu}_m)^2}{\sum_{n=1}^{N_\tau} \sum_{t=1}^{T^n} \sum_{j} \delta(q_t = j) \gamma_{j,m}(\boldsymbol{x}_t)}$$

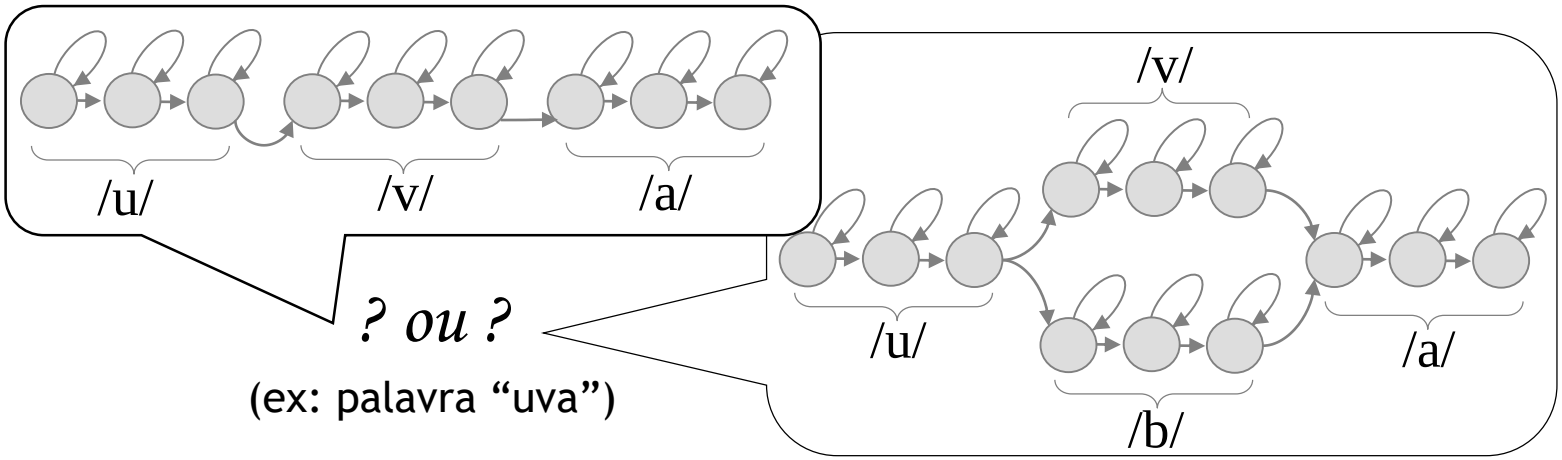
# HMMs na modelação acústica

- **Objectivo**: construir um HMM para modelar cada palavra do vocabulário

$$\omega_c \leftrightarrow (HMM)_c, \quad c=1,\ldots,C$$

- **Questões essenciais**

1. **Qual a topologia da cadeia de Markov?**

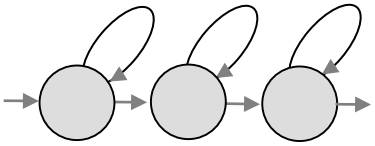



2. **Como modelar a verosimilhança em cada estado?**

   espaço acústico discreto ou contínuo? misturas baseadas em que formas funcionais básicas? partilha de parâmetros entre estados do mesmo modelo e/ou entre modelos? ...

# HMMs na modelação acústica – Topologia

- **A Topologia da cadeia de Markov** deve adequar-se à estrutura temporal das realizações acústicas mais significativas da palavra
- O número de estados deve ser suficiente para modelar com precisão os diferentes segmentos do sinal (é típico usar um modelo Bakis com 3 estados para cada símbolo da transcrição ortográfica da palavra)



- A complexidade das interligações deve permitir variações devidas a múltiplos falantes, aos diferentes "acentos" regionais, etc. (embora frequentemente se defina apenas a estrutura adequada à pronunciação *standard*, sendo a variabilidade modelada por "misturas" mais complexas em cada estado)
- Nota: nos Modelos Escondidos Semi-Markov modela-se explicitamente a distribuição estatística de permanência em cada estado, obtendo-se uma modelação temporal mais precisa dos períodos de quase-estacionaridade do sinal

# HMMs na modelação acústica – Verosimilhança

O processo de cálculo da verosimilhança de observação do sinal acústico determina, geralmente, um de três tipos diferentes de HMM: Discretos, Contínuos, ou Semi-Contínuos

- **HMM Discreto (DHMM)**
  - o espaço acústico é "partido" em sub-domínios disjuntos, cada qual representado por um centróide (o conjunto de todos os centróides define o ***codebook***);

    em cada estado está definido o "peso" de cada sub-domínio
  - no reconhecimento, em cada estado:

    1) o vector de é atribuído ao sub-domínio com o centróide mais próximo

    2) e a verosimilhança de observação toma o valor do respectivo "peso"

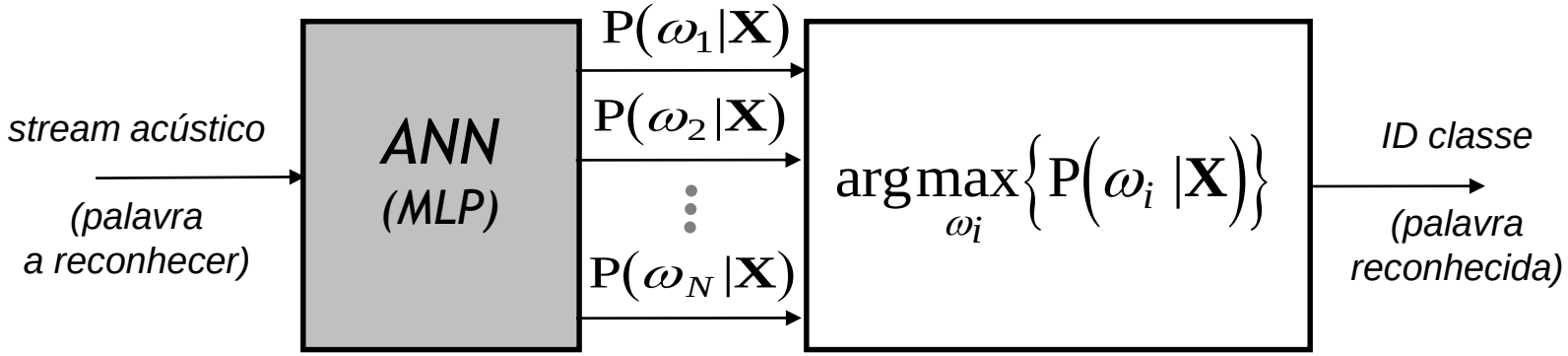
  - **vantagem**: dispensa a avaliação de funções de Gauss (ou outras)
  - **desvantagens**: erro de quantificação vectorial; falta de "suavidade" dos modelos; *codebook* estabelecido no início e geralmente não submetido a qualquer processo de adaptação

# HMMs na modelação acústica – Verosimilhança

- **HMM Contínuo (CHMM)**
  - em cada estado existe uma mistura de funções elementares (em geral é uma mistura gausseana, ou seja, uma combinação linear de funções de Gauss) específica desse estado
  - no reconhecimento, para o vector de características dado, o valor dessa mistura é avaliado
  - **vantagem**: em princípio, é o método que permite modelos acústicos mais precisos se os dados de treino são suficientes
  - **desvantagem**: grande exigência de dados de treino; elevados custos de processamento devido à avaliação das misturas específicas de cada estado

- **HMM Semi-contínuo (SCHMM)**
  - existe um único conjunto de funções elementares (geralmente, funções de Gauss) – chamado *codebook* - que são partilhadas por todos os estados
  - são geralmente um bom compromisso entre a precisão dos CHMM e a acentuada partilha de parâmetros dos DHMM

# Reconhecimento baseado em MLPs




- Classificação é baseada num processo com duas fases
  - conhecida a sequência de *frames* correspondentes à palavra a reconhecer (*stream* acústico), a rede neuronal calcula a probabilidade ***a posteriori*** relativa a cada classe (palavra do vocabulário)
  - conhecidas as probabilidades ***a posteriori*** de todas as classes, o módulo final de decisão selecciona, como resposta do sistema, a classe correspondente ao maior valor de probabilidade ***a posteriori***

# HMMs *versus* MLPs no Reconhecimento da Fala

- Características positivas dos ANN para o RAF (podem ser vantagens, relativamente aos HMM)
  - o critério *standard* de treino é discriminativo
  - podem, em princípio, estabelecer qualquer transformação não linear da entrada (muito importante no RAF, onde as superfícies de decisão apresentam formas muito complexas)
  - dispensam muitas das assunções sobre diferentes partes do sistema que em outras tecnologias é necessário formular previamente
  - podem acomodar facilmente entradas contextuais e realimentação

- Grande desvantagem dos ANN, relativamente aos HMM, no RAF:
  - inaptidão para suportar adequadamente a dimensão temporal do sinal da fala, onde é necessário classificar padrões de comprimento variável e apresentando distorções no alinhamento temporal

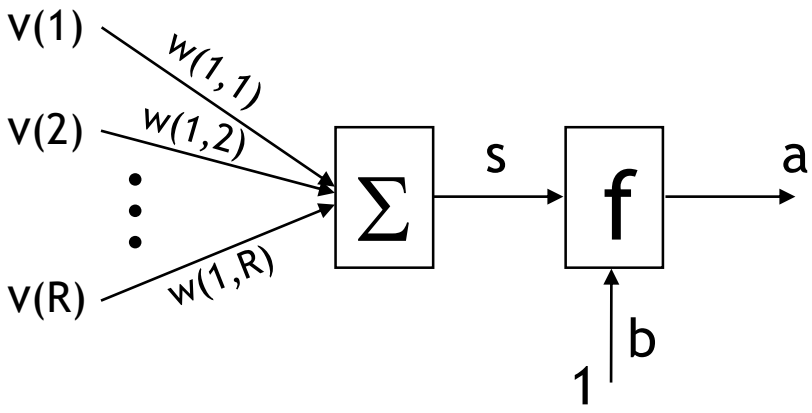
    (soluções como os TDNN, com elevado potencial em algumas aplicações, não se têm revelado consistentes em sistemas de maior dimensão)

# Híbridos HMM/MLPs no Reconhecimento da Fala

- Abordagens híbridas tentam combinar as características mais favoráveis das tecnologias HMM e ANN
  - os HMM são utilizados na modelação temporal do sinal através do processo “cadeia de Markov”
  - os ANN são utilizados na modelação local (em cada estado da “cadeia de Markov”) do espaço acústico (em substituição das tradicionais misturas gausseanas e permitindo uma utilização mais eficaz da informação de contexto)
- Os resultados obtidos em algumas aplicações superam os alcançados por sistemas baseados apenas nos HMM (actualmente, a tecnologia *standard*)

# Neurónio Artificial

- Neurónio com R entradas e *bias* (b)



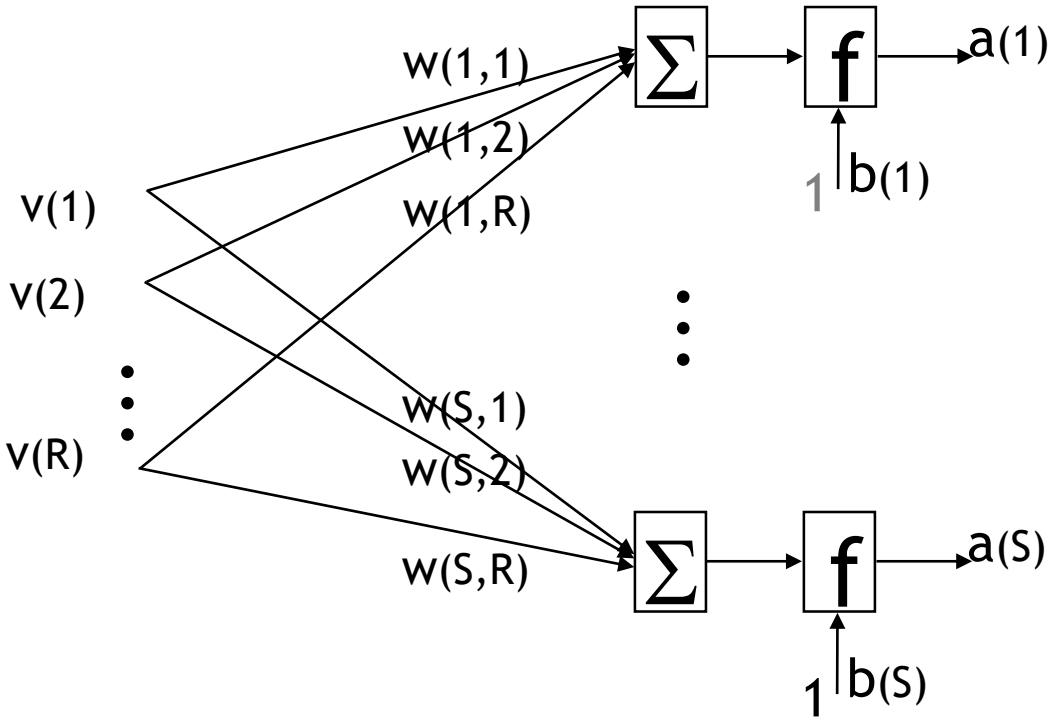

$$s = \sum_{i=1}^{R} w(1,i)\, v(i) = W*V$$

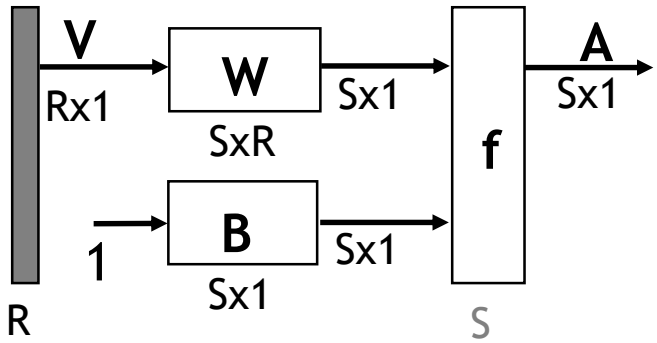
$$a = f(W*V + b)$$

se f é a função *log-sigmoide*

$$a = \frac{1}{1 + e^{-(W*V+b)}}$$

# Rede com uma camada de neurónios

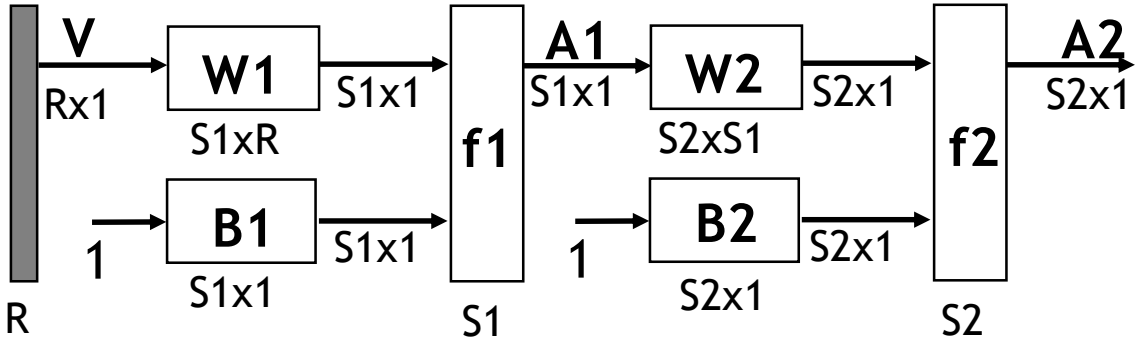



$$A = f(W*V+B)$$

# Rede com duas camadas* de neurónios

- S1 neurónios na 1ª camada e S2 na 2ª camada




$$A2=f2(W2*f1(W1*V+B1)+B2)$$

- Um MLP de 2 camadas - por exemplo, com a 1ª composta por neurónios *log-sigmóide* e a 2ª por neurónios *lineares* - pode aproximar qualquer função, linear ou não (embora a precisão possa ser limitada por aspectos práticos como, por exemplo, o limite de dados ou o tempo de treino)

* em geral, c múltiplas camadas de neurónios, ou multi-layer perceptron (MLP)

# MLPs: Treino

- O critério de treino é discriminativo (...) baseando-se na minimização de uma medida da distância entre a saída calculada pelo MLP em resultado da matriz *Exemplos* e a matriz *Alvo*

$$E = \sum_{q=1}^{Q} \sum_{n=1}^{N} \left[ A2(n,q) - Alvo(n,q) \right]^2$$

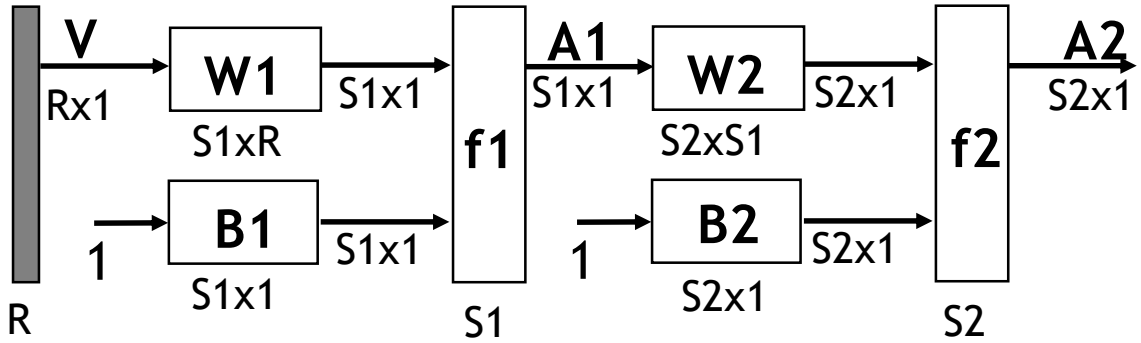
- A minimização de ***E*** , através da adaptação dos parâmetros **W1**, **W2**, **B1** e **B2**, é geralmente efectuada pelo algoritmo ***backpropagation***, o qual se baseia numa técnica de gradiente
- Técnicas para acelerar o treino: inicialização Nguyen-Widrow; *learning-rate* adaptável; momento
- Situações a evitar
  - sub-treinamento
  - sobre-treinamento

# MLPs no reconhecimento de palavras isoladas

- Tarefa: IWR com vocabulário de **N** palavras
- A representação de cada palavra deve ter igual "comprimento": **R=LxD** (necessidade de técnicas de "segmentação do traço")

$$\mathbf{V} = \left[\mathbf{x}_1^1, \mathbf{x}_1^2, ..., \mathbf{x}_1^D,\ \mathbf{x}_2^1, \mathbf{x}_2^2, ..., \mathbf{x}_2^D,\ .....,\ \mathbf{x}_L^1, \mathbf{x}_L^2, ..., \mathbf{x}_L^D\right]^T$$

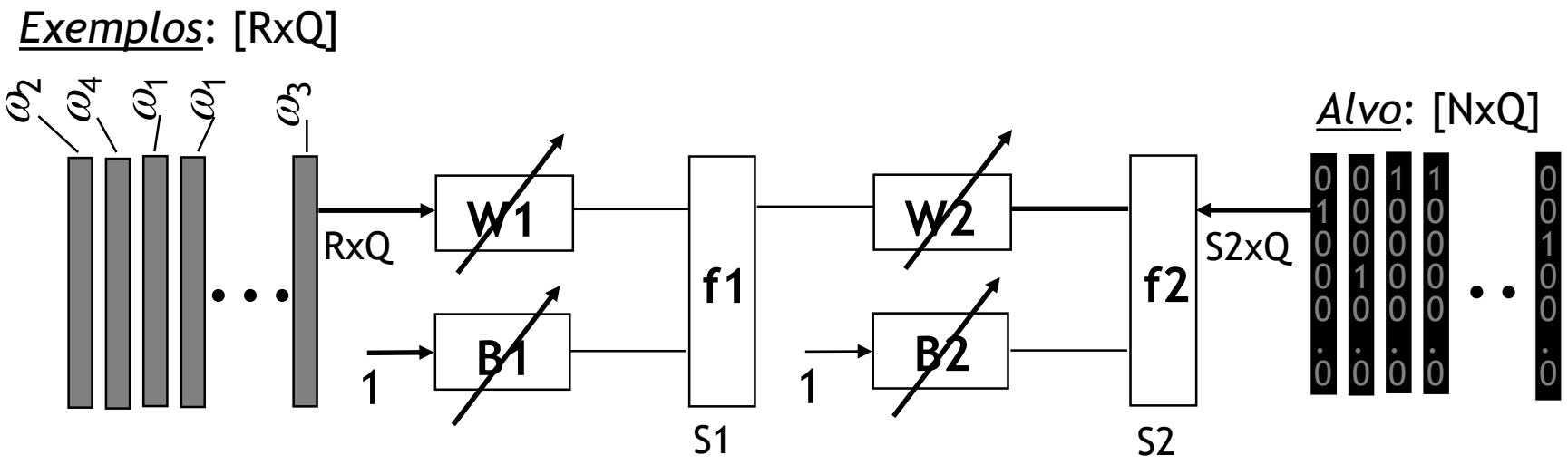
- Utilizando um MLP com 2 camadas para o classificador
  - **S1** deve ser suficiente para modelar as (complexas) fronteiras de classificação no espaço de representação acústica
  - **S2** deve ser igual ao número de classes, **N** (sendo reconhecida a classe que corresponde ao maior elemento de **A2**)



# MLPs no reconhecimento de palavras isoladas

- Os parâmetros do MLP
  - **W1** : matriz (S1xR) dos "pesos" da 1ª camada
  - **B1** : vector (S1x1) "bias" da 1ª camada
  - **W2** : matriz (S2xS1) dos "pesos" da 2ª camada
  - **B2** : vector (S2x1) "bias" da 2ª camada

  são treinados a partir de um conjunto de exemplos
- No treino ***batch***, em cada ciclo de treino são "processados" todos os exemplos, só então os parâmetros são adaptados



# Referências principais

Reconhecimento da Fala, Diamantino Freitas, Vitor Pêra, apontamentos Processamento da Fala 2009/2010, DEEC-FEUP.