
Técnicas Avanzadas de Sistemas Inteligentes:
Reconocimiento Automático de Habla

Máster Universitario de
Investigación en TIC

Valentín Cardeñoso Payo
Universidad de Valladolid

Para hoy ...

- Introducción al uso de habla en interfaces
- Qué es ASR
- Aplicaciones
- Dificultades
- Estrategias de solución
- Modelos ocultos de Markov

Lenguaje Natural

- Medio de comunicación genuinamente humano
 - Naturalidad
 - Modo más natural ‘cara a cara’
- Variantes posibles y aspectos relevantes:
 - Escrito
 - Pantalla
 - Teclado / ratón / ...
 - Hablado
 - Altavoces
 - Micrófono

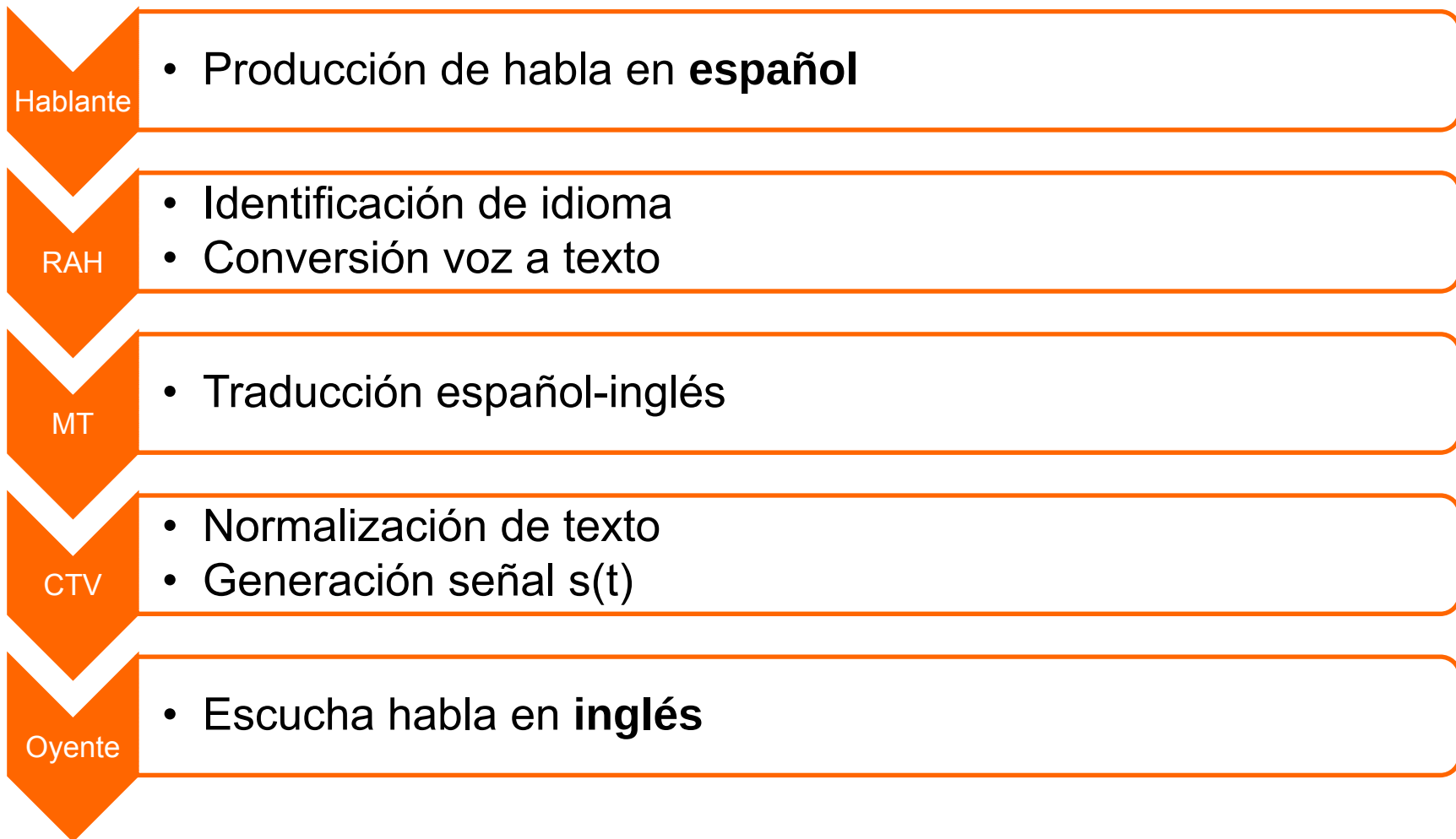
Lenguaje Natural Escrito

- Salida:
 - Tipos de letra
 - Distribución espacial contenidos
 - Iconografía
 - Generación de texto
- Entrada:
 - Juego de caracteres
 - Elementos gráficos (teclado/ratón):
 - Botones
 - Desplegables
 - Listas
 - Reconocimiento/interpretación del guión escrito

Lenguaje Natural Hablado

- Salida (altavoces):
 - Conversión Texto-Voz:
 - Normalización texto
 - Anotación supra-segmental (prosodia/entonación)
 - Síntesis de voz (generación de señal de salida)
- Entrada (micrófono[s])
 - Reconocimiento Automático de Habla
 - Transcripción voz->texto (Decodificación)
 - Captura y acondicionamiento de señal
 - Extracción de características
 - Selección de mejores candidatos
 - Emisión de representación textual asociada

Traducción voz-voz?



RAH(ASR): A vista de pájaro

- Voz:
 - Fenómeno ondulatorio
 - Energía
 - Frecuencia
- Acondicionamiento
 - Paso banda
 - Eliminación ruidos
 - Normalización
- Estacionareidad:
 - Secuencia vectores $p(t)$
- Comparación con DB
 - Modelado
 - Secuencia óptima

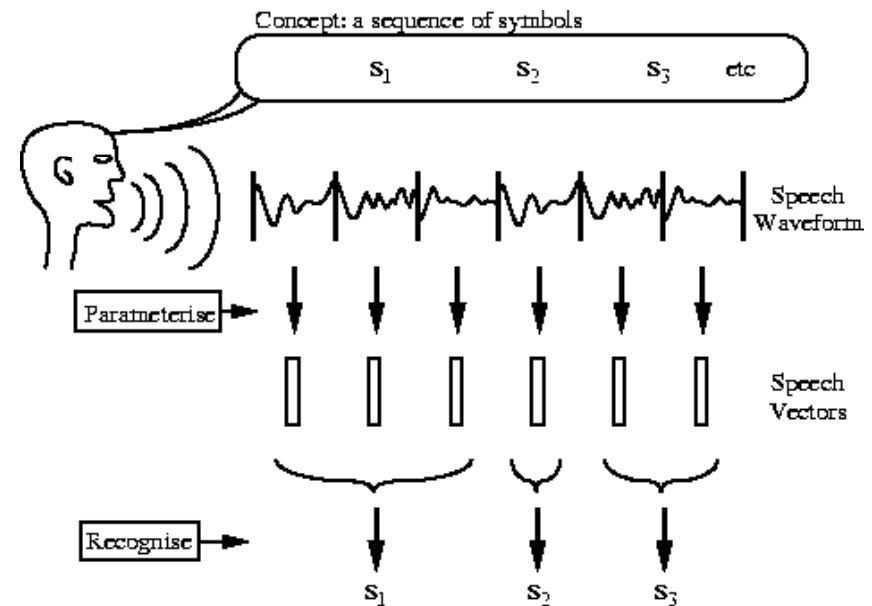


Fig. 1.1 Message Encoding/Decoding

source: HTK book

Aplicaciones

- Máquina de dictado (transcripción)
- Control y gobierno (Command & Control)
- Interfaz hablado con la computadora.
- Dispositivos activados por voz:
 - Teléfono, TV, VCR, ...
- Entornos ‘manos y ojos libres’:
 - Conducción vehículos
 - Sala de control seguridad crítica
- Ayuda a los discapacitados:
 - Dispositivos ayuda a la movilidad
 - Sustitución de dispositivos hápticos
- Detección de palabras clave (Keyword spotting)
- Resumen de documentos sonoros
- Sistemas de Información y Reserva
 - Banca, Viajes, Operadoras telefónicas

Para el cuaderno ...

- Localizar en la web:
 - Servicios vocales para TV o móvil:
 - Viajes
 - Comida
 - Películas
 - ...
- Servicios de atención vocal:
 - Experiencia personal
 - Ejemplos conocidos

Variantes del problema

- Tipo de habla:
 - Palabras aisladas (IWR)
 - Palabras conectadas (CWR)
 - Habla continua (CSR)
 - Habla espontánea
 - Detección de palabras clave
- Dependencia del hablante
 - Dependiente de hablante
 - Adaptable al hablante
 - Independiente de hablante
 - Multilocutor

Variantes del problema ...

- Talla del vocabulario
 - Pequeño (<100 palabras)
 - Mediado ($100 < x < 1000$)
 - Grande ($x \sim 1000$)
 - Muy grande ($x \sim > 10000$)
 - Vocabulario abierto (OOV)
- Canal
 - Escritorio / Estudio / Oficina
 - Exterior / Entorno ruidoso (factoría, vehículo)
 - Banda estrecha (móvil)

Dificultad del ASR

- **Variabilidad** (ambigüedad-perplejidad):
 - Dependiente de hablante: fisiológica, emocional, cultural.
 - Continuidad de la señal: ¿fronteras de segmentos?
 - Segmento = unidad lingüística (fonema, difonema, sílaba, ...)
 - Ambiente: ruido, efecto Lombard, acústica del medio
 - Canal: compresión / ruido aditivo / convolución
 - Dispositivo: direccionalidad / tipo de sensor / ...
 - Contexto fonético

Modelo esquemático de fonación

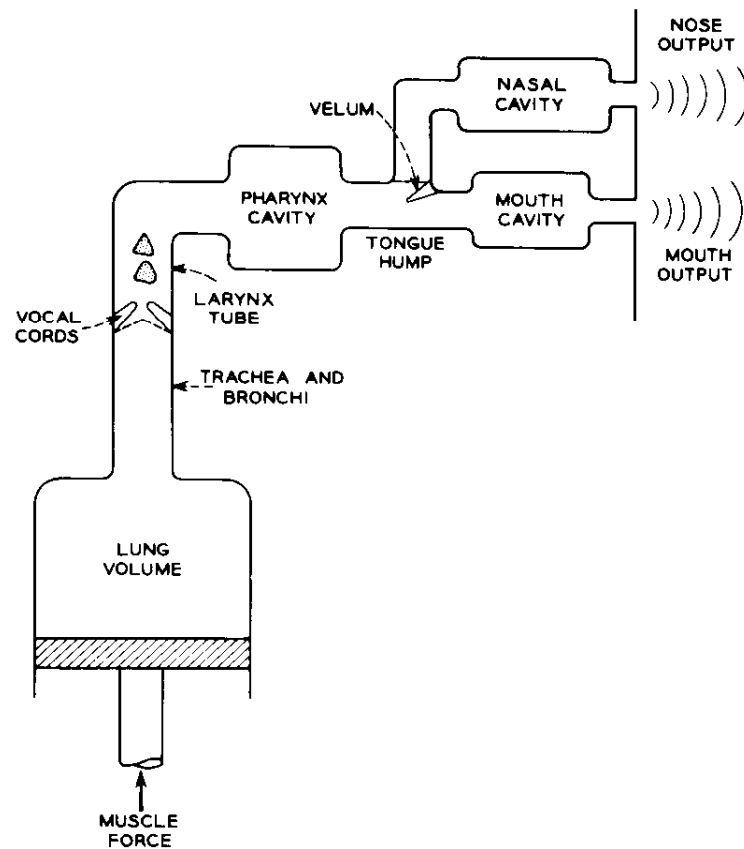


Figure 2.6 Schematic representation of the complete physiological mechanism of speech production (after Flanagan [3]).

De la voz a los patrones

- Tratamiento en dominio de tiempo
 - Acondicionamiento
 - Filtrado
- Espectrogramas
- Cepstrum
- FFT de tiempo corto
- Comparación de plantillas
 - Cuantificación vectorial
 - Alineamiento temporal (lineal y no lineal)

ASR: Estrategias de solución

- Acústico-Fonética: Basadas en reglas (heurísticas)
 - Clasificación acústico-fonética
- Clasificación
 - VQ
 - K-NN
 - ANN
- Ajuste de patrones (PM o PR)
 - Deterministas:
 - Alineamiento lineal
 - Pandeo dinámico temporal (DTW)
 - Estocásticos:
 - HMM
- Técnicas de AI:
 - Integración de conocimiento
 - Niveles estratificados

**RECONOCER
SECUENCIAS**

ASR: Aproximación fonética acústica

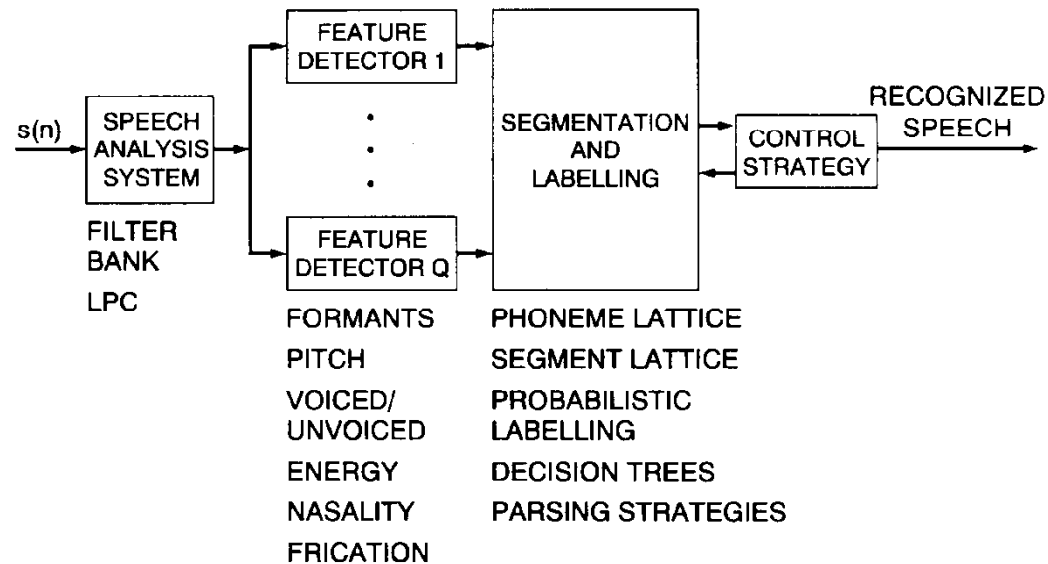


Figure 2.32 Block diagram of acoustic-phonetic speech-recognition system.

ASR: Rasgos acústico-fonéticos

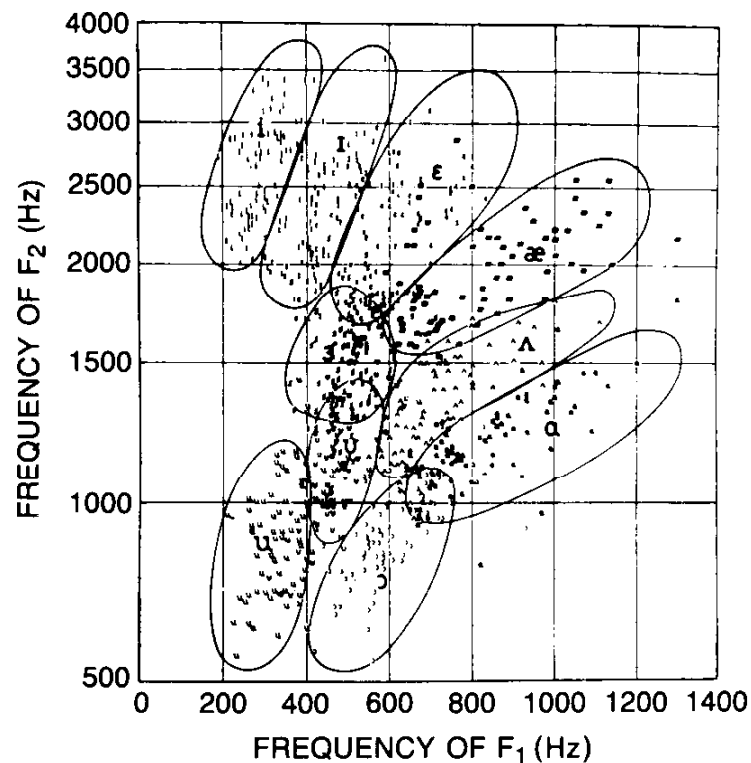


Figure 2.16 Measured frequencies of first and second formants for a wide range of talkers for several vowels (after Peterson & Barney [7]).

- Mapa vocálico
 - Diferenciación sonidos
 - Formantes (frecuencia)

ASR: Reconocedor palabras aisladas

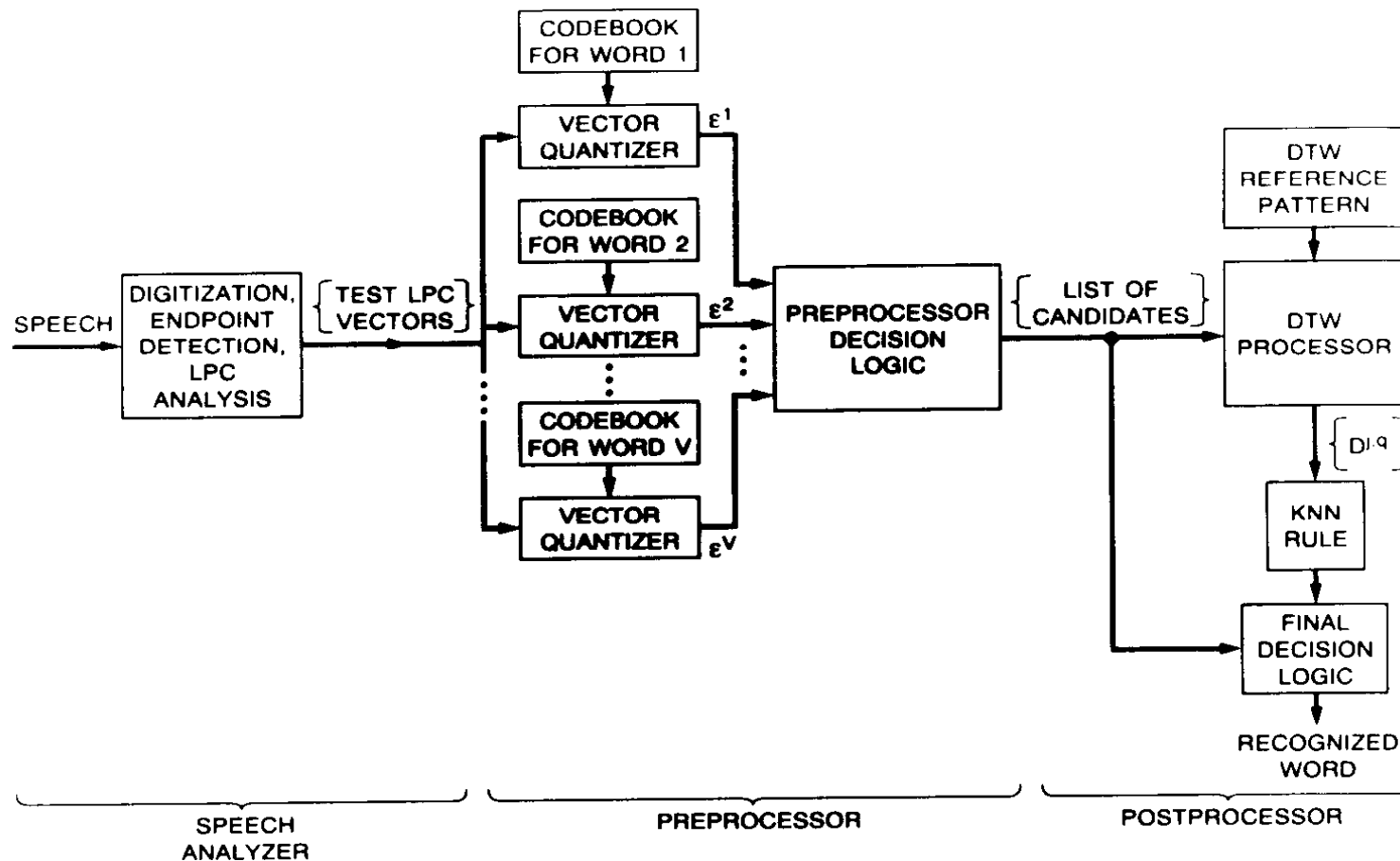


Figure 5.4 Block diagram of isolated word recognizer incorporating a word-based VQ preprocessor and a DTW-based postprocessor (after Pan et al. [4]).

ASR: Reconocimiento de Patrones Secuenciales (PR)

- Objetivo:
 - *Reconocer la secuencia de palabras (u otras unidades) partiendo de la forma de onda de la señal sonora.*

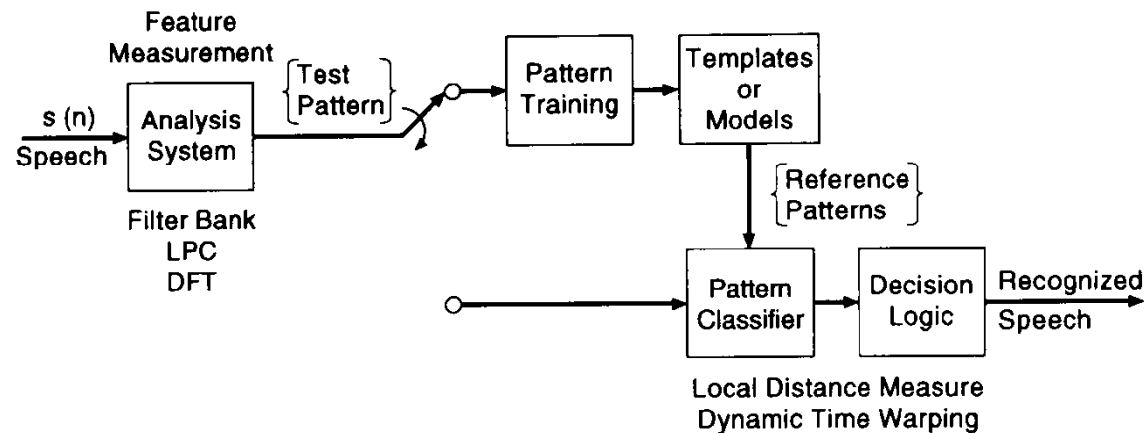


Figure 2.37 Block diagram of pattern-recognition speech recognizer.

- Dos fases esenciales de funcionamiento ASR:
 - Entrenamiento (aprendizaje)
 - Reconocimiento (puesta a prueba)

ASR: Ventajas de PR

- Sencillez:
 - Fácil de comprender
 - Fundamentos formales/matemáticos sólidos
- Robustez:
 - Invariante a cambios de vocabulario, usuarios, juegos de rasgos, algoritmos de comparación, reglas de decisión.
 - Metamodelo vs. Modelo
- Alto rendimiento
 - Demostrado en la práctica
 - Margen para mejora claro

ASR: Medidas de distancia

- Rasgos
 - Bancos de filtros
 - Codificación de Predicción lineal
 - Análisis cepstral
 - Cepstrum en escala Mel
 - Predicción perceptual lineal
- Métricas
 - Distancia Euclídea, p-Norma
 - Mahalanobis
 - Distancia Itakura-Saito

ASR: Predictor Lineal (LPC)

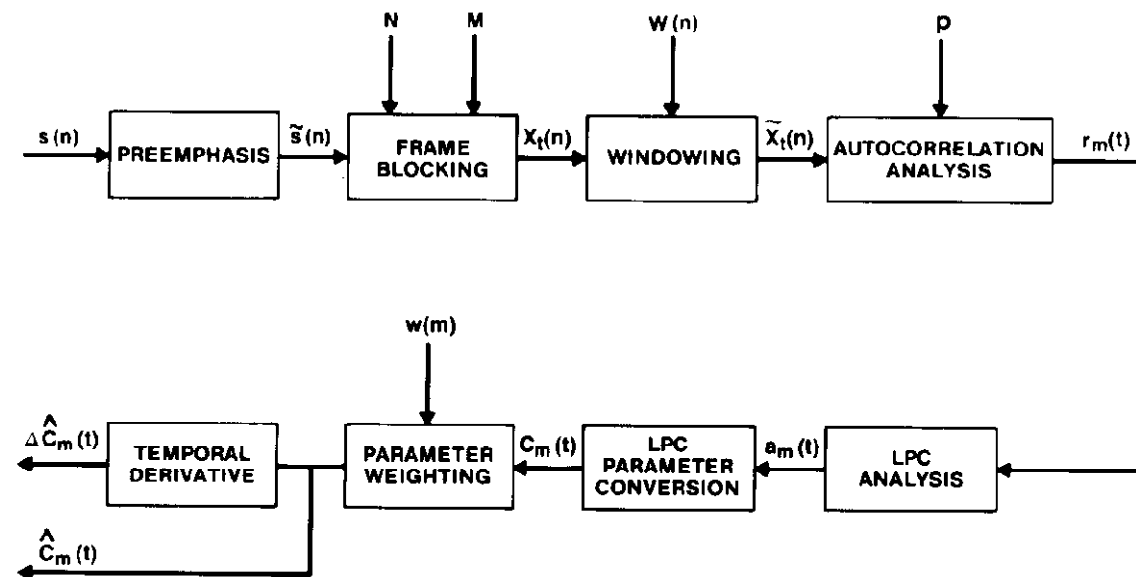


Figure 3.37 Block diagram of LPC processor for speech recognition.

ASR: Modelos perceptuales (no lineales)

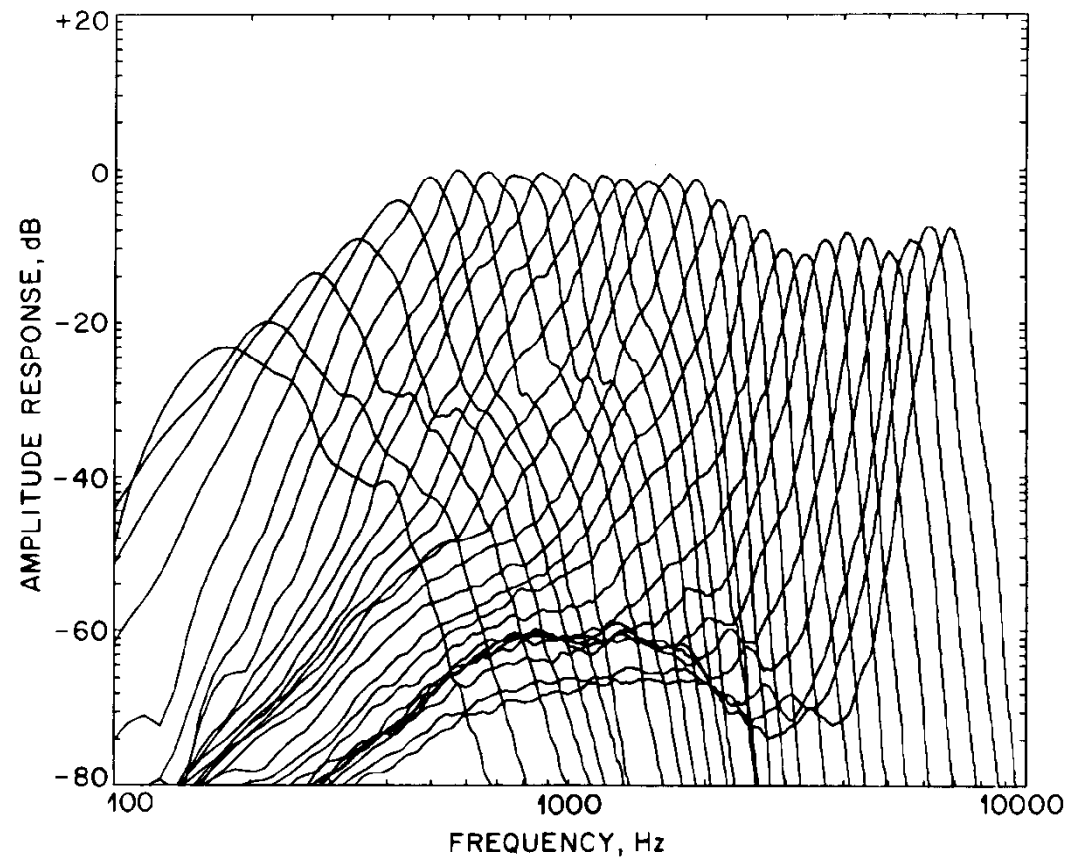


Figure 3.50 Frequency response curves of a cat's basilar membrane (after Ghitza [13]).

ASR: Alineamiento no lineal

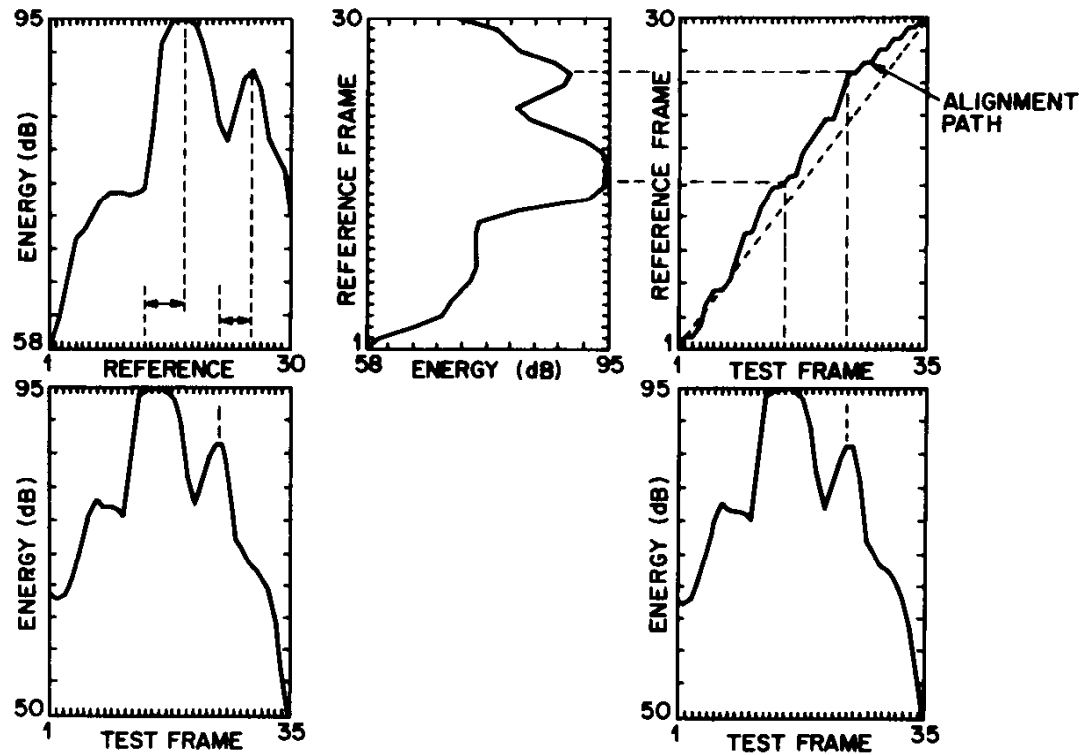


Figure 4.51 Example illustrating the need for nonlinear time alignment of two versions of a spoken word.

ASR: Limitaciones PR

- Rendimiento = f (Talla Entrenamiento)
- Patrones de referencia = f (Medio, Entorno)
- Carga computacional $\sim N^\circ$ patrones
- Generalidad vs. Especificidad
 - Independencia del nivel descriptivo
 - Dificultad de incorporar conocimiento específico.

ASR: Integración de conocimiento

- Acústico
 - Evidencias espectrales: presencia/ausencia de rasgos
- Léxico
 - Combinación de evidencias acústicas en palabras.
- MorfoSintáctico
 - Combinación de palabras en frases ‘correctas’
- Semántico
 - Combinación condicional correcta de significado:
 - Significado: representación simbólica estructurada
 - Datos + Relaciones (Hechos y Reglas)
- Pragmático
 - Normas de uso contextual (histórico)
 - Desambigüación

ASR: Esquema top-down

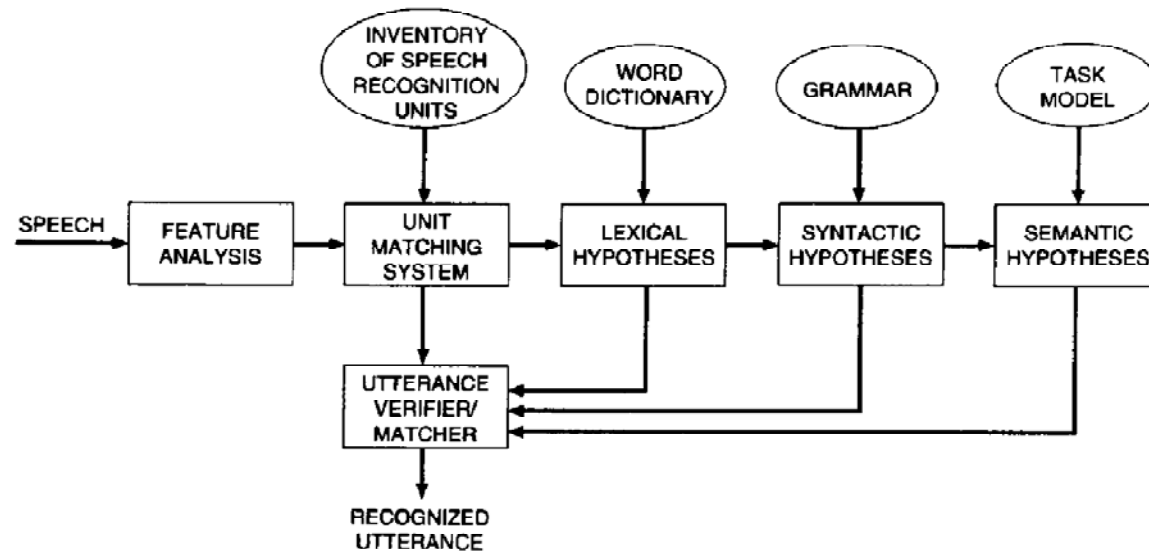


Figure 2.40 A top-down approach to knowledge integration for speech recognition.

- Generación de hipótesis.
- Validación con datos (hechos) encontrados.

ASR: Esquema bottom-up

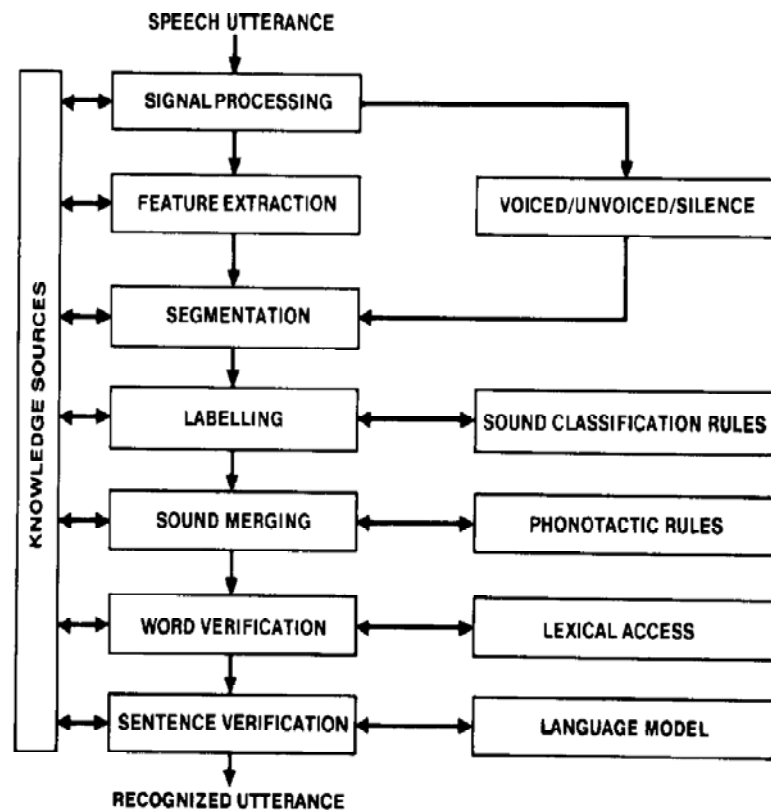


Figure 2.39 A bottom-up approach to knowledge integration for speech recognition.

- Constructivista
- Síntesis controlada
- Restricciones
- Menor acoplamiento

ASR: Esquema de pizarra

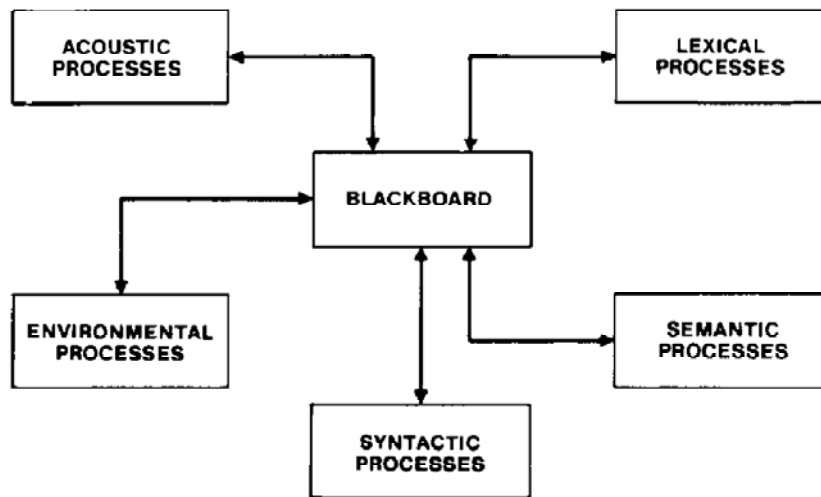


Figure 2.41 A blackboard approach to knowledge integration for speech recognition (after Lesser et al. [11]).

- Dirigido por datos
- KS independientes
- Paradigma individual:
 - Hipótesis/prueba
- Propagación asíncrona:
 - Costes
 - Méritos globales

ASR: Diagrama conceptual (ANNs?)

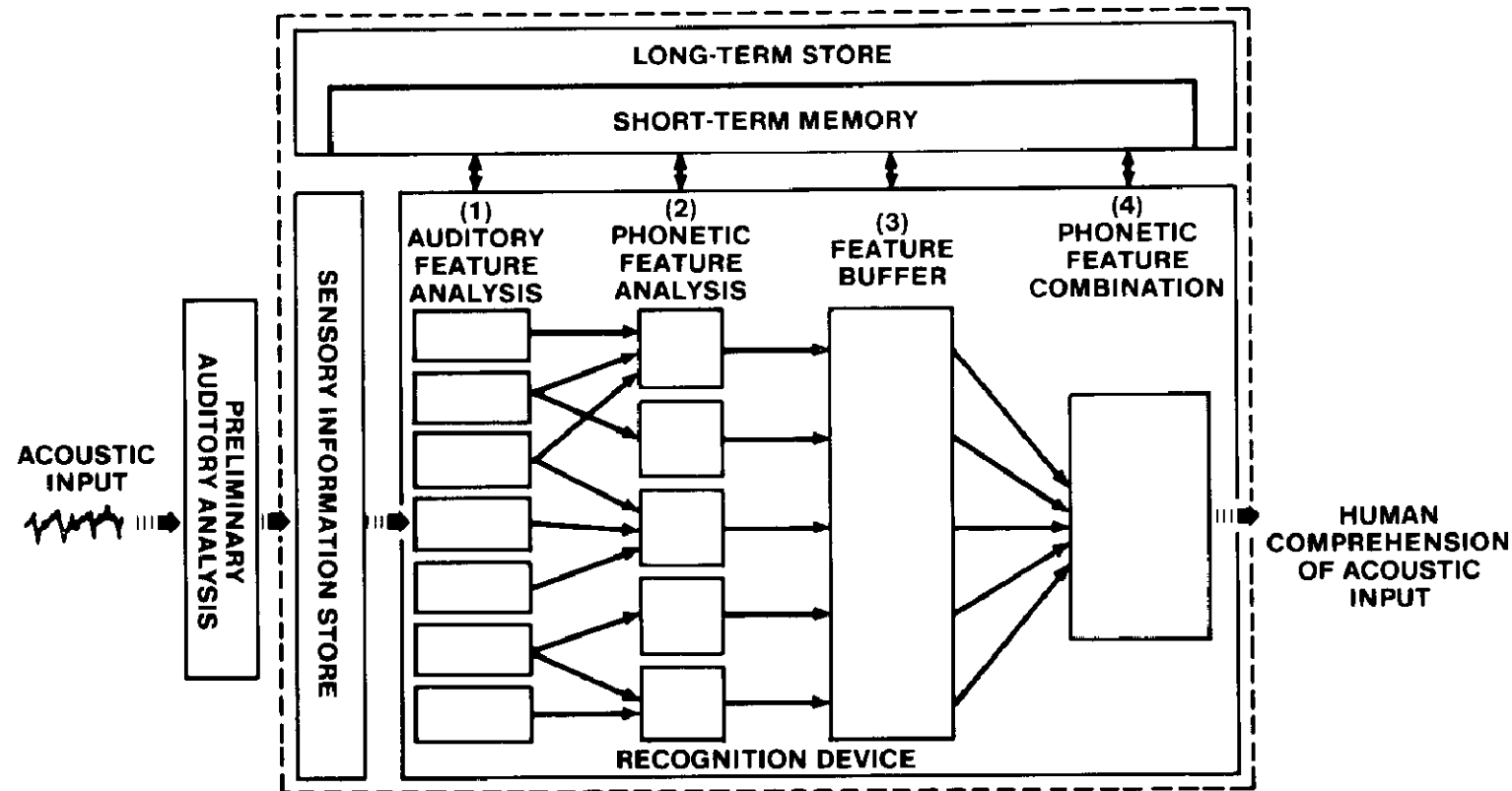


Figure 2.42 Conceptual block diagram of a human speech understanding system.

ASR: Modelos estocásticos

- PR (plantillas):
 - Caracterización estadística implícita limitada
- Suposición (variabilidad):
 - Señal hablada $\equiv$ Proceso aleatorio paramétrico
 - Parámetros = $f(\text{Observación})$
- Soporte:
 - Baum et al (60's)
 - Baker (CMU) y Jelinek (IBM) (70's)
 - Rabiner et al. (Bell) (80's), Young (Camb.) (90's)

HMM: Intro (¿Qué nos oculta Markov?)

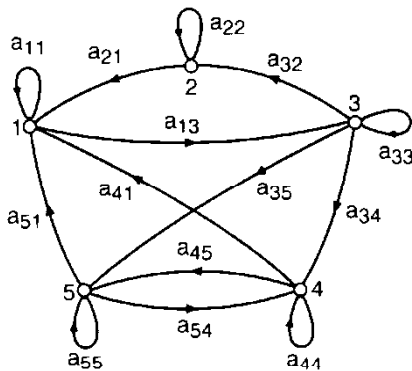


Figure 6.1 A Markov chain with five states (labeled 1 to 5) with selected state transitions.

Interpretación(Estado) = Salida

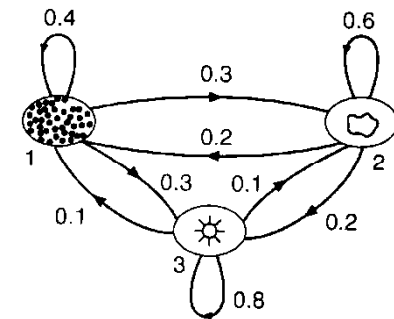
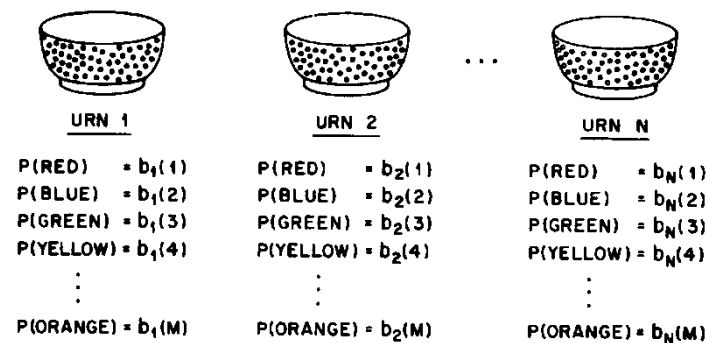


Figure 6.2 Markov model of the weather.

$$\Pr(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \Pr(X_{n+1} = x | X_n = x_n).$$



$O = \{\text{GREEN, GREEN, BLUE, RED, YELLOW, RED, } \dots, \text{BLUE}\}$

Figure 6.4 An N -state urn-and-ball model illustrating the general case of a discrete symbol HMM.

¿Qué oculta HMM?

Estado $\langle 1:N \rangle$ Salida

HMM: Componentes

- N: Conjunto de estados (finito)
- M: Alfabeto de salida (finito +)
- A: Distribución de probabilidades de transición
- B: Distribución de probabilidad de emisión
- π : Distribución de probabilidad de estado inicial
- Modelo:

$$\mu = (A, B, \pi)$$

HMM: Generación observaciones

1. Elegir estado inicial $q[1]$ usando π
2. Hacer $t=1$
3. Elegir $\mathbf{o}[t] = \mathbf{v}[k]$ usando B
4. Cambiar a $\mathbf{q}[t+1]$ usando A
5. Hacer $t=t+1$
6. Ir a 3 si $t < T$
7. Terminar

HMM: Problemas

- Problema 1: **Evaluación**

- IN:
 - Secuencia de observación $\mathbf{O} = (\mathbf{o}[1]\mathbf{o}[2] \dots \mathbf{o}[n])$
 - Modelo $\mu = (A, B, \pi)$
- OUT:
 - Probabilidad $P(\mathbf{O} \mid \lambda)$

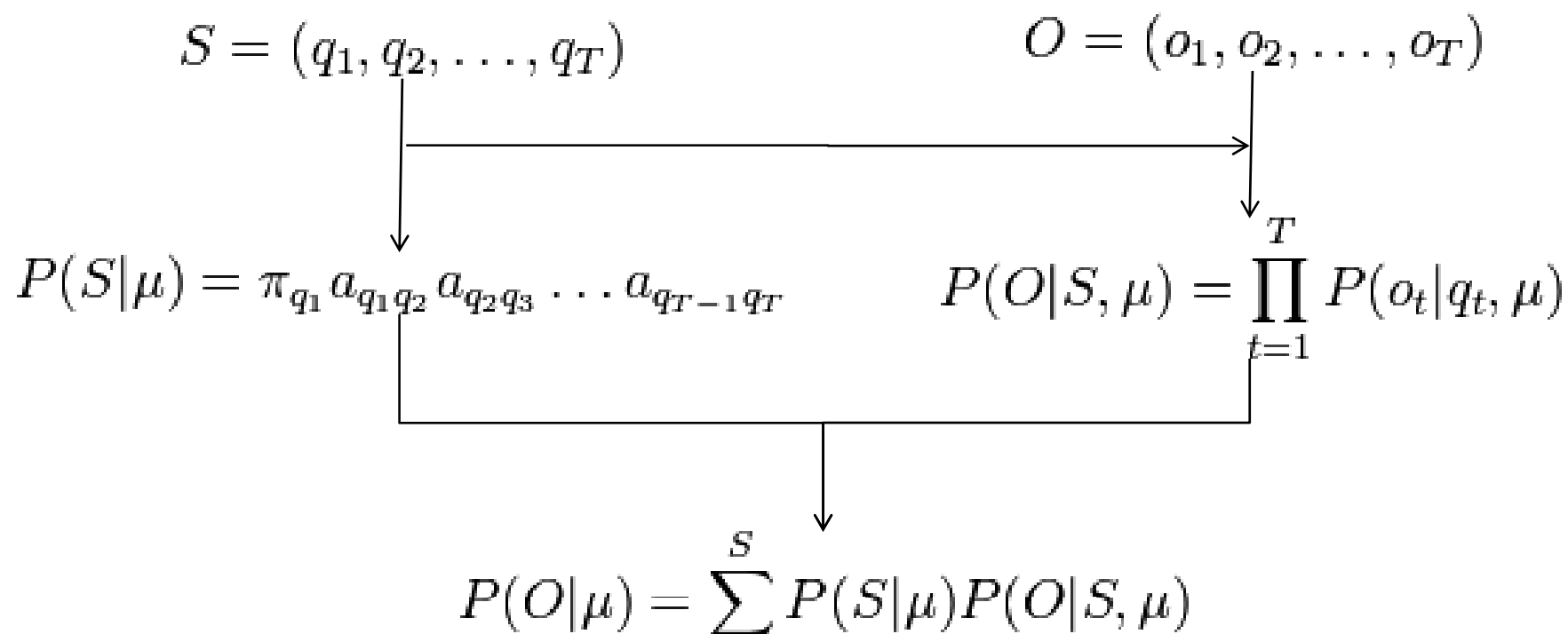
- Problema 2: **Camino óptimo**

- IN:
 - Secuencia de observación $\mathbf{O} = (\mathbf{o}[1]\mathbf{o}[2] \dots \mathbf{o}[n])$
 - Modelo $\mu = (A, B, \pi)$
- OUT:
 - Secuencia de estados $\mathbf{q} = (q[1]q[2] \dots q[T])$ 'óptima'.

- Problema 3: **Ajuste de modelo**

- IN:
 - Secuencia(s) de observación $\mathbf{O} = (\mathbf{o}[1]\mathbf{o}[2] \dots \mathbf{o}[n])$
- OUT:
 - Modelo $\mu = (A, B, \pi)$ que maximiza $P(\mathbf{O} \mid \lambda)$

HMM: Problema 1 (Scoring)



HMM: Solución Problema 1

Forward

$$\alpha_t(i) = P(o_1, o_2, \dots, o_t, q_t = i | \mu)$$

$$\alpha_1(i) = \pi_i b_i(o_1), \quad 1 \leq i \leq N$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(o_{t+1})$$
$$t = 1, 2, \dots, T-1 \quad 1 \leq j \leq N$$

$$P(O|\mu) = \sum_{i=1}^N \alpha_T(i)$$

Backward

$$\beta_t(i) = P(o_{t+1} o_{t+2}, \dots, o_T | q_t = i, \mu)$$

$$\beta_T(i) = 1, \quad 1 \leq i \leq N$$

$$\beta_t(i) = \sum_{j=1}^N a_{ij} \beta_{t+1}(j) b_j(o_{t+1})$$
$$t = T-1, T-2, \dots, 1 \quad 1 \leq i \leq N$$

$$P(O|\mu) = \sum_{i=1}^N \beta_1(i) \pi_i b_i(o_1)$$

HMM: Problema 2 (Optimum Path)

$$O = (o_1, o_2, \dots, o_T)$$

!! Secuencia Estados Oculta !!

Secuencia Esperada = Más probable

$$S = (q_1, q_2, \dots, q_T)$$

Maximiza

$$P(O|S, \mu) = \prod_{t=1}^T P(o_t|q_t, \mu)$$

HMM: Solución Problema 2 (Viterbi)

$$\delta_1 i = \pi_i b_i(o_1) \quad 1 \leq i \leq N$$

Inicio

$$\delta_{t+1}(j) = \left[\max_{1 \leq i \leq N} \delta_t(i) a_{ij} \right] b_j(o_{t+1})$$
$$\varphi_{t+1}(j) = \arg \max_{1 \leq i \leq N} \delta_t(i) a_{ij} \quad t = 1, 2, \dots, T-1 \quad 1 \leq j \leq N$$

Iteración

$$q_T^* = \arg \max_{1 \leq i \leq N} \delta_T(i)$$

Terminación

$$q_t^* = \varphi_{t+1}(q_{t+1}^*) \quad t = T-1, T-2, \dots, 1$$

Secuencia (inversa)

HMM: Problema 3 (Model estimation)

- IN:
 - Observación O (o conjunto de ellas en secuencia)
- OUT:
 - Modelo $\mu = (A, B, \pi)$
 - Maximiza $P(O \mid \mu)$
- Procedimiento:
 - Reestimación Baum-Welch (ver paper)

HMM: Solución P. 3 (Baum-Welch)

$$O = (o_1, o_2, \dots, o_T)$$

$$\xi_t(i, j) = \frac{P(q_t = i, q_{t+1} = j, O | \mu)}{P(O | \mu)} = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O | \mu)} \quad \xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{k=1}^N \sum_{l=1}^N \alpha_t(k) a_{kl} b_l(o_{t+1}) \beta_{t+1}(l)}$$

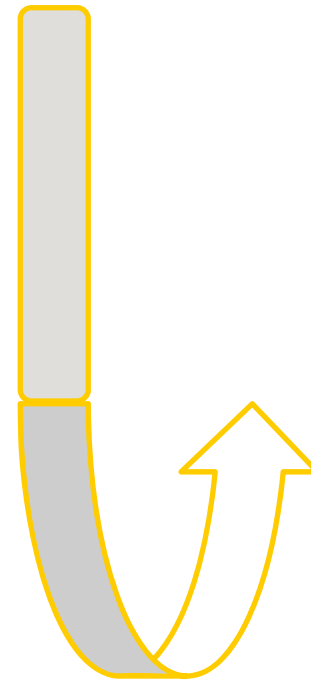
$$\gamma_t(i) = \sum_{j=1}^N \xi_t(i, j)$$

$$t = 1 \quad \bar{\pi}_i = \gamma_1(i) \quad 1 \leq i \leq N$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad 1 \leq i \leq N \quad 1 \leq j \leq N$$

**Convergencia
Asegurada**

$$\bar{b}_j(o_k) = \frac{\sum_{t=1: o_t=o_k}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)} \quad 1 \leq j \leq N \quad 1 \leq k \leq N$$



HMM: Variantes y Extensiones

- Transiciones:
 - Ergódico (poco empleado)
 - Left-to-Rigth (Bakis)
- Densidad continua
- Duración explícita de estados
- Criterio de optimalidad:
 - ML (Baum-Welch)
 - Cada clase (modelo) por separado: poca discriminación
 - MMI (métodos de Gradiente)
 - Todas las clases a la vez.
 - Minimizar la incertidumbre condicional de una clase (dado el resto)

HMM: Herramientas

- HMM Toolbox para Matlab
 - Útil para jugar con problemas de monedas/urnas
- Hidden Markov Model Toolkit (HTK)
- Sphinx
 - Muy usados en ASR
- Incorporación API programas propios:
 - GHMM Library
 - Biblioteca en C:
 - <http://www.kanungo.com/software/software.html>

Referencias

- Libro de Rabiner y Juang.
 - “Fundamentals Of Speech Recognition”. Prentice-Hall 1993.
- Artículos proporcionados en el entorno.
 - Tutorial de HMM (con corrección de errores)
- Wikipedia (para empezar) y otros web:
 - [http://es.wikipedia.org/wiki/Modelo oculto de M%C3%A1rkov](http://es.wikipedia.org/wiki/Modelo_oculto_de_M%C3%A1rkov)
 - [http://es.wikipedia.org/wiki/Algoritmo de avance-retroceso](http://es.wikipedia.org/wiki/Algoritmo_de_avance-retroceso)
 - ...
 - <http://jedlik.phy.bme.hu/~gerjanos/HMM/node2.html>
- Pon las tuyas ... (para el cuaderno)

Actividades

- ‘Para el cuaderno’ en esta presentación
- Elaborar un resumen de HMMs y solución de los tres problemas base empleando artículo Rabiner.
- Usar HTK: Realizar el tutorial.