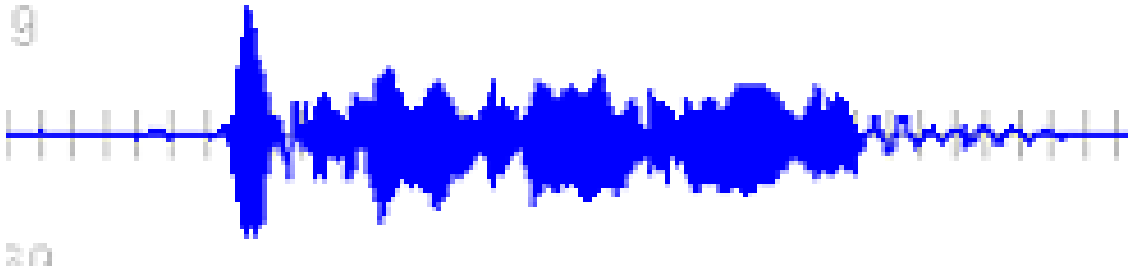
Minicurso en Reconocimiento Automatico del Habla

Jorge Luis Guevara Díaz

Reconocimiento automático del habla

- Basicamente consiste en:

Convertir habla



a texto automáticamente (computadora)

"hola como estan?"

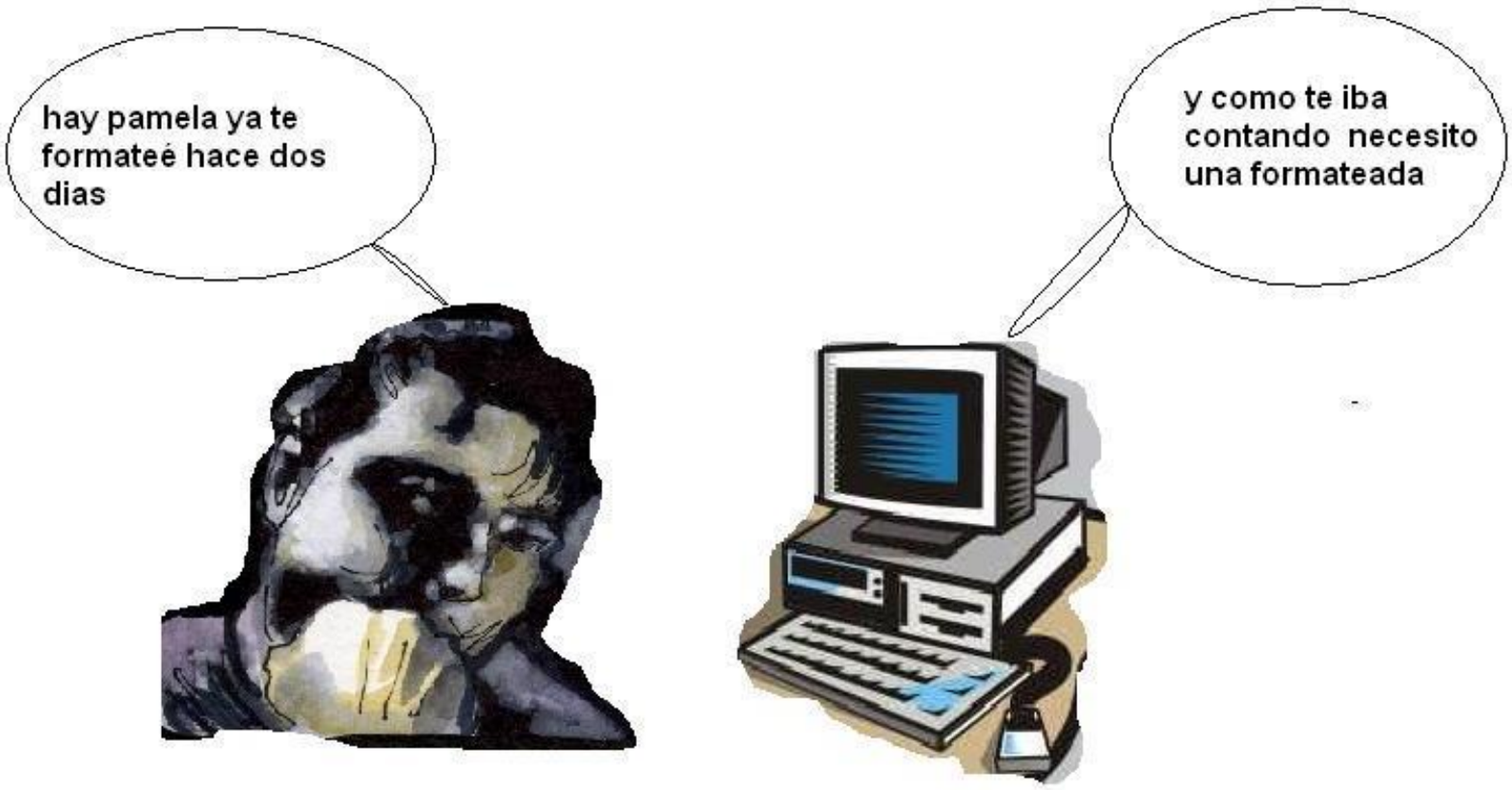
Justificacion

- El habla es natural en los humanos

- No requiere entrenamiento sofisticado previo
- Puede ser usada de manera paralela con otras actividades

Justificacion

- seria natural usarla para comunicarnos con las máquinas de manera rápida.



Importancia

- Acceso remoto a computadoras mediante telefono

- Procesamiento indexacion y comprension de audios con habla humana, para posterior transcripción automática.

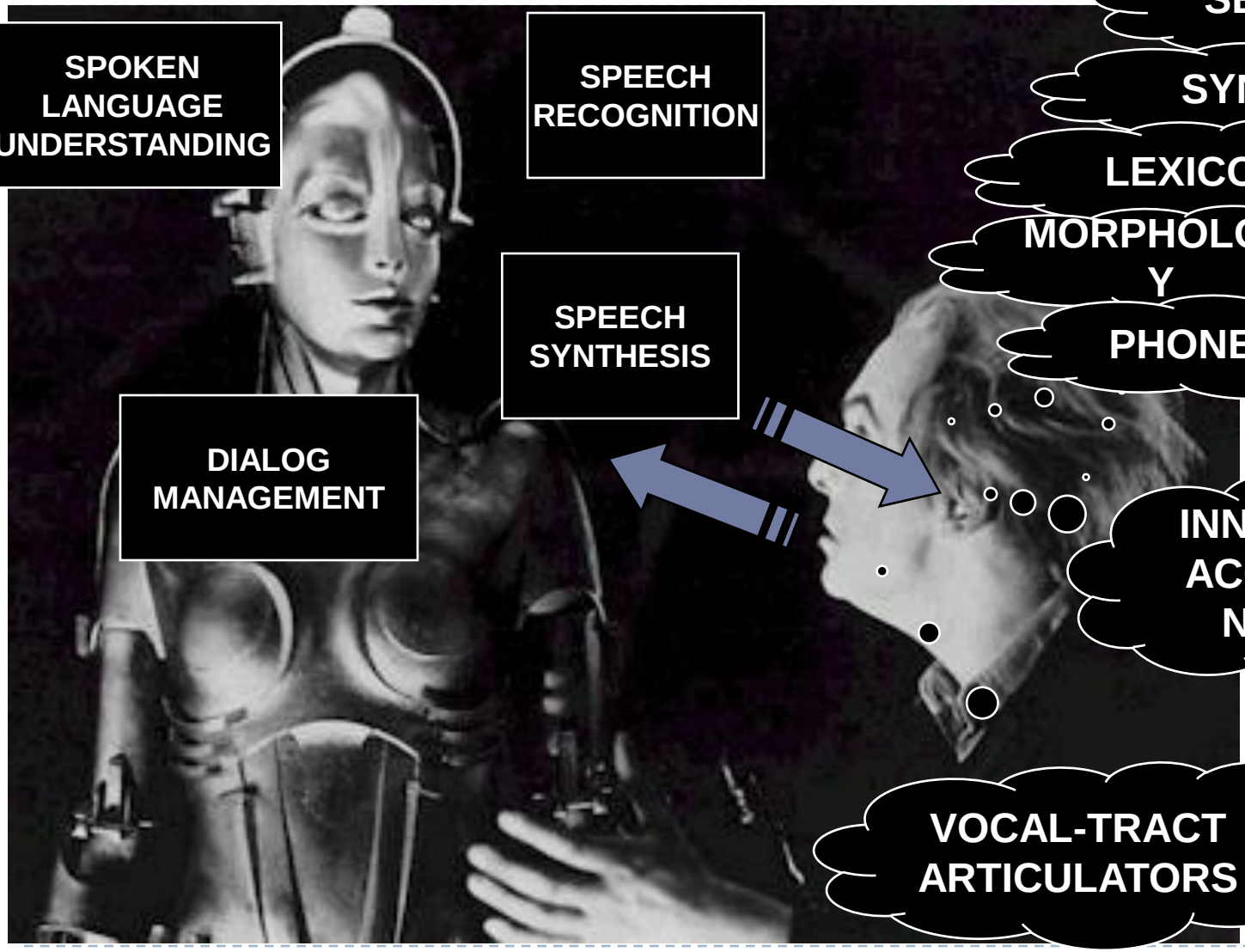
Importancia

- Parte de una interfaz humano-máquina que utilice el habla como medio de comunicación

Interfaz humano computador mediante habla

- Reconocimiento automático del Habla (speech recognition)
- Entendimiento del Lenguaje Natural (natural spoken understanding)
- Sintesis de habla (speech synthesis)

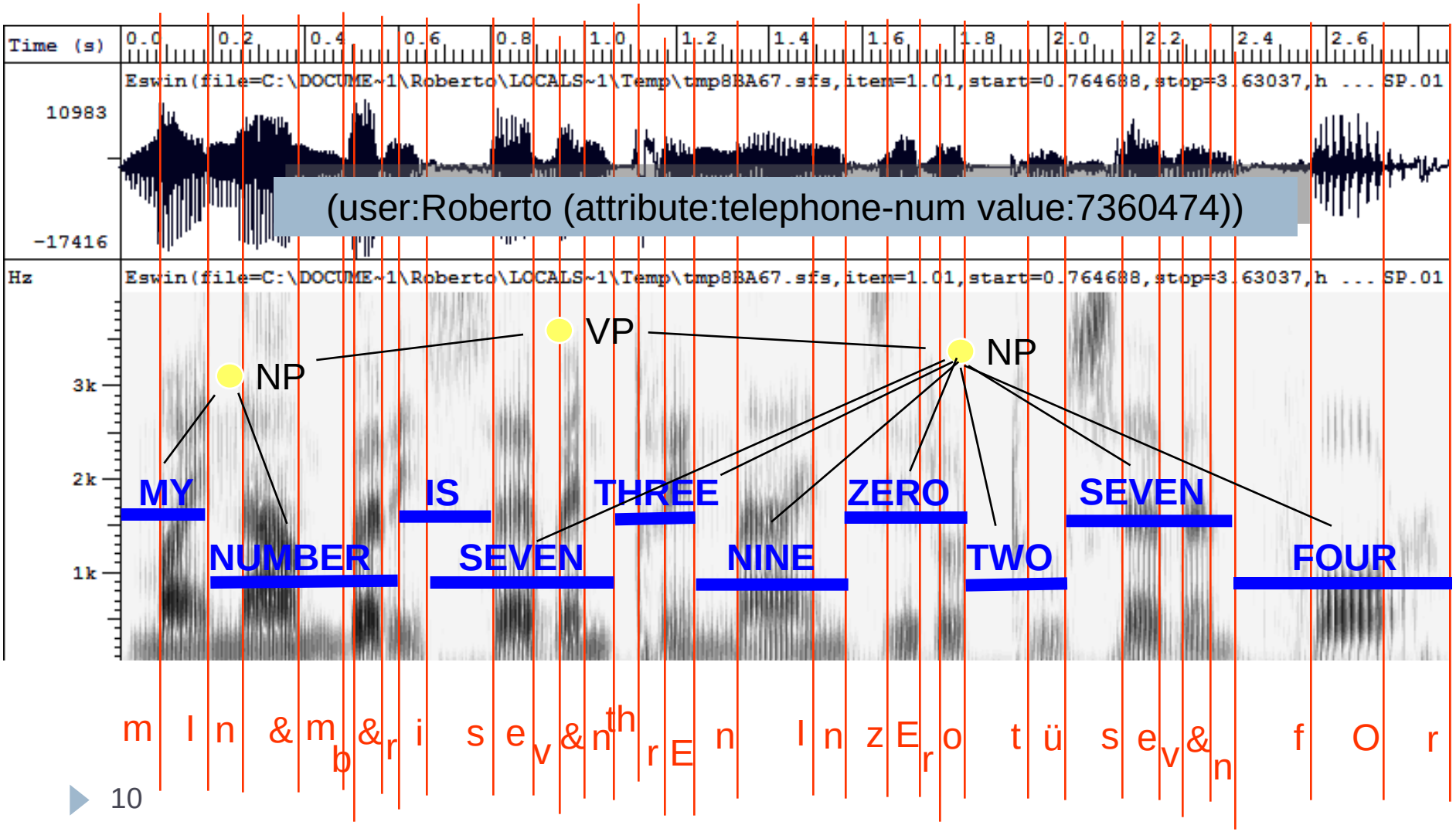
Cadena del Habla



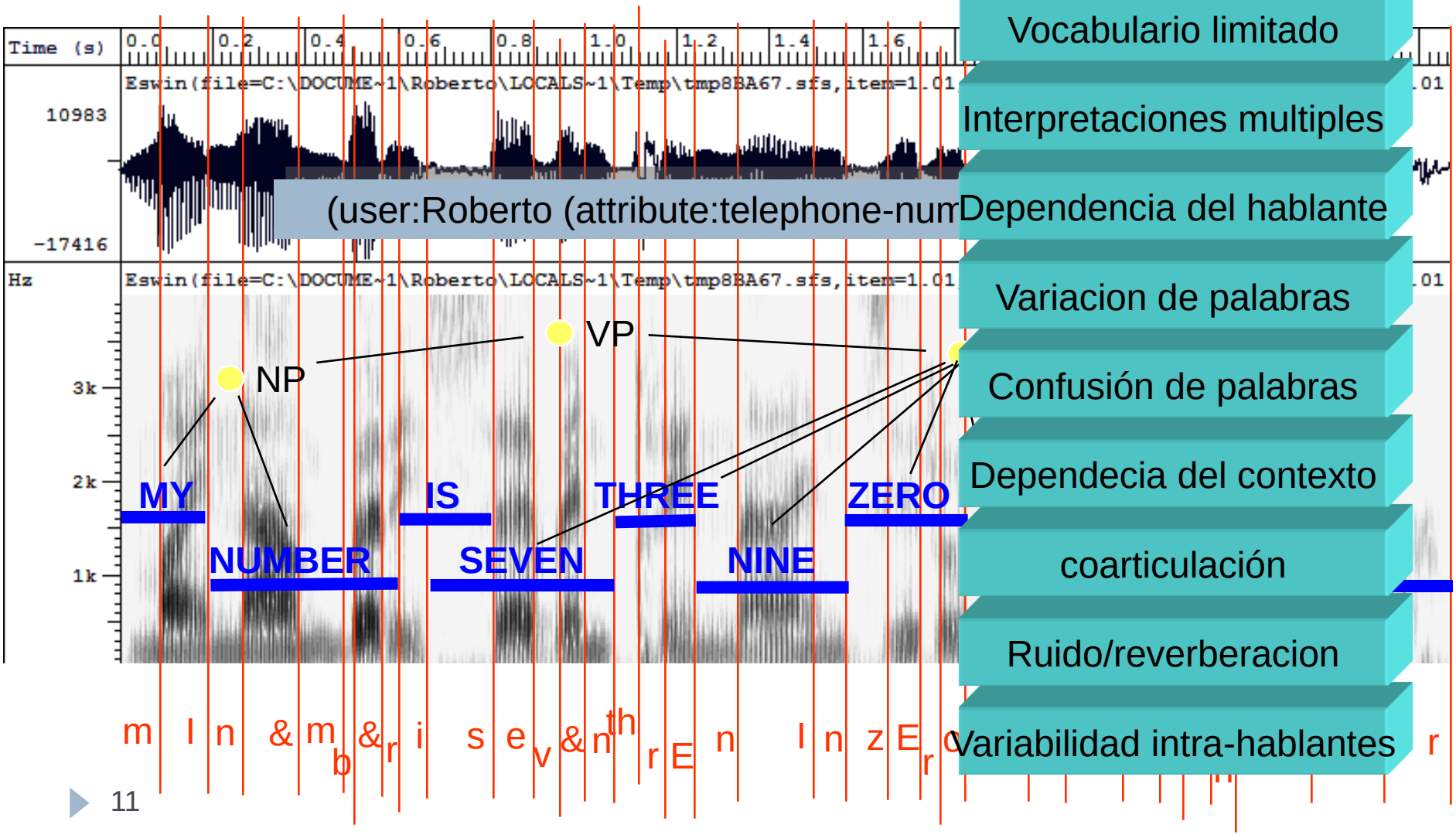
RAH es un curso multidiciplinario

- Procesamiento digital de señales
- Lingüística
- Procesamiento del lenguaje natural
- Reconocimiento de patrones,
- Inteligencia artificial
- etc

Porque el RAH es tan complicado



Porque el RAH es tan complicado

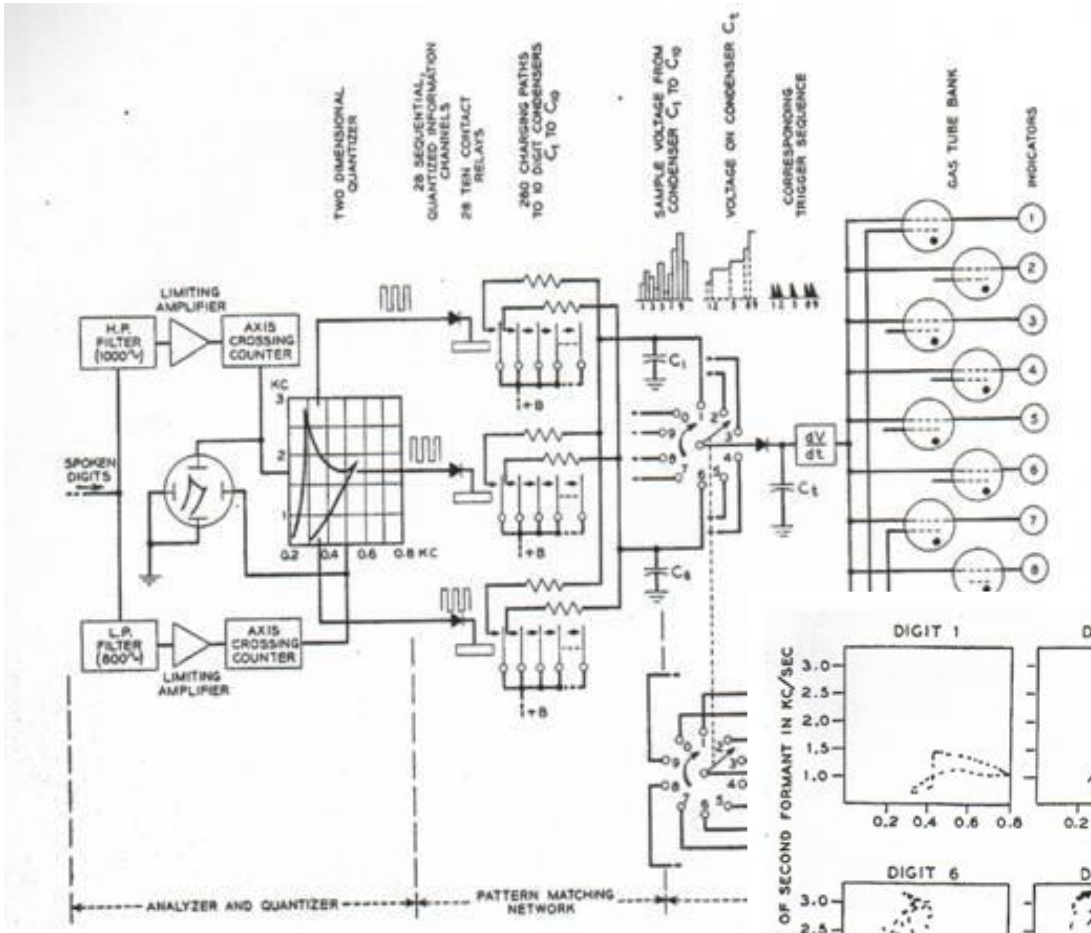


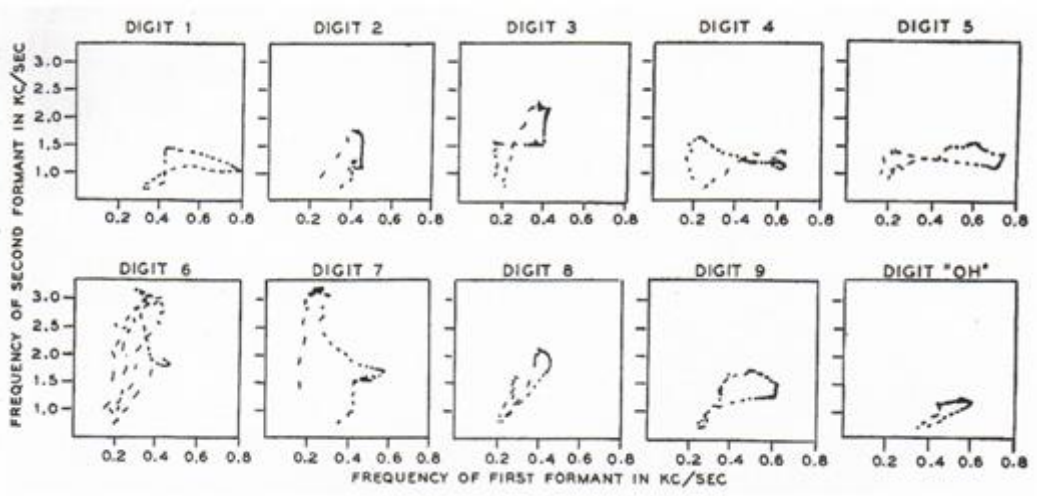
Breve Historia

El inicio

- Radio Rex (1920)
  - Reconocedor independiente del hablante de una sola palabra "Rex"
  - Funcionaba si suficiente energía era detectada a los 500 Hz proveniente de la letra "e"

- 1952 – Automatic Digit Recognition (AUDREY)
  - Davis, Biddulph, Balashek (Bell Laboratories)





1960's – Procesamiento del habla y computadoras digitales

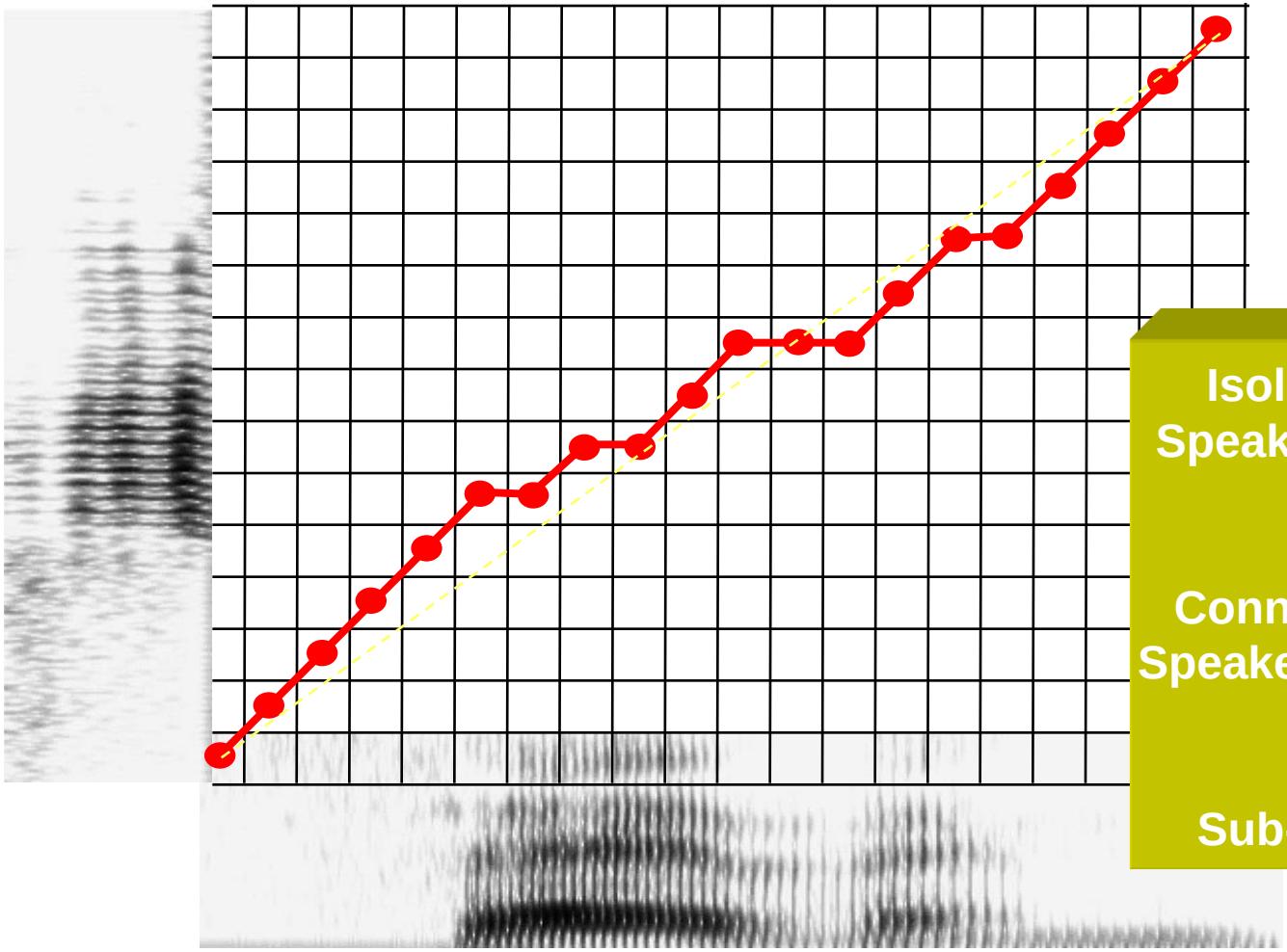
- Convertirores AD/DA y computadoras digitales aparecen en los laboratorios

James Flanagan
Bell Laboratories

Años 1920 a 1960

- Metodos ad hoc
  - Procesamiento de la señal demasiado simple, extracción de carácterísticas muy básica
    - Se detectaba la energía a diferentes bandos de frecuencia, o se buscaban frecuencias dominantes
  - Muchas ideas son introducidas, pero no se usan todas juntas
    - Entrenamiento estadístico, modelado del lenguajes
  - Vocabulario pequeño
    - Digitos, si/no, vocales
  - No se testeaban los algoritmos con muchos hablantes (<10)

1970's – Dynamic Time Warping



T.K. Vyntsyuk (1968)
H. Sakoe,
S. Chiba (1970)

Isolated Words
Speaker Dependent

⇩

Connected Words
Speaker Independent

⇩

Sub-Word Units

Punto de Quiebre

- Jhon Pierce, Bell Labs 1969

- "RAH para propositos generales esta muy lejano, RAH para propositos especiales es muy limitado, deberia ser adecuado preguntar por que estan trabajando en el campo y que es lo que quieren lograr"

- "Esas consideraciones hacen creer que un RAH general sea simplemente imposible a menos que tenga inteligencia y conocimiento de lenguaje comparado a los hablantes nativos del inglés"

1969 – Whither Speech Recognition?

General purpose speech recognition seems far away. Social-purpose speech recognition is severely limited. It would seem appropriate for people to ask themselves why they are working in the field and what they can expect to accomplish…

It would be too simple to say that work in speech recognition is carried out simply because one can get money for it. That is a necessary but not sufficient condition. We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn't attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%. To sell suckers, one uses deceit and offers glamour…

Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. The typical recognizer gets it into his head that he can solve "the problem." The basis for this is either individual inspiration (the "mad inventor" source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach).

The Journal of the Acoustical Society of America, June 1969

J. R. Pierce
Executive Director,
Bell Laboratories

Punto de Quiebre

- Bell Labs suspendio la investigación por muchos años

- Darpa 1971-1976
  - Fundan ASR research
  - Meta: integrar conocimiento del habla, lingüística e Inteligencia Artificial para hacer un punto de quiebre en RAH

  - Gran vocabulario: 1000 palabras, sintáxis artificial
  - Tiempo real

Punto de Quiebre

- Cuatro competidores
  - SDC (24%)
  - BBN's *HWIM (44%)*
  - *CMU's Hearsay II (74%)*
  - *CMU's HARPY (95% !)*

  - Tres usaban reglas derivadas a mano, puntajes basados en conocimiento del habla y del lenguaje

  - Harpy (CMU): integraba todas las fuentes del conocimiento en redes de estado finito que fueron entrenados estadísticamente

- Harpy ganó

Desarrollo de métodos estadísticos

- RAH es visto como:
  - Buscar la secuencia de palabra mas probable dada la señal de audio, dada cierta informacion de la distribución de probabilidad
  - Se entrena la distribucion de las probabilidades automáticamente de habla transcrita
  - Se usa un monto mínimo de conocimiento explícito del lenguaje y del habla

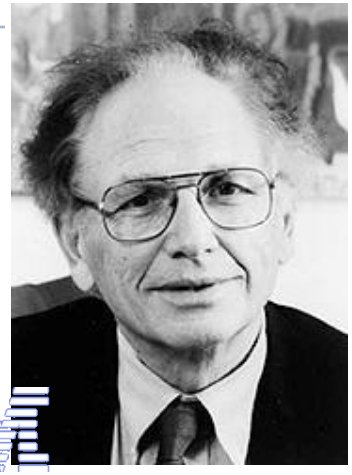
Nacimiento del moderno RAH 1970-1980

- Algoritmo EM
- Modelos n grams
- Mixturas gaussianas
- Modelos Ocultos de Markov
- Decodificacion de Viterbi, etc

- Primer sistema en tiempo real de dictado (IBM 1984)

1980s – enfoque estadístico

- Hidden Markov Models Leonard Baum en IDA, Princeton 1960s
- Fred Jelinek y Jim Baker, IBM T.J.Watson Research
- Bases de los modernos reconocedores

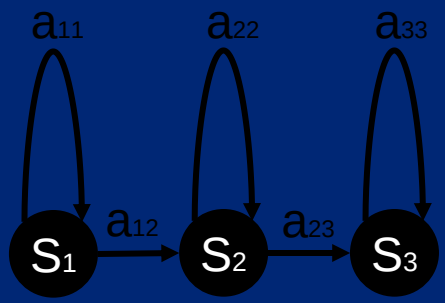


Fred Jelinek

Jim Baker

$$\hat{W} = \arg\max_{W} P(A \mid W) P(W)$$

HMMs acústicos

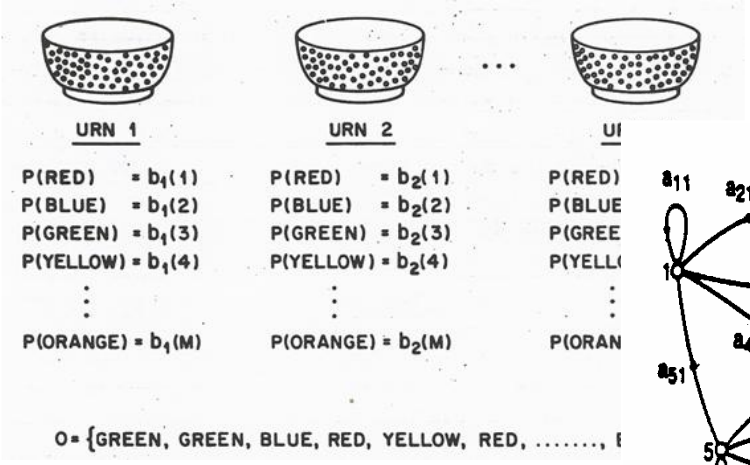


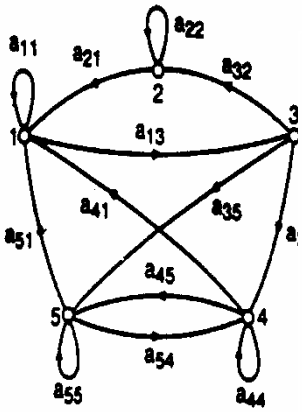
Word Tri-grams

$$P(w_t \mid w_{t-1}, w_{t-2})$$

1980-1990 – Enfoques Estadístico

- Lawrence Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceeding of the IEEE, Vol. 77, No. 2, February 1989.





Markov Assumption:

$$P[q_t = j | q_{t-1} = i, q_{t-2} = k, \ldots] = P[q_t = j | q_{t-1} = i]$$

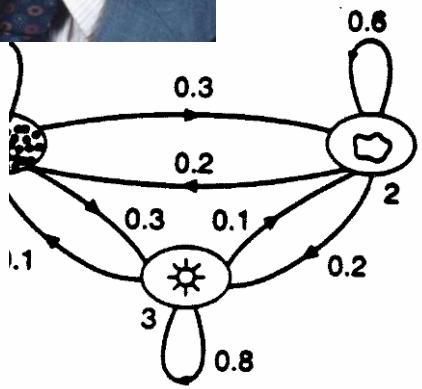
Set

$$a_{ij} = P[q_t = j | q_{t-1} = i] \quad 1 \le i, j \le N$$

Such that

$$a_{ij} \ge 0 \qquad \forall i, j$$

$$\sum_{j=1}^{N} a_{ij} = 1 \qquad \forall i$$



Estado del Arte

- Bajo condiciones de ruido
- Gran vocabulario
  - ~20,000-60,000 palabras a más…
- Independiente del hablante (vs. dependiente del hablante)
- Habla continua (vs palabras aisladas)
- Conversación multilenguaje
- Entrenamiento en 1971 (Darpa) 1 hora de habla. Hoy miles horas de habla

No todos los reconocedores son creados igual

- Dependiente del hablante vs independiente del hablante
  - Reconoce un solo hablante o muchos
- Pequeño vocabulario vs gran vocabulario
  - Diferente tamañoñ de vocabulario
- Dominio restringido vs dominio no restringido
  - Reserva de pasajes o dictado de email
- Aislado vs continuo
  - Pausado entre cada palabra o hablado naturalmente
- Leido vs espontaneo
  - Noticias de periodico o conversaciones de telefono

Reconocedores Comerciales

- 1995 Dragon, IBM, palabras aisladas, gran vocabulario, sistema de dictado
- 1997 Dragon, IBM, palabras continuas, gran vocabulario, sistema de dictado
- 1990 Existe software dependiente del hablante, de pequeño vocabulario disponible para uso mediante telefono
- 1990 Existe software dependiente del hablante, de gran vocabulario, de dominio limitado disponible para uso mediante telefono

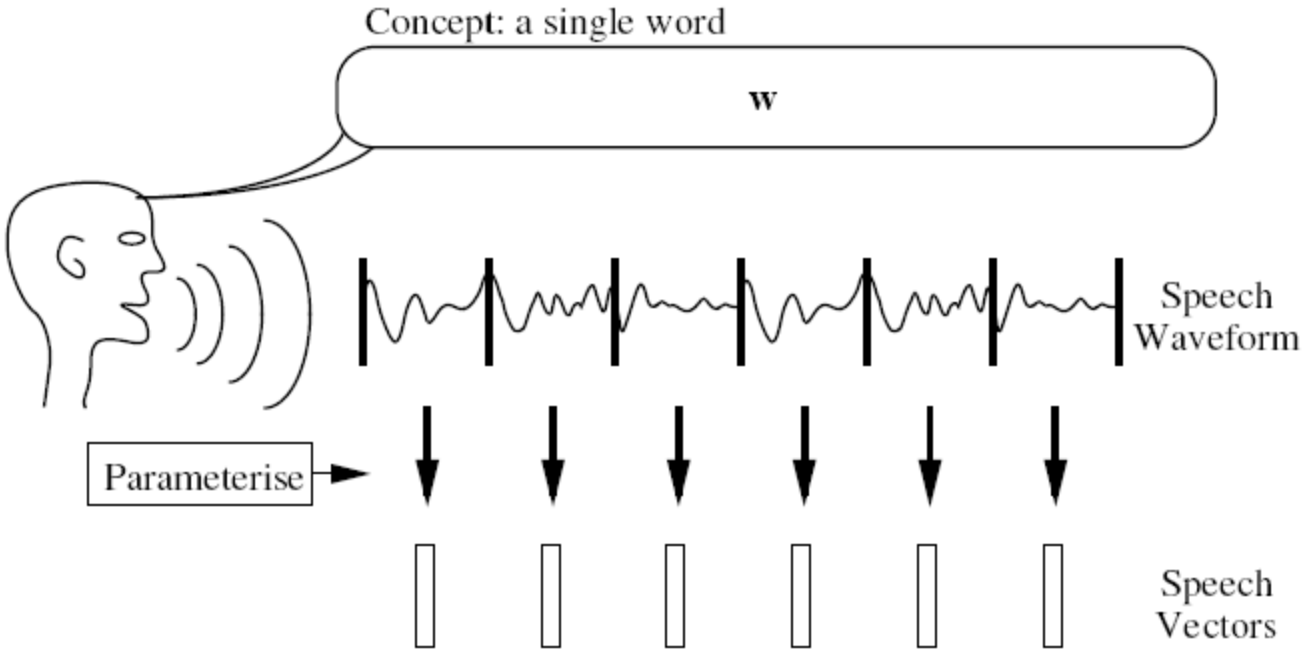
Algortimos para RAH

Formulacion

- Cada palabra hablada W es representada por una secuencia de vectores de habla ó observaciones O

$$O = o_1, o_2, \ldots, o_T$$

- O_t es el vector de habla en el tiempo t



Formulacion

- La entrada acustica O es una secuencia de observaciones acusticas individuales
  - $O = o_1, o_2, o_3, \ldots, o_t$
- Una sentencia W es una secuencia de palabras:
  - $W = w_1, w_2, w_3, \ldots, w_n$

Formulacion

- El RAH puede ser formulado como

$$\arg\max_i \{P(w_i|\boldsymbol{O})\}$$

- wi es la i'ava palabra del vocabulario, esta probabilidad no es computable directamente entonces usando la regla de Bayes

$$P(w_i|\boldsymbol{O}) = \frac{P(\boldsymbol{O}|w_i)P(w_i)}{P(\boldsymbol{O})}$$

Formulacion

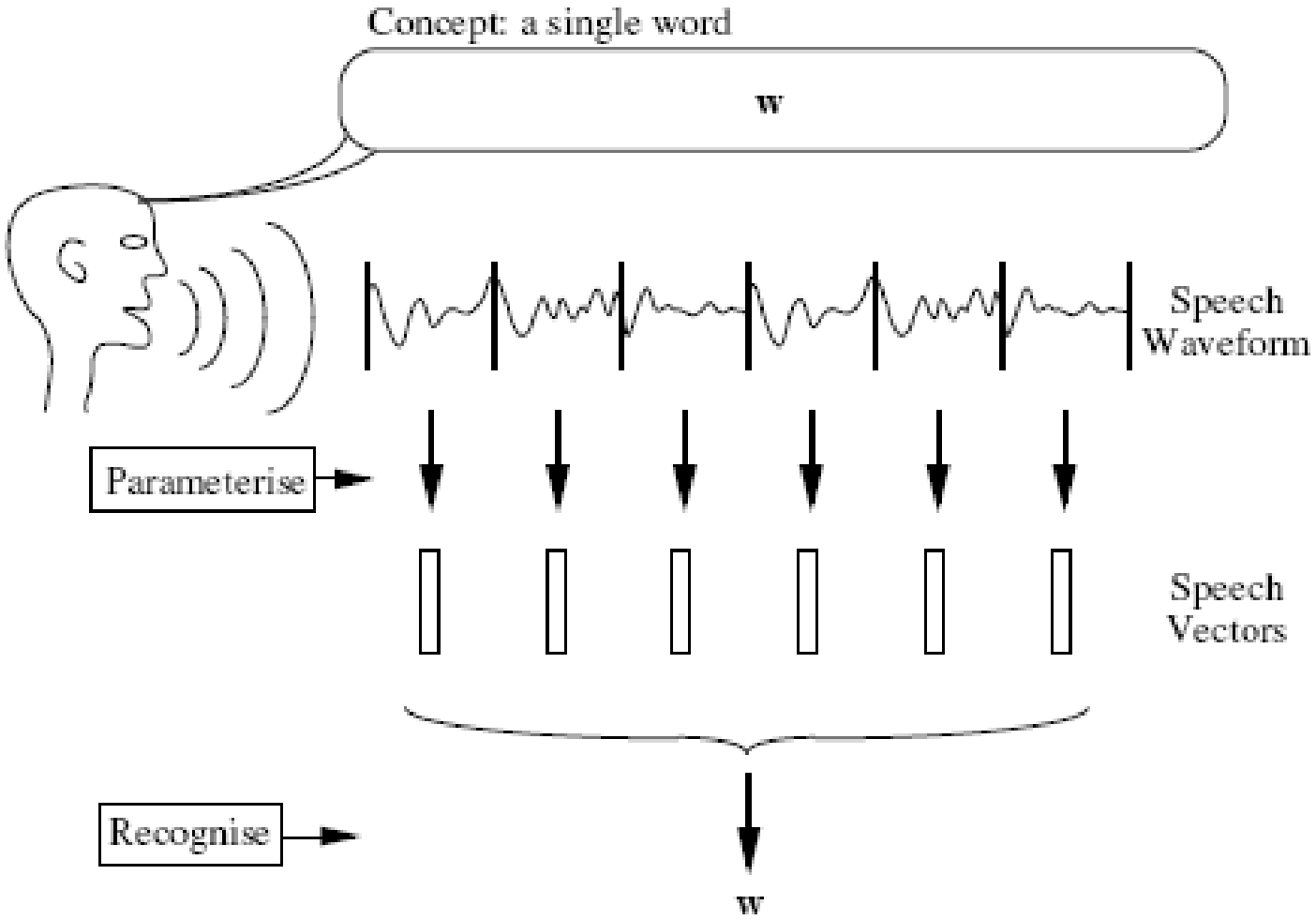
$$P(w_i|\boldsymbol{O}) = \frac{P(\boldsymbol{O}|w_i)P(w_i)}{P(\boldsymbol{O})}$$

- Dado un conjunto de probabilidades a priori

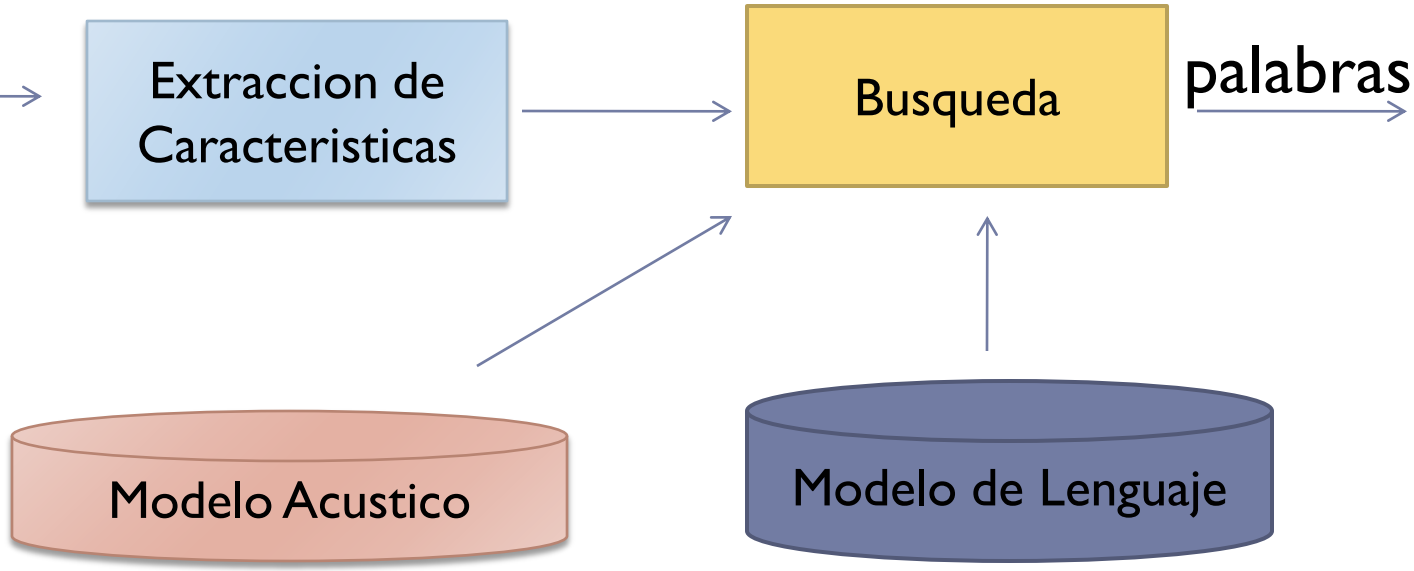
  $P(w_i)$,

  la mas probable palabra depende del likelihood

  $P(O| w_i)$


Concept: a single word
w
Speech Waveform
Parameterise
Speech Vectors
Recognise
w

Reconocimiento automatico del habla arquitectura



$$W = \arg\max_W P(O \mid W)\, P(W)$$

Extraccion de carácterísticas

- Objetivo: dada una señal acustica de entrada obtener una codificacion caracteristica asociada para dicha señal

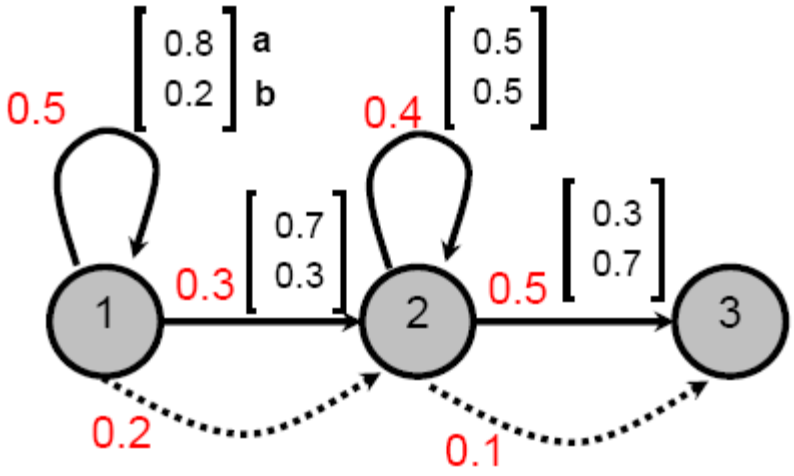
- Input: Señal de habla
- Output: O

- Principales algoritmos
  - MFCC
  - PLP
  - LPC
  - LPC-Cepstrum
  - Basados en wavelets

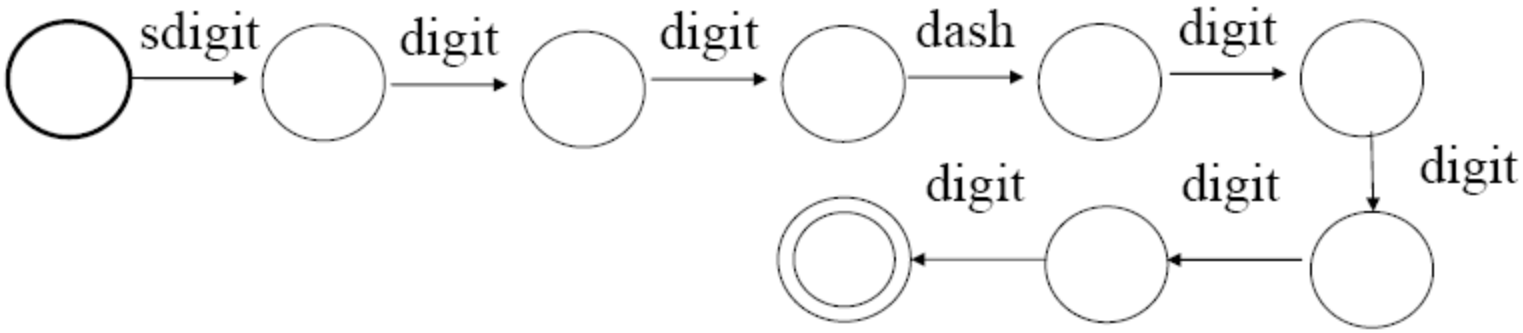
Modelo acustico

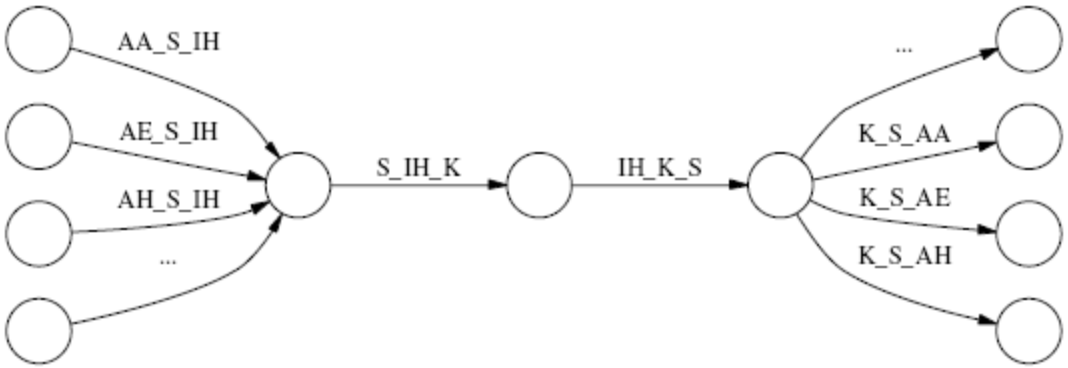
- Objetivo: construir modelos estadisticos de palabras que funcionen como generadores de las observaciones O

- Modelos ocultos de Markov

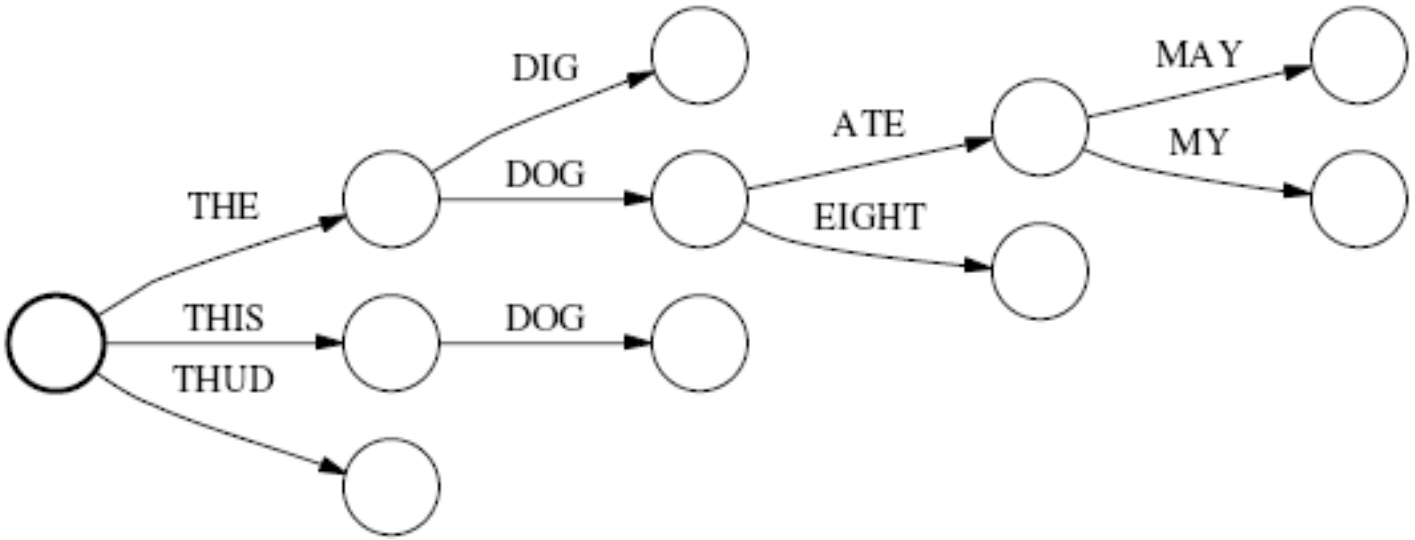


Modelo de Lenguaje





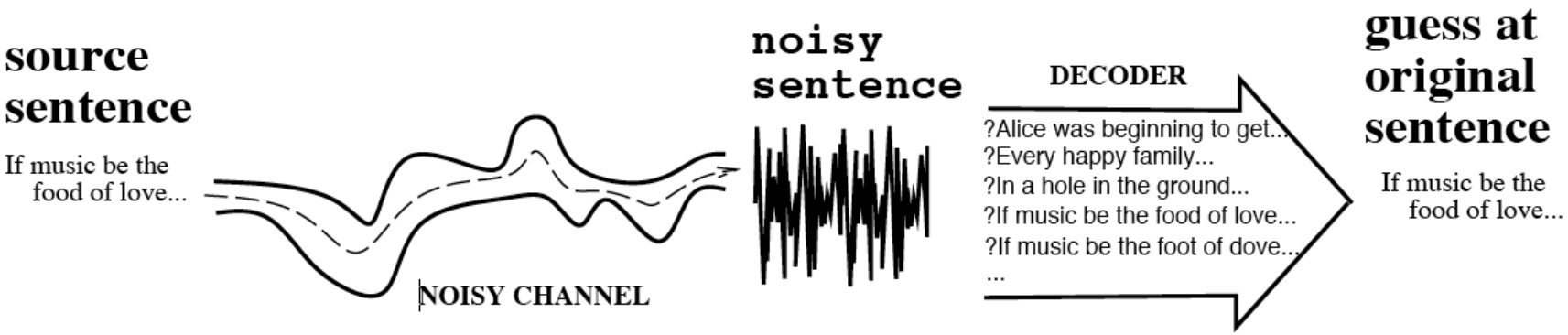
Búsqueda



Construyedo un sistema de RAH

- Construir un modelo estadistico del proceso habla-a-palabras
  - Coleccionar mucha habla y transcribir todas las palabras
  - Entrenar el modelo con el habla etiquetada
- Paradigma:
  - Aprendizaje de Máquinas Supervisado + Búsqueda
  - Modelo del Canal Ruidoso

El Modelo del Canal Ruidoso



- Buscar atravez del espacio de todas las posibles sentencias.
- Escoger aquella que es más probable dada la señal de habla entrante

Formulacion

- Cual es la sentencia mas probable de todas las sentencias de un lenguaje L, dado alguna entrada acustica O?

- La entrada acustica O es una secuencia de observaciones acusticas individuales
  - $O = o_1, o_2, o_3, \ldots, o_t$

- Una sentencia W es una secuencia de palabras:
  - $W = w_1, w_2, w_3, \ldots, w_n$

Componentes de un sistema RAH

- Cuerpo para entrenamiento y testing
- Modelo de pronunciación
- Modelo acustico
- Modelo de lenguaje
- Componente de extraccion de características
- Algoritmos eficientes de búsqueda en el espacio de hípótesis

Cuerpo para entrenamiento y testing

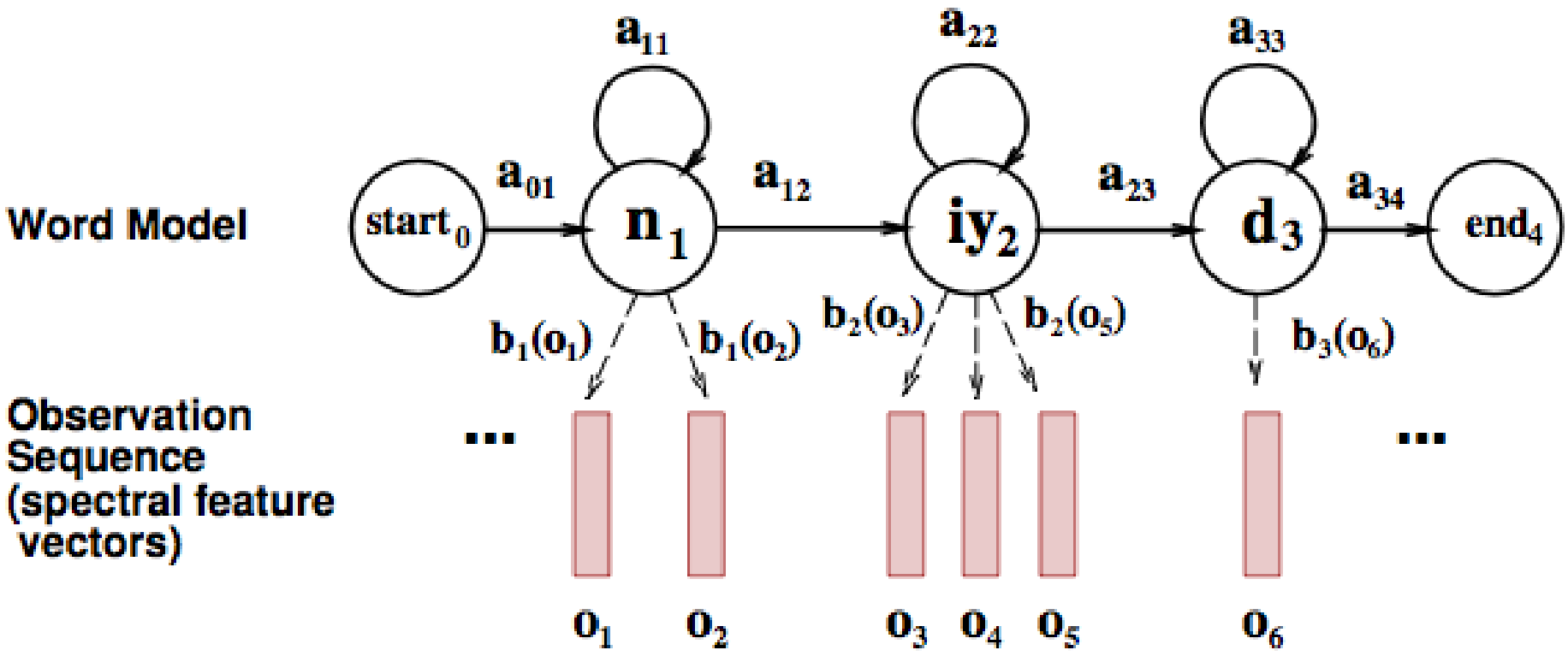
- Recolectar un cuerpo apropiado para la tarea de reconocimiento a mano
  - Poca cantidad de habla + transcripciones foneticas asociadas a los sonidos con símbolos (Modelo acustico)
  - Mucha cantidad de habla (>= 60 hrs) + transcripciones ortograficas asociadas a las palabras con sonidos (Modelo acústico)
  - Textos muy grandes para identificar probabilidades ngram o construir una gramática (Modelo del Lenguaje)

Modelo acústico

- Objetivo: Modelar el likelihood de sonidos dadas características espectrales, modelos de pronunciacion, y contexto a priori

- Usualmente representado como un Modelo Oculto de Markov
  - Estados representan los fonos u otras subunidades de palabras
  - Probabilidades asociadas a las transiciones en los estados: que tan probable es tener un sonido después de otro?
  - Likelihoods asociados a Observaciones/salida : Que tan probable es que un vector de caracteristicas espectrales sea observado en el fono del estado i, dado el fono en el estado i-1?

HMM Palabra


Word Model
Observation Sequence (spectral feature vectors)
start0
n1
iy2
d3
end4
a01
a11
a12
a22
a23
a33
a34
b1(o1)
b1(o2)
b2(o3)
b2(o5)
b3(o6)
o1
o2
o3
o4
o5
o6

Modelo acústico

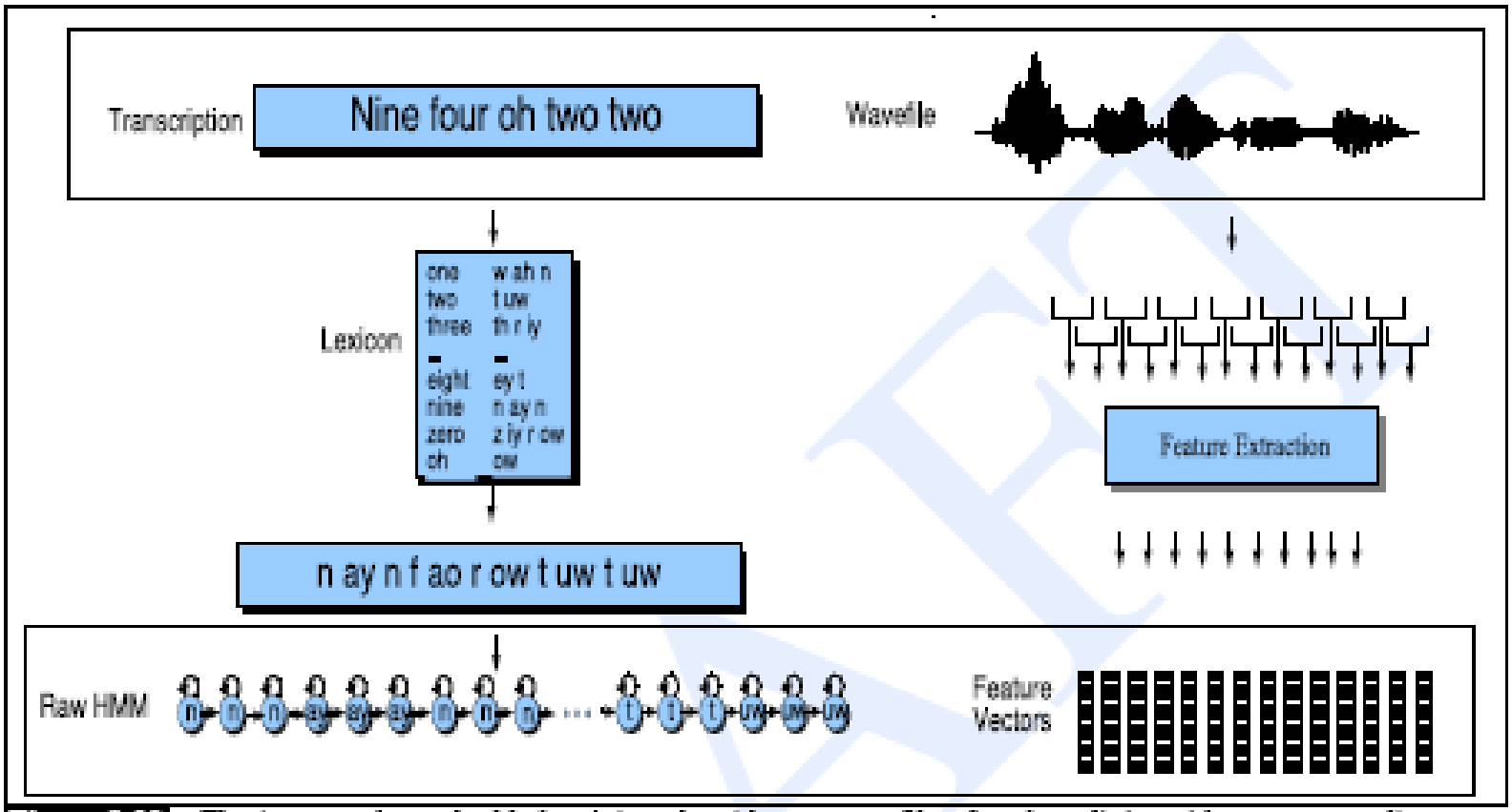


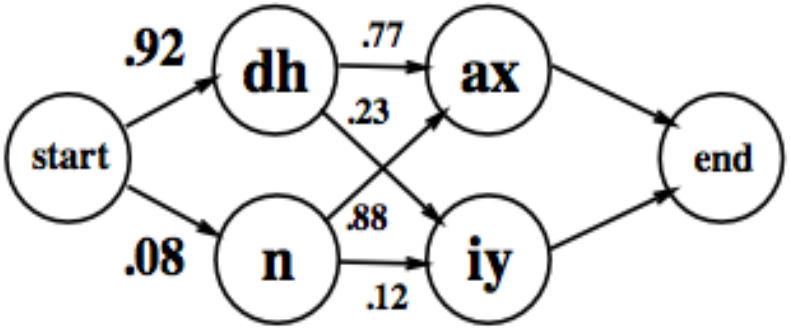
Figure 9.32 The input to the embedded training algorithm; a wavefile of spoken digits with a corresponding transcription. The transcription is converted into a raw HMM, ready to be aligned and trained against the cepstral features extracted from the wavefile.

Modelo de Pronunciacion

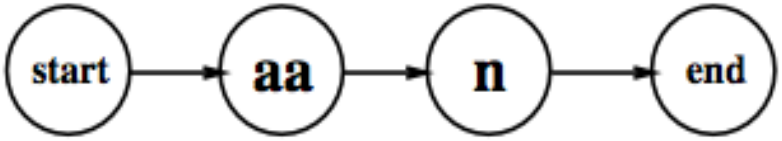
- Modela el likelihood de las palabras dada la red de hipotesis de fonos candidatas
  - Multiples pronuniciaciones para cada palabra
  - Puede ser un autómata o un simple diccionario

- Las palabras se obtienen del cuerpo de entrenamiento
- Las pronunciaciones se obtienen de un diccionario de pronunciacion o de un sistema TTS

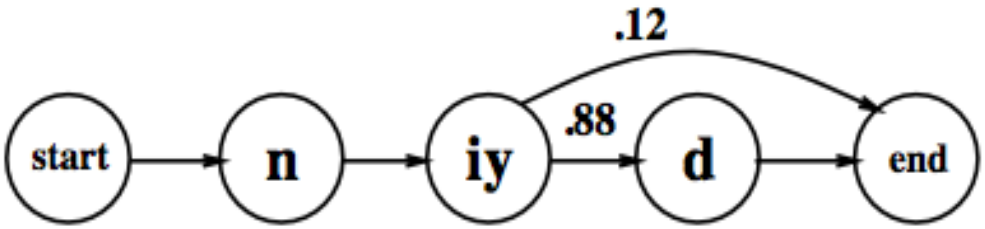
RAH Lexicon: Modelos de Markov para pronunciacion



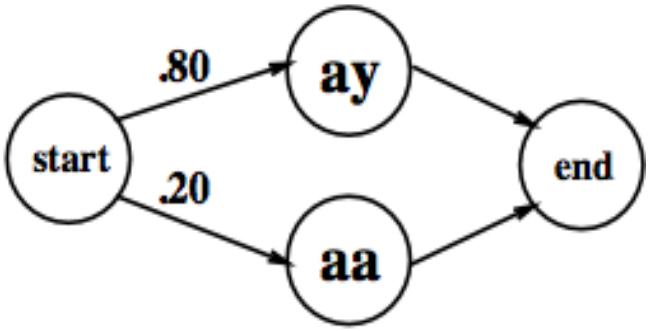
Word model for "the"



Word model for "on"



Word model for "need"



Word model for "I"

Modelo de Lenguaje

- Modela el likelihood de una palabra dada la palabra(s) previa
- Modelos n-gram:
  - Calcula las probabilidades bigram o trigram del texto de entrenamiento : que tan probable es que una palabra siga a otra? Que sigan dos palabras previas?
- Gramaticas
  - Gramáticas de estado finito o gramáticas libres del contexto
- Fuera del vocabulario problema (OOV)

Busqueda/Decodificacion

- Buscar la mejor hipotesis P(O|W) P(W) dado
  - Una secuencia de vectores de caracteristicas acústicas (O)
  - Un HMM entrenado (AM)
  - Lexicon (PM)
  - Probabilidades de secuencias de palabras (LM)
- For O
  - Calcular la secuencia de estados mas probable en HMM dadas las probabilidades de transiciones y observaciones
  - Luego asignar las palabras a los estados
  - N best vs. 1 best vs. lattice output
  - Poda: beam search

Recursos:

Visite mi web page

- http://jorge.sistemasyservidores.com/

Visite proyectos de alumnos en IA

- http://jorge.sistemasyservidores.com/ia-2008unt/index.html

Visite proyectos de alumnos en RAH

- http://jorge.sistemasyservidores.com/topicosiii-2007ii/index.htm

- Proyectos realizados con alumnos

- *Extracción de características de la señal de voz utilizando LPC-Cepstrum - Jorge Velarde, Jhon Franko, Pretel Jesús, Alicia Isolina*
- *Prediccion Lineal Perceptual PLP - Alan Alfredo Collantes Arana Dany Richard Sari Bustos*
- *Audio files compression through wavelets - Fredy Carranza-Athó*
- *Máquinas de Sopoerte Vectorial en el Reconocimiento Automático del Habla - Juan Carlos Federico Roeder Moreno*
- *Efectos de las diferentes transformadas del coseno en RAH Márquez Fernández, Luz Victoria*

- CONTACTO:
- mail
  - jorgeluis.guevaradiaz@gmail.co,
  - jorge.jorjasso@gmail.com
  - jorjasso@hotmail.com
  - jguevara@unitru.edu.pe
  - jgd@upnorte.edu.pe
- Telefono
  - 044 – 949434637
  - 044 – 270219