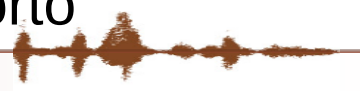
Áudio Computacional

Reconhecimento automático da Fala (reconhecimento de fala contínua, bases de dados)

Mestrado Integrado em Engenharia Electrotécnica e Computadores

Faculdade de Engenharia da Universidade do Porto

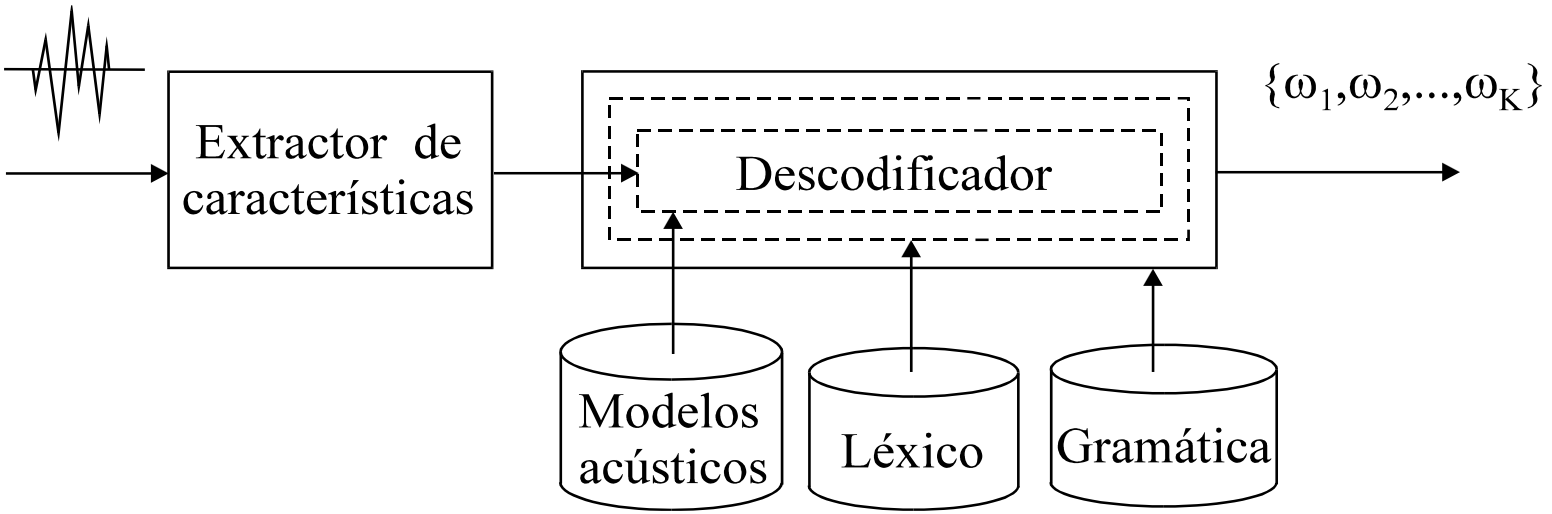


Reconhecimento de fala-contínua

Perspectiva histórica

- Sistemas de reconhecimento de palavras isoladas (IWR)
- Sistemas de reconhecimento *connected-speech* (ConectedSR)
  - modelo acústico
    - vocabulário pequeno e recursos de treino suficientes para …
    - treinar modelos-de-palavras
  - modelo linguístico
    - aplicações “artificiais”, com gramáticas sintáticas muito rudimentares (ex: descodificador *one-state* com mecanismo *level-building* para reconhecimento de sequências de palavras com comprimento fixo)
    - tentativas (com resultados muito deficientes) para introduzir conhecimento aos níveis semântico e pragmático
- **Sistemas de reconhecimento de fala contínua** (CSR)
- Sistemas de reconhecimento de fala espontânea

Sistema *standard* para CSR



- É descodificada a frase correspondente à hipótese de reconhecimento que apresenta o maior valor do produto:

probabilidade *apriori* (mod. linguístico: gram, léxico) X verosimilhança (modelo acústico)

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

Unidades acústicas elementares

- A **dimensão elevada do vocabulário** (para além de aumentar a confusão entre as palavras) impossibilita a modelação individual de cada palavra, devido a limitações das bases de dados para treino e do espaço de memória para armazenar os modelos
- Para atenuar estes problemas são geralmente usados **modelos acústicos elementares** - **princípio da partilha dos parâmetros** -, geralmente definidos ao nível fonético (opção natural dado o vasto conhecimento apriorístico que existe sobre os fonemas), ou silábico, duo-silábico, UAs, etc.
- Mas o aumento dos dados disponíveis para treinar cada modelo tem, como contrapartida, a sua **perda de nitidez**
- Torna-se necessário modelar as unidades elementares com **informação de contexto** (compromisso entre o treino robusto e a sensibilidade aos efeitos de co-articulação): tri-fonemas, di-fonemas; fonemas dependentes da palavra; etc. ...
- Implicando o uso de técnicas para controlar o elevado número de unidades elementares, quando modeladas com contexto

Modelos acústicos elementares

- Para obter uma boa aproximação ao classificador Bayesiano é fundamental que o modelo acústico global estime correctamente a verosimilhança global (sobretudo quando o modelo não é treinado discriminativamente)
- O modelo acústico global é criado concatenando, de acordo com a gramática e o léxico do sistema, os modelos acústicos correspondentes às unidades elementares
- A tecnologia actualmente dominante na modelação das unidades acústicas elementares baseia-se nos modelos escondidos de Markov (dos tipos SC-HMMs e C-HMMs)
- As fraquezas mais significativas dos HMM têm sido identificadas e algumas soluções têm sido propostas, mas por regra tornam os sistemas substancialmente mais “pesados” pelo que são geralmente evitadas
- Destaque para os “modelos de segmentos” (SSM) e os sistemas híbridos HMM-ANN

Léxico e Gramática

- A gramática e o léxico devem representar as dependências sintácticas e lexicais que se verificam na realidade (apesar dos modelos linguístico e acústico serem por norma treinados isoladamente, devem articular-se para permitir a correspondência correcta entre os modelos acústicos e as observações reais)
- A introdução de restrições linguísticas no reconhecimento reduz a **perplexidade** associada ao processo de decisão, conduzindo a melhores taxas de reconhecimento
- **Gramáticas**: determinísticas e estocásticas (N-Grams; têm sido propostas gramáticas capazes de modelar dependências muito complexas, e de longo alcance, entre as palavras, mas geralmente apresentam um elevadíssimo numero de parâmetros)
- **Léxicos**: apenas uma transcrição fonética por palavra; ou de pronunciação múltipla (a fala contínua não adere geralmente às formas canónicas de pronunciação); modelos estocásticos, treinados de maneira automática, ou baseados num conjunto de regras fonológicas extraídas do conhecimento apriorístico.

Treino e Descodificação

- Treino *standard* de reconhecedores baseados em HMMs
  - "semear" os modelos (exige apenas um conjunto limitado de dados acústicos previamente segmentados e "etiquetados")
  - treino embebido dos modelos (grande vantagem: dispensa a prévia segmentação e anotação da BD)
- No processo de treino aceita-se
  - o treino independente dos modelos acústico e linguístico, o que é reconhecido ser uma solução sub-óptima
  - a aproximação do critério da máxima probabilidade *a posteriori* pela máxima verosimilhança (degrada muito a capacidade discriminativa dos modelos acústicos, mas a solução ideal é impraticável)
  - (frequentemente) a aproximação do critério da máxima verosimilhança (nos HMMs, a tecnologia *standard*), considerando apenas o "melhor caminho" do modelo acústico global (o treino Viterbi é bastante mais rápido que o treino Baum-Welsh)

Treino e Descodificação

- Treino *on-line*: em alguns sistemas, posteriormente ao procedimento de treino inicial, os parâmetros são novamente adaptados *on-line* utilizando um conjunto de dados reduzido

- O problema da descodificação consiste na pesquisa do espaço das frases possíveis de maneira a encontrar aquela que corresponde à maior probabilidade conjunta
- Uma busca exaustiva é apenas praticável em problemas de pequeno vocabulário
- Um sistema baseado numa estrutura integrada de conhecimento permite a utilização de algoritmos eficientes de pesquisa (como o algoritmo Viterbi), assentes em técnicas de programação dinâmica para pesquisa de grafos
- Técnicas “suplementares” (pesquisa em feixe; descodificação em passos múltiplos, etc.) são geralmente usadas para reduzir muito o espaço de busca, sem deteriorar significativamente as taxas de reconhecimento

Estado da arte

- Caracterização dos sistemas do ponto de vista aplicacional
  - léxicos com **mais de 20K palavras**
  - **independência do falante**
  - **ambiente acústico favorável**
- Soluções técnicas
  - **arquitecturas integradas de conhecimento**
  - decisão baseada em **formalismos estatísticos**
  - representação do sinal acústico baseada na **análise espectral** e **cepstral** do sinal, sendo associadas **características dinâmicas**
  - modelo acústico baseado na tecnologia dos **HMM** ou numa abordagem híbrida **HMM/ANN**
  - **unidades acústicas elementares** definidas ao nível **fonético** e modeladas com **contexto** (até penta-fonemas!)
  - modelo linguístico baseados em **gramáticas do tipo 3-Gram** (**ou** até mesmo gramáticas mais complexas)
  - **léxico** composto por **modelos estocásticos** e/ou **regras fonológicas**
  - **enormes bases de dados**, acústicos e de texto, são utilizados para treinar estes sistemas.

Bases de dados
e
ferramentas de desenvolvimento

Bases de dados – fases de desenvolvimento

- Especificação (*database design)*
  - Conteúdo (modo de fala, léxico, nº falantes, etc.)
  - Ambiente acústico (constante ou variável, níveis de ruído, etc.)
  - Dispositivo de gravação (nº microfones, posição, sampling rate, conversão AD, etc.)
  - Formato e estrutura (organização dos dados acústicos, informação de segmentação e anotação, etc.)
- Gravação (aquisição material acústico ou áudio-visual)
- Segmentação e anotação (eventualmente, segmentação ao nível da frase, da palavra, do fonema, etc.; transcrições a diversos níveis, podendo incluir anotação prosódica)
- Documentação (produção de um *manual* da base de dados)
- Validação (um centro independente analisa todo o material – sinais, informação de segmentação e anotação, documentação, etc. - da base de dados construída e, de acordo com os critérios de validação estabelecidos, produz um relatório de avaliação e valida ou rejeita a base de dados

Bases de dados e ferramentas de desenvolvimento

- Bases de dados para processamento da fala – exemplos de catálogos na Web
  - http://www.ldc.upenn.edu/
  - http://cslu.cse.ogi.edu/corpora/corpCurrent.html

    ...
- Ferramentas de desenvolvimento – exemplos na Web
  - http://htk.eng.cam.ac.uk/ - **HTK** (Hidden Markov Model Toolkit)
  - http://cslu.cse.ogi.edu/toolkit/ - CSLU Toolkit
  - http://www.cstr.ed.ac.uk/projects/speech_tools/ - Edinburgh Speech Tools Library
  - http://cmusphinx.sourceforge.net/html/cmusphinx.php - CMU Sphinx Group Open Source Speech Recognition Engines
  - Matlab
    - http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html - Voicebox toolbox
    - http://www.ai.mit.edu/~murphyk/Software/HMM/hmm.html - HMM Toolbox

    ...

Referências principais

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