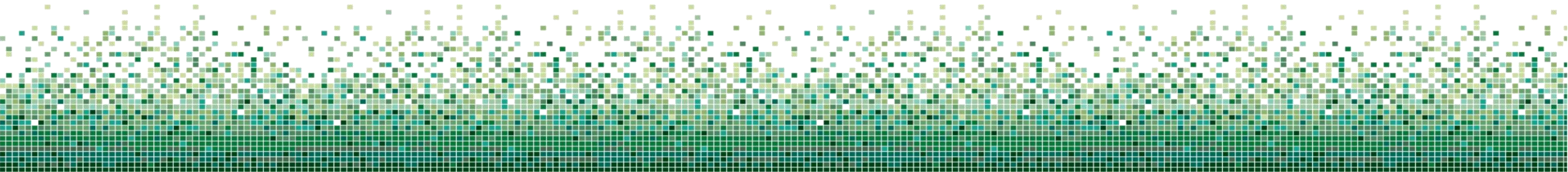
Inteligencia Artificial & Tecnologías del Habla

Dr. Bioing. Diego Evin

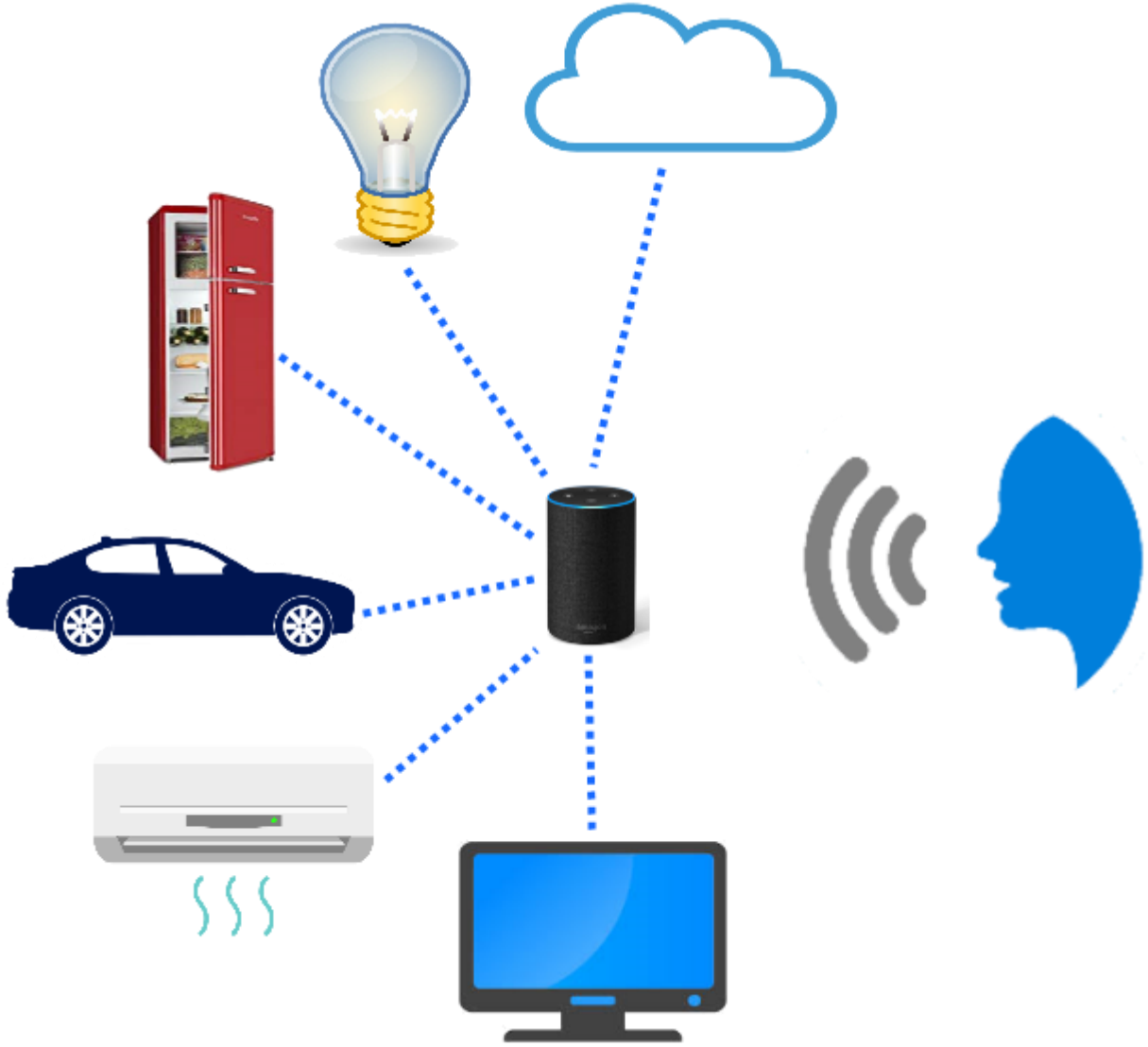
CINTRA (CONICET – UTN)
Universidad Nacional de Entre Ríos



Motivación

- **Habla** elemento esencial en comunicaciones humanas e interacciones hombre-máquina

- **Transformación digital** uso de tecnología digital para resolver problemas tradicionales



Justificación

- Intuitivo (Modo más natural de interacción del ser humano)
- Veloz (3x más rápido que tipeo, 65% más preciso)
- Uso de canales digitales que los clientes ya manejan

Tecnologías del Habla

Areas de Investigación

- **Reconocimiento del Habla (ASR)**
- **Síntesis del Habla (TTS)**
- **Identificación de Hablantes (SID)**
- **Codificación de Habla**
- **Realce del Habla**
- **Identificación del lenguaje hablado (LID)**
- **Interacción Multimodal**

Tecnologías del Habla

HCI

- Agentes inteligentes
- IVRs
- Comandos por voz
- Navegación por voz

Comunicación

- Filtrado
- Encoders
- Realces

Biometría y Clínica

- Reconocimiento de hablantes
- Detección de patologías
- Terapéuticas

Entretenimiento

- Síntesis de Canto
- Conversión de voces
- Avatares
- Video Juegos
- Juguetes

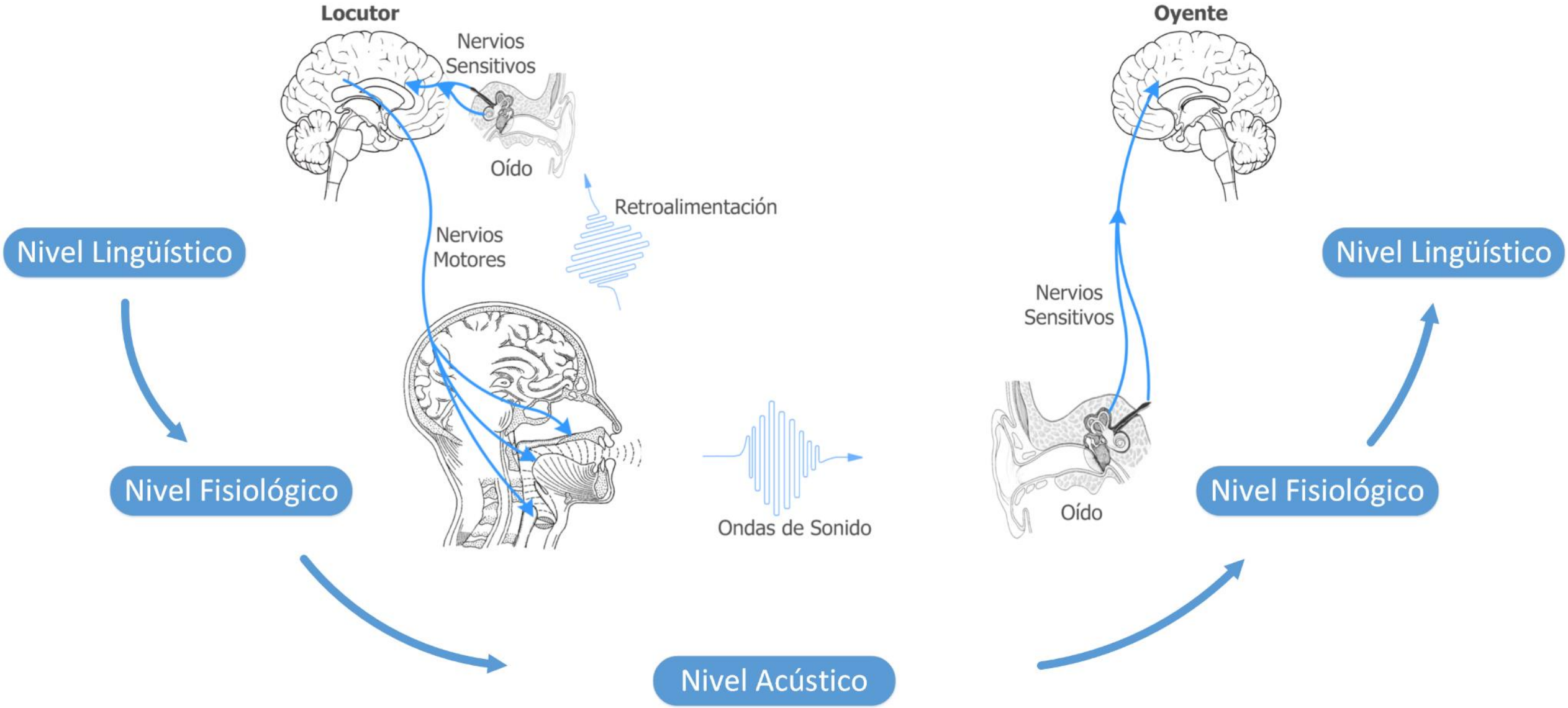
Educación

- Enseñanza Idiomas
- Canto

Varios

- Traducción habla-habla
- Analítica de habla
- Monitoreo de medios

Reconocimiento Humano del Habla (RHH)



RHH – Niveles Lingüísticos

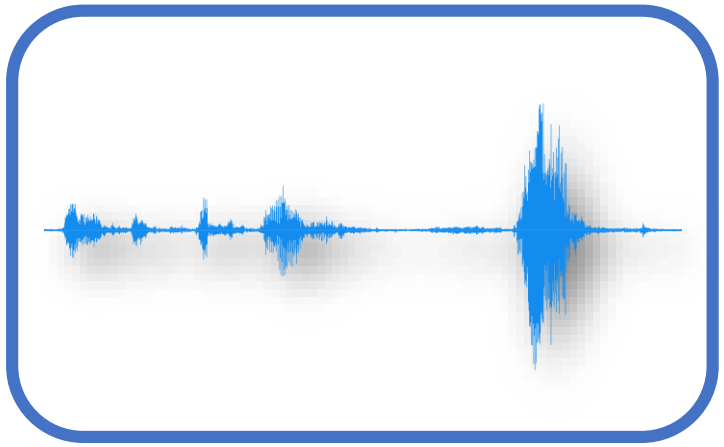
Fonética	• Realización física de sonidos de una lengua (fonos)
Fonología	• Función de sonidos a nivel abstracto o mental (fonemas)
Morfología	• Formación las palabras (morfemas)
Prosodia	• Uso de melodía, acentos, duraciones para agregar información
Sintaxis	• Estructura del lenguaje (combinar palabras en oraciones)
Semántica - Pragmática	• Significado de las oraciones y reglas conversacionales

RHH - Complejidad

- Hablante nativo ~ 600/900 sonidos por minuto
- Cada sonido uno de al menos 24 fonemas (Castellano)
- Alrededor de 42.000 lemas posibles * (persona de 20 años)
- Las palabras pueden admitir diferentes pronunciaciones
- No hay delimitación de inicio y fin entre palabras
- Además discriminar voz/ruido, seguir a un hablante específico, …

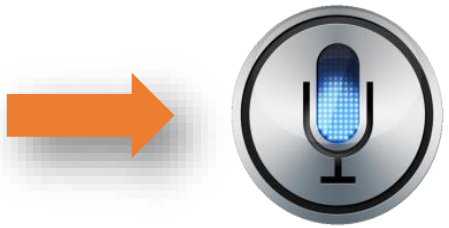
* Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in psychology*, *7*, 1116.

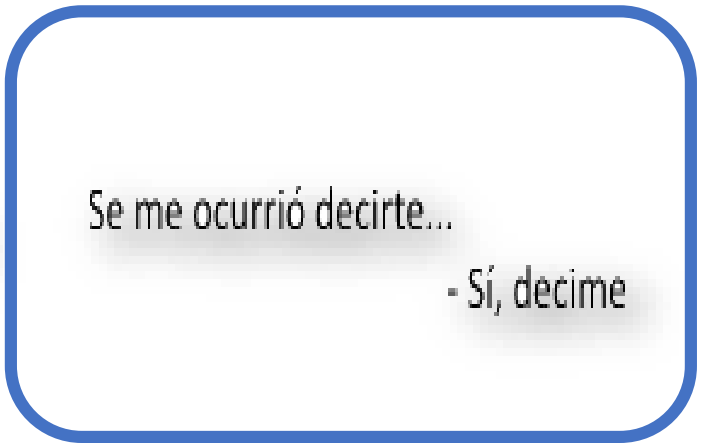
Reconocimiento Automático del Habla (RAH)



Audio de Entrada

Sistema de Reconocimiento





Texto Reconocido

Tareas Relacionadas

- Detección de Habla
- Detección de Lenguaje
- Identificación de Hablantes
- Comprensión del Habla

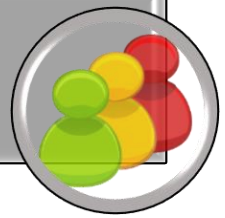
RAH Dificultad del Problema - Variabilidad

- Estilos y velocidad
- Coarticulación
- Salud, Emociones
- Interlocutor
- Edad

Variabilidad Intra-Locutores

- Diferencia anatómica
- Socio-Culturales
- Acentos

Variabilidad Inter-Locutores



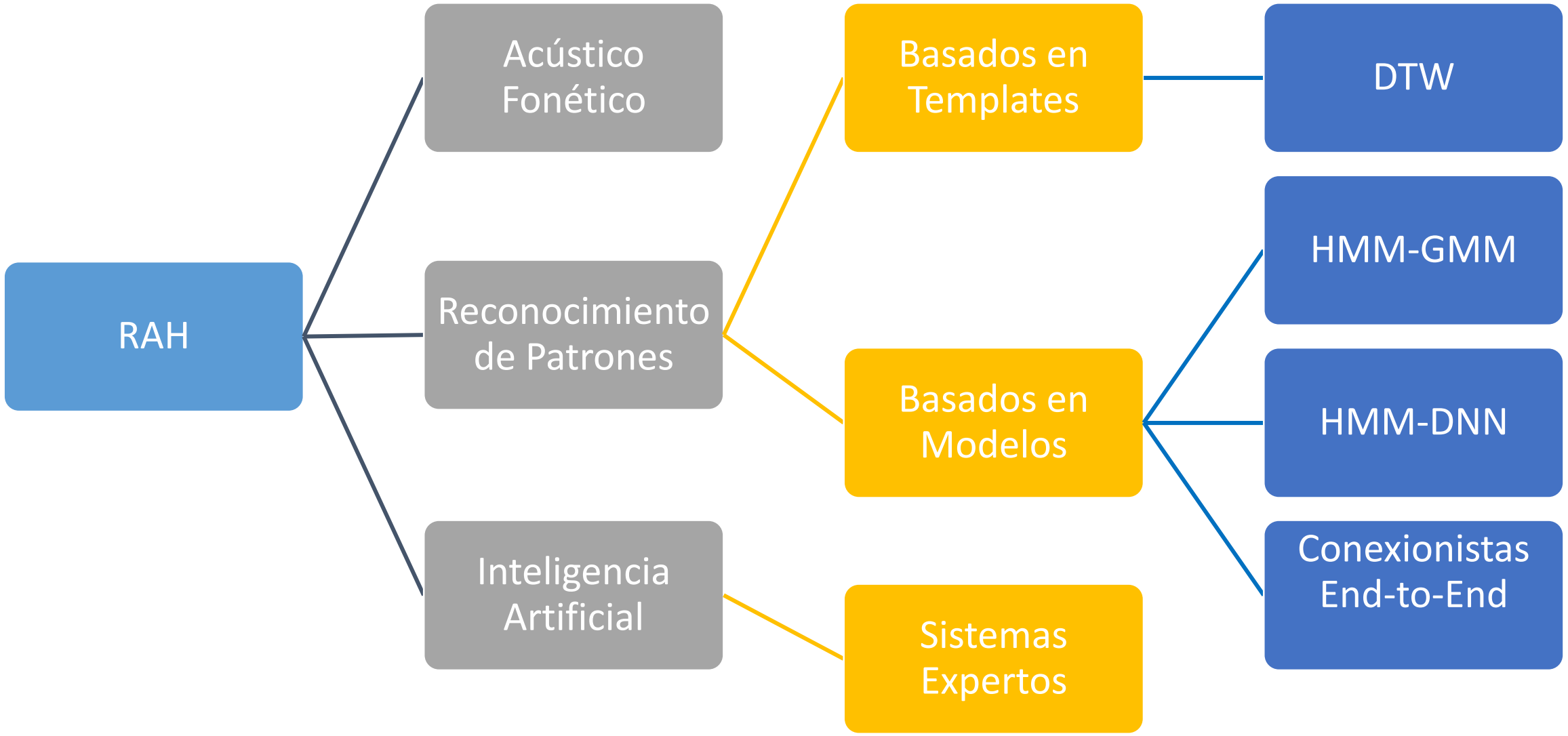
- Del mismo hablante
- De otros hablantes
- Del medio acústico

Ruidos e Interferencias

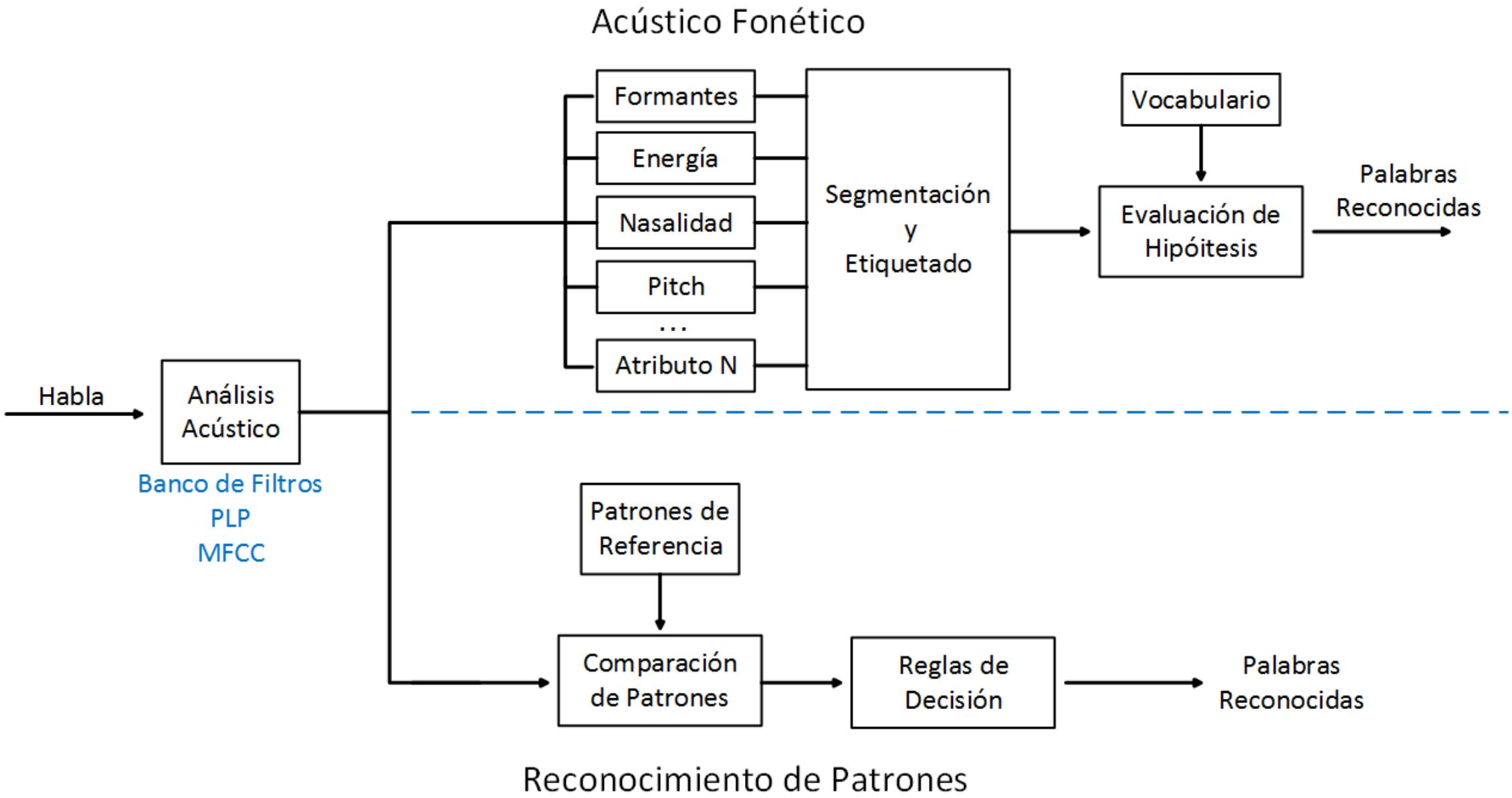
- En sensores
- Canales
- Codecs

Variaciones en el Medio

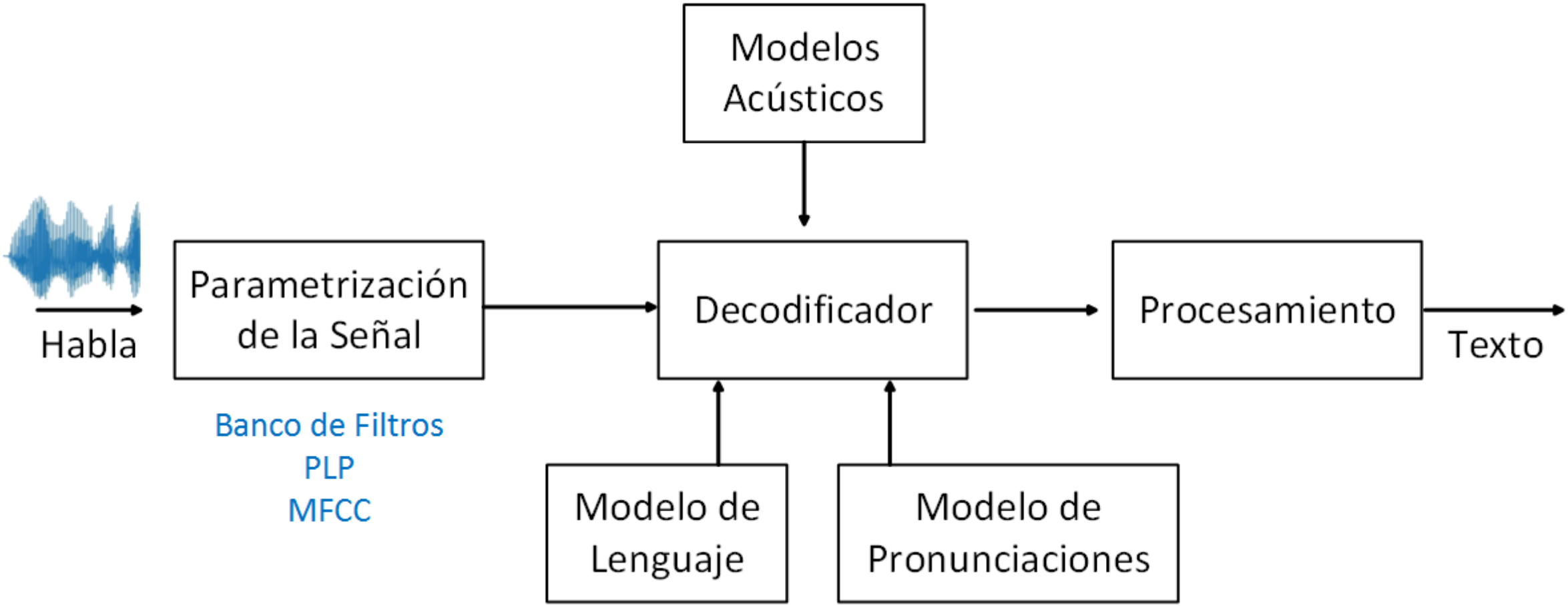
Aproximaciones al RAH



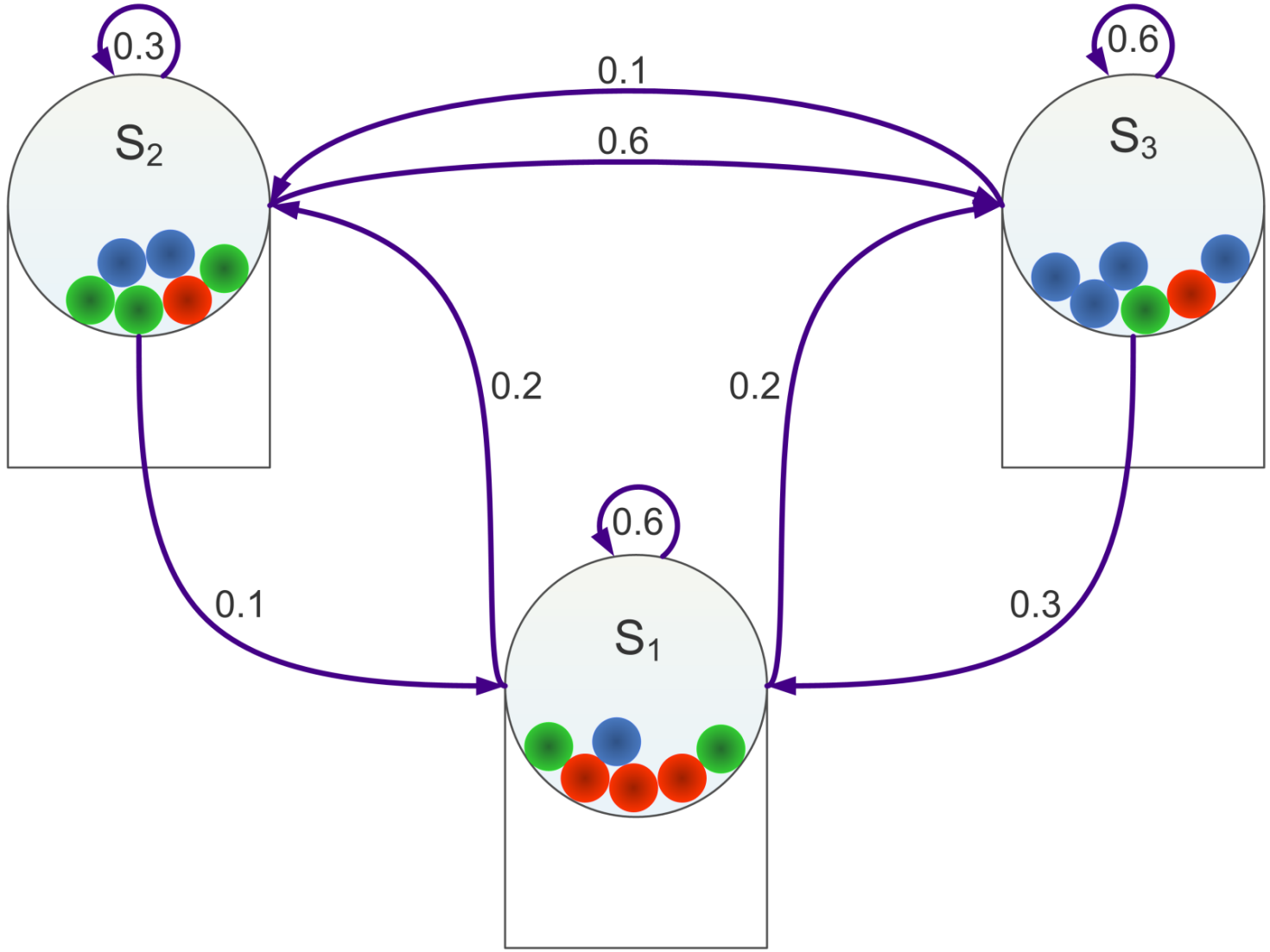
Aproximaciones al RAH



RAH – Esquema Estándar



Modelos Ocultos de Markov (HMM)



Modelos Ocultos de Markov (HMM)

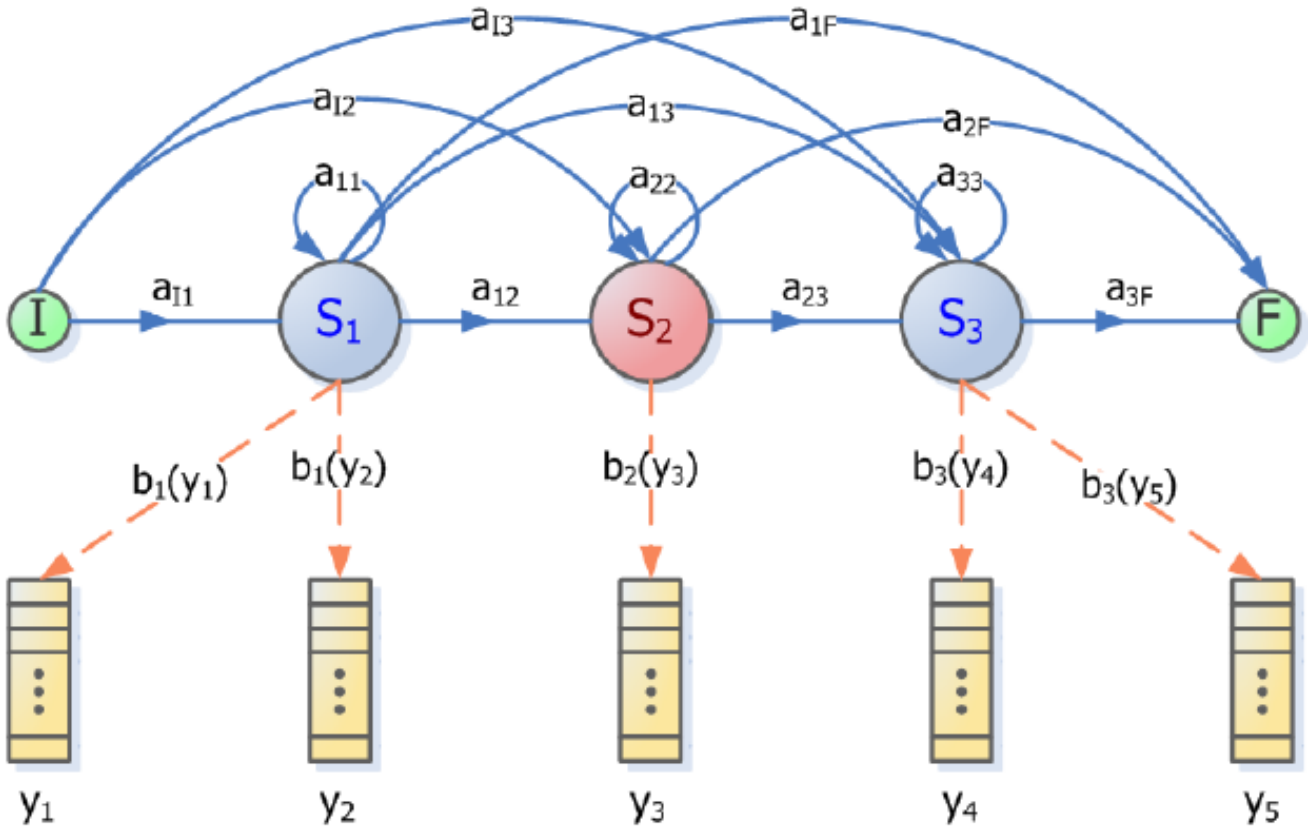
Modelan procesos temporales discretos bivariados: $\{S_k, O_k\}$

Cada HMM está caracterizado por la tupla $\lambda(\boldsymbol{S}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}, \boldsymbol{O})$

- $\boldsymbol{S} = \{S_1, S_2, \cdots, S_N\}$ estados posibles del modelo
- $\boldsymbol{A} = \{a_{ij}\}$ matriz de transiciones entre estados
- $B = \{b_j(k)\}$ probabilidad de emisión del símbolo y_k al activarse S_j
- $\pi = \{\pi(i)\}$ distribución de probabilidades para estados iniciales
- $O = \{o_1, o_2, \cdots, o_M\}$ posibles observaciones de las emisiones

RAH Basado en GMM-HMM

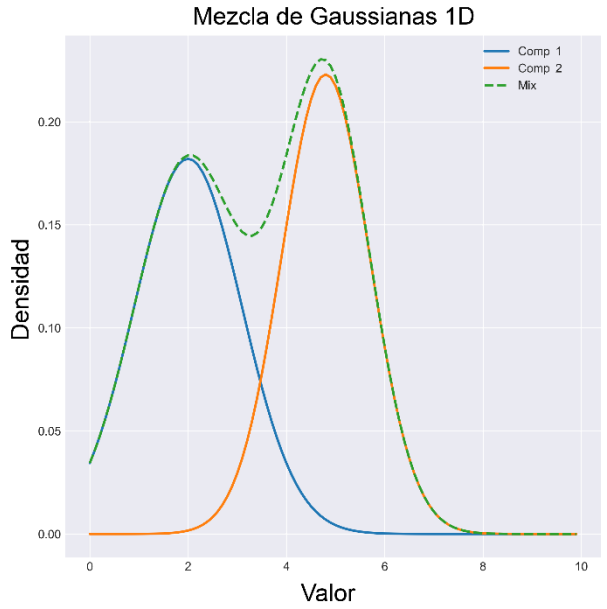
- HMM resuelven implícitamente segmentación y clasificación de unidades
- Entrenamiento escalable, métodos eficientes de aprendizaje y decodificación
- Estado del arte por más de 30 años

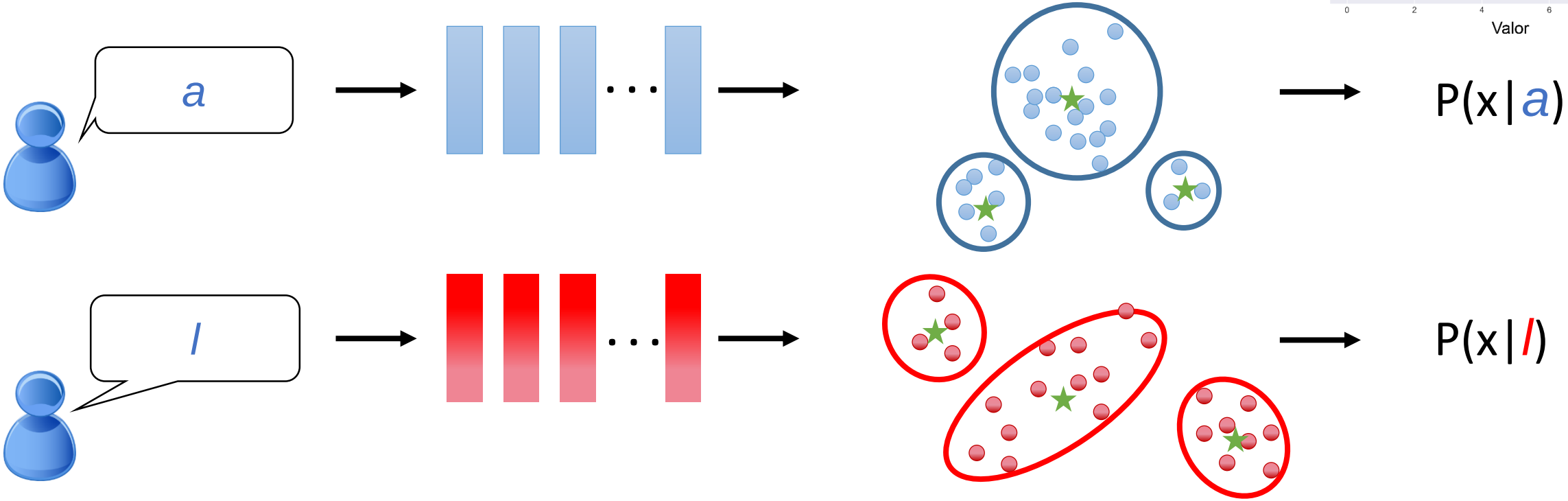


Modelos Ocultos de Markov

RAH Basado en GMM-HMM

Asume distribución estacionaria de atributos acústicos en cada estado





Gaussian Mixture Model (GMM)

RAH Basado en GMM-HMM

Dados:

$\boldsymbol{O}$: secuencia de T atributos acústicos (MFCC, PLP, LPCC)

$\boldsymbol{W}$: secuencia de N palabras de un vocabulario fijo y conocido

$\boldsymbol{P(W|O)}$: probabilidad condicional de la secuencia de palabras $\boldsymbol{W}$, dada la secuencia observada $\boldsymbol{O}$

Resultado del reconocedor: $W = \underset{W}{\operatorname{argmax}} P(W|O)$

"Cuál es el texto más probable de todo el lenguaje, teniendo como evidencia los rasgos acústicos observados"

RAH Basado en GMM-HMM

Usando la regla de Bayes:

$$P(W|O) = \frac{P(O|W)\,P(W)}{P(O)}$$

Donde:

$P(W)$: Probabilidad a priori de observar la secuencia de palabras **W**

$P(O|W)$: Probabilidad a priori de que cuando una persona pronuncia la secuencia de palabras **W** se observe la secuencia de evidencias acústicas **O**

$P(O)$: Probabilidad de la secuencia de evidencias acústicas **O**

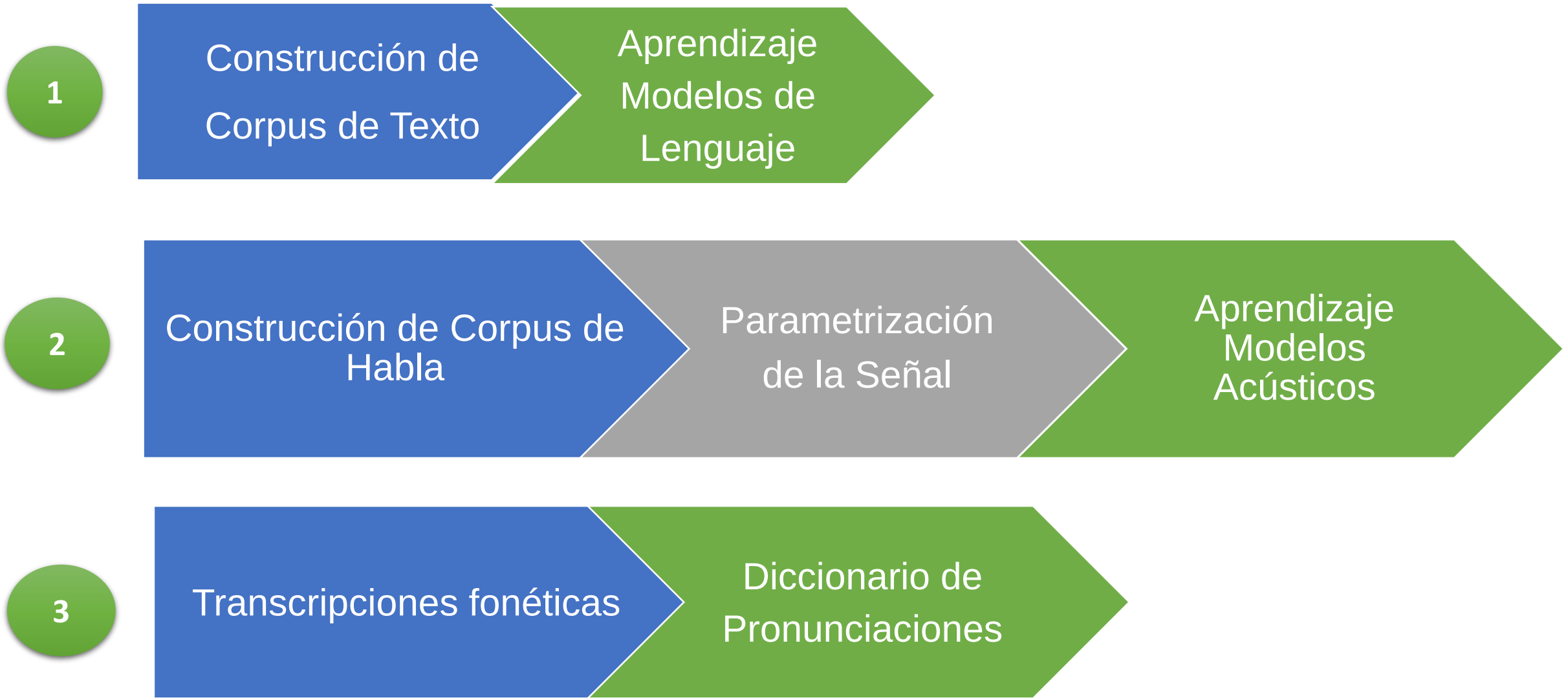
Se puede reescribir una expresión más práctica para el reconocedor:

$$\overline{W} = \underset{W}{\operatorname{argmax}}\, P(O|W)P(W)$$

Modelo Acústico

Modelo de Lenguaje

Reconocimiento Automático del Habla

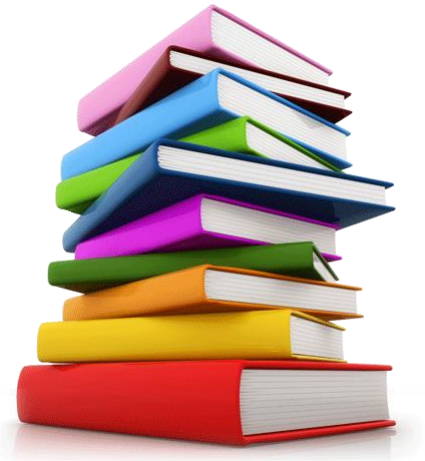


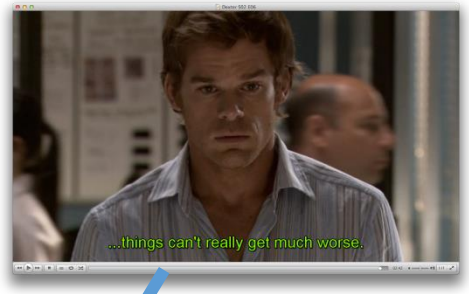
Corpus de Texto

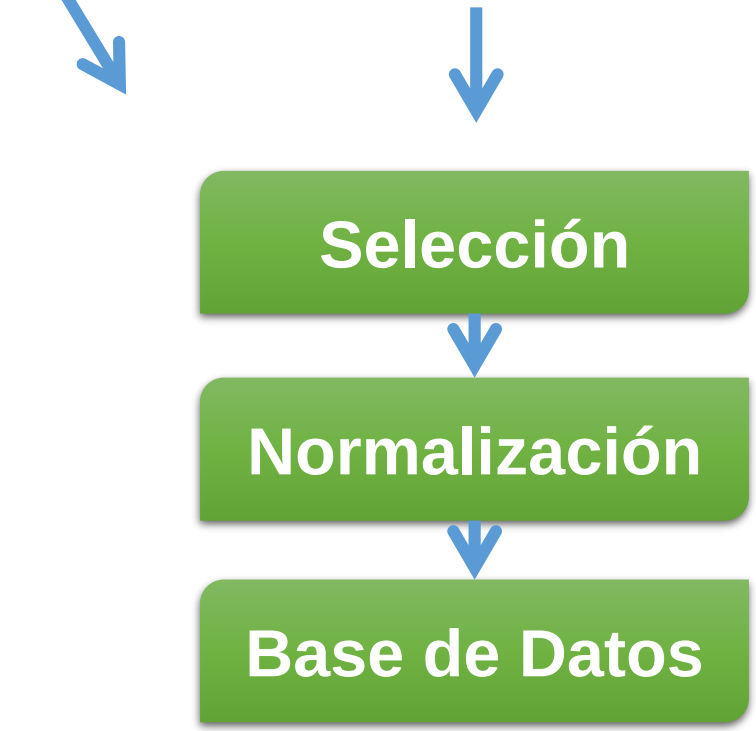
Objetivo: obtener datos sobre cómo se usan las palabras en el dominio de interés

Métodos: Web as a Corpus, libros, diarios, subtítulos, etc.

Desafíos : Conseguir el datasets de contenido similar a las secuencias de palabras a reconocer, adaptar corpus





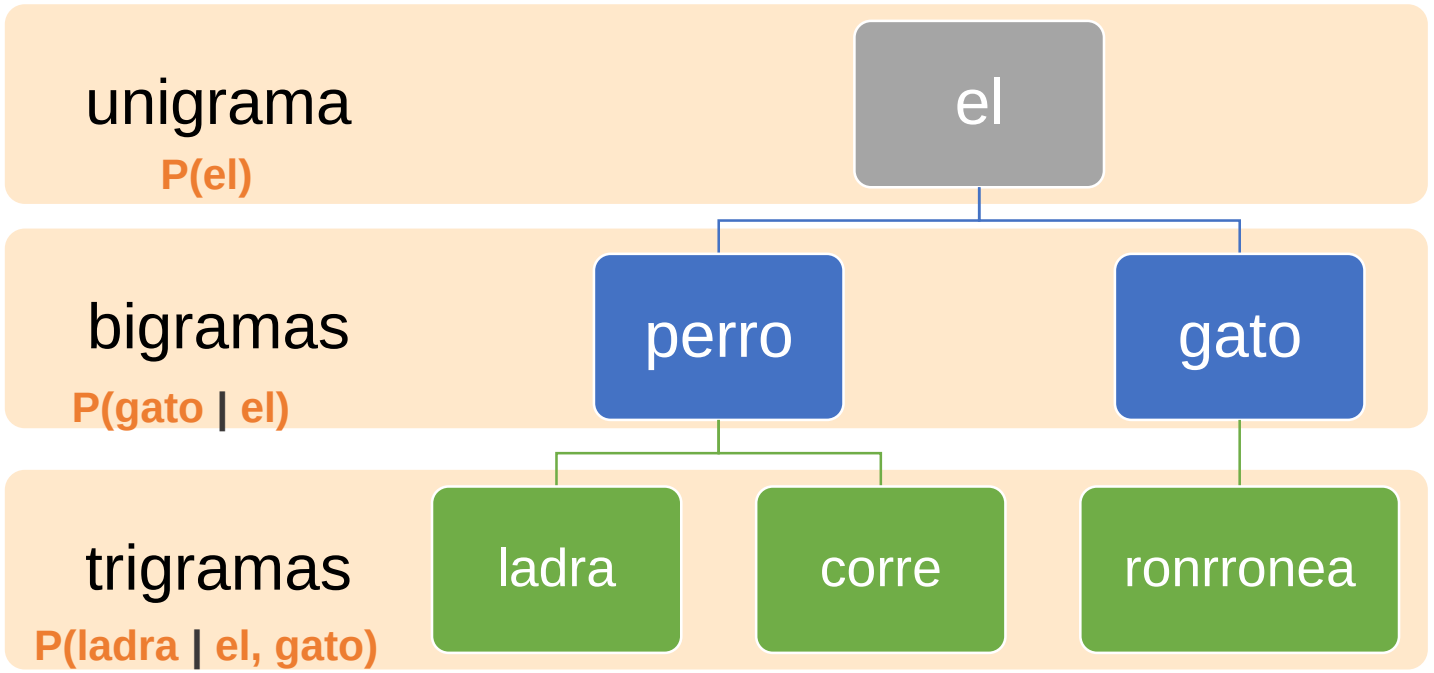


Modelo de Lenguaje

Objetivo: Modelar frases válidas según la sintaxis

Métodos: Basados en reglas (CFG), estadísticos (N-gramas), conexionistas (RNN)

Desafíos : Cómo construir rápida y eficientemente un modelo de lenguaje para una tarea nueva (otro contexto)



Corpus de Habla

Contenido por frase:

file.lab: transcripción ortográfica

file.wav archivo con la señal acústica

file.gra anotación ortográfica con

alineamiento temporal a nivel palabras

file.fon anotación fonética con

segmentación temporal

Tipos de Frases:

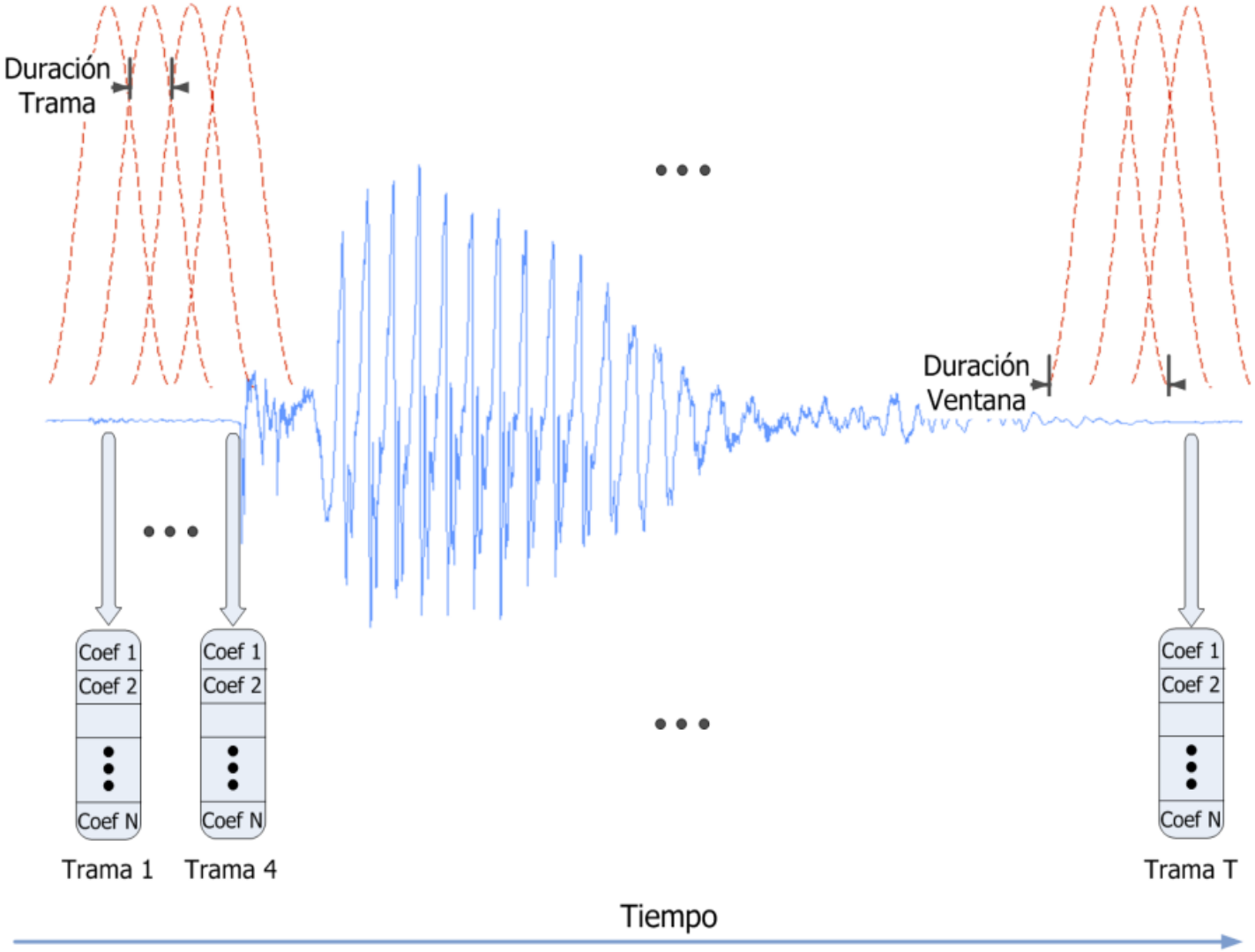
- Balanceadas fonéticamente: buena cobertura de cada combinación de sonidos
- De múltiples hablantes: para aprender un modelo acústico que sirva para todos
- Diversidad prosódica: para aprender cómo varían los sonidos con la entonación
- Diversas pronunciaciones dialectales
- …

Parametrización del Habla

Objetivo: extraer características robustas y relevantes para clasificar

Métodos: Análisis STF (MFCC, LPC, Rasta), compensaciones no lineales y normalización. Cada ventana se representa usando ~40 rasgos.

Desafíos : compresión, robustez al entorno, dispositivos, locutores, ruido y ecos.

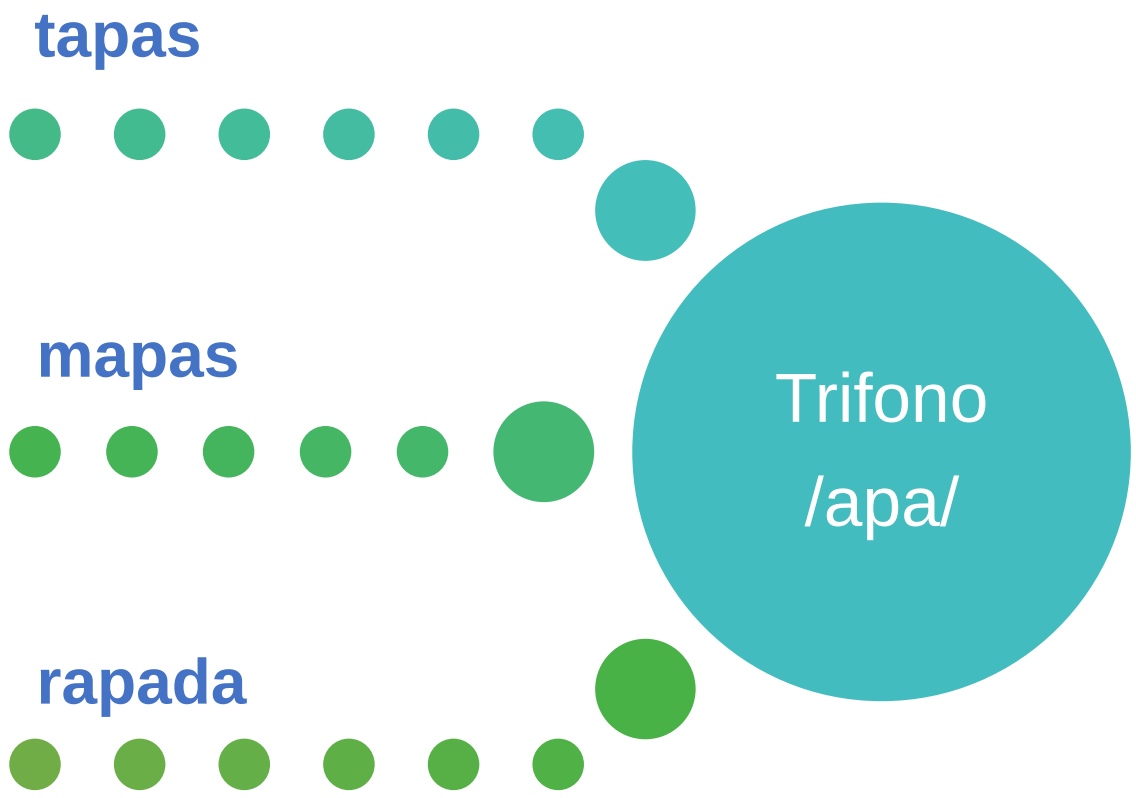


Entrenamiento Modelos Acústicos

Objetivo: Obtener modelos que permitan caracterizar los sonidos del habla

Métodos: Se representa cada unidad acústica con HMMs . Probabilidades de emisión fdp GMM o ANN

Desafíos : precisión, robustez al entorno, dispositivos, locutores, ruidos y ecos.

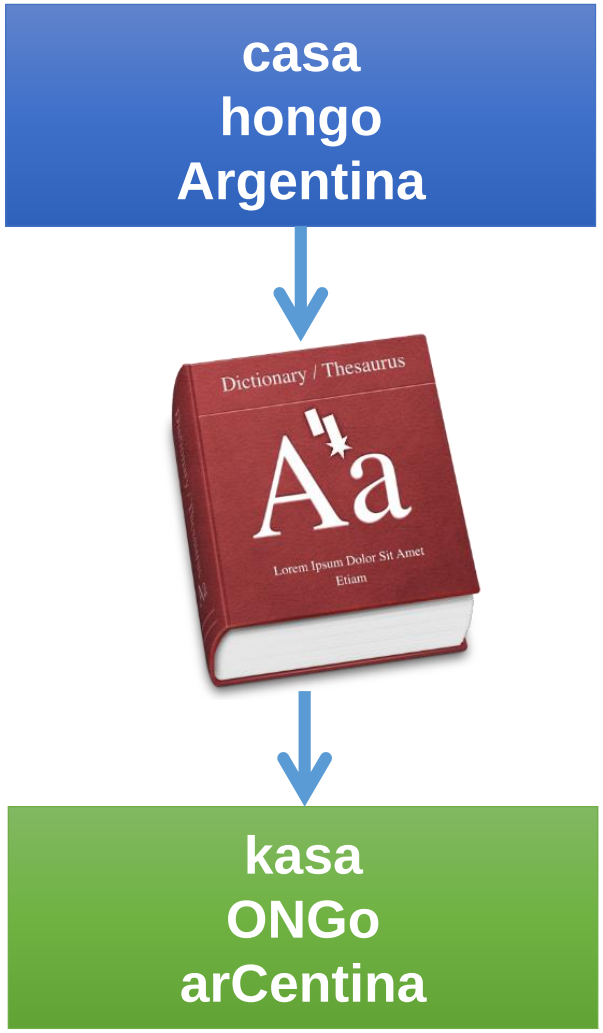


Modelo de Pronunciaciones

Objetivo: Indicar cómo se conforma cada palabra a reconocer en término de las unidades usadas en los modelos acústicos

Métodos: Basados en reglas, o en Machine Learning

Desafíos : generar de manera automática un lexicón, agregar nuevas variantes dialectales y pronunciaciones

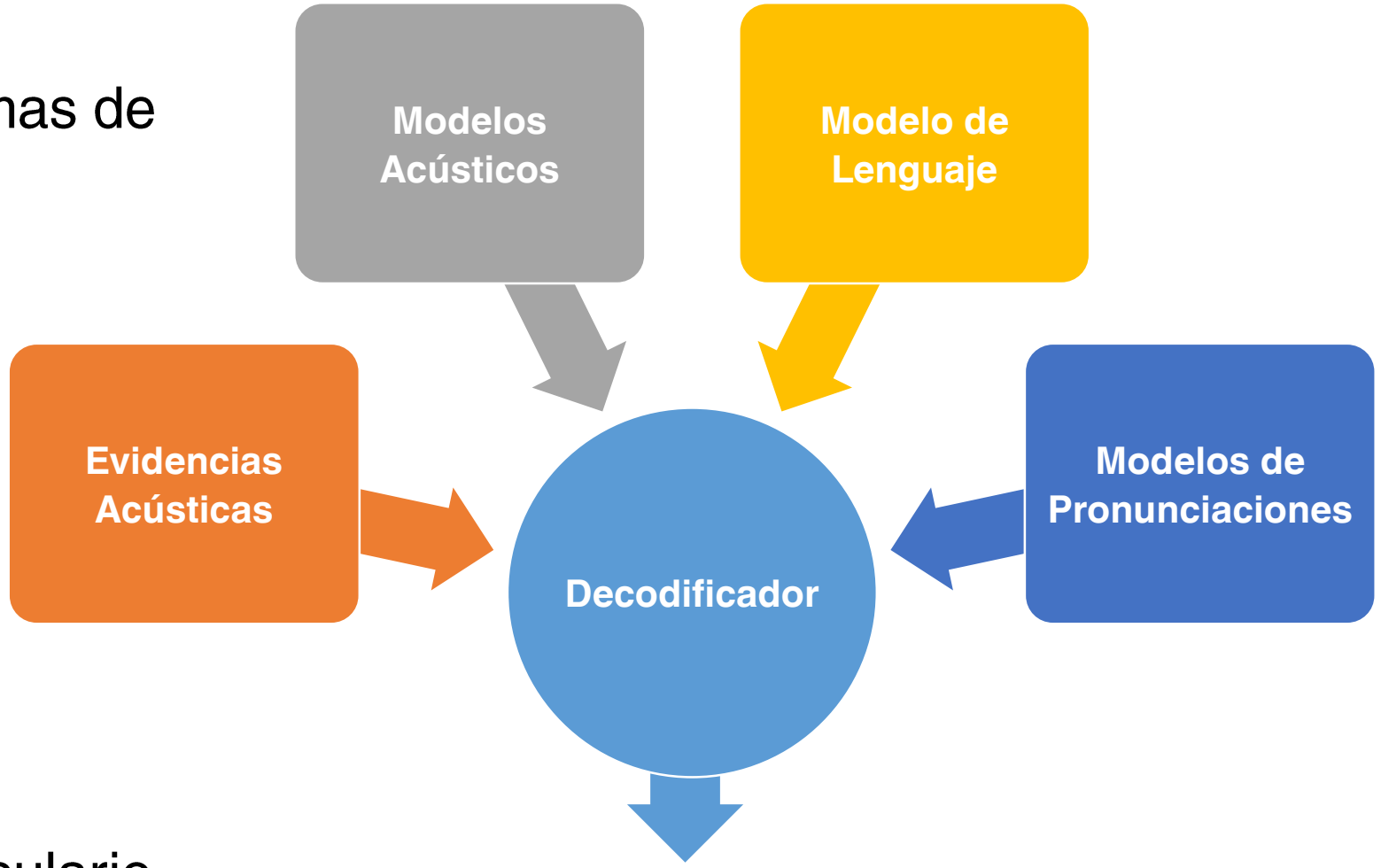


Decodificador

Objetivo: generar secuencia óptimas de palabras combinando el modelo acústico, de lenguaje y de pronunciaciones

Métodos: Viterbi, WFST

Desafíos : Construir estructuras eficientes para decodificación y búsqueda en tareas de gran vocabulario y modelos de lenguajes complejos.



Problemas de RAH

Tipo de Habla

- Aislada
- Conectada
- Continua
- Diálogo

Vocabulario

- Pequeño
- Mediano
- Grande

Conocimiento del Usuario

- Dependiente
- Adaptable
- Independiente

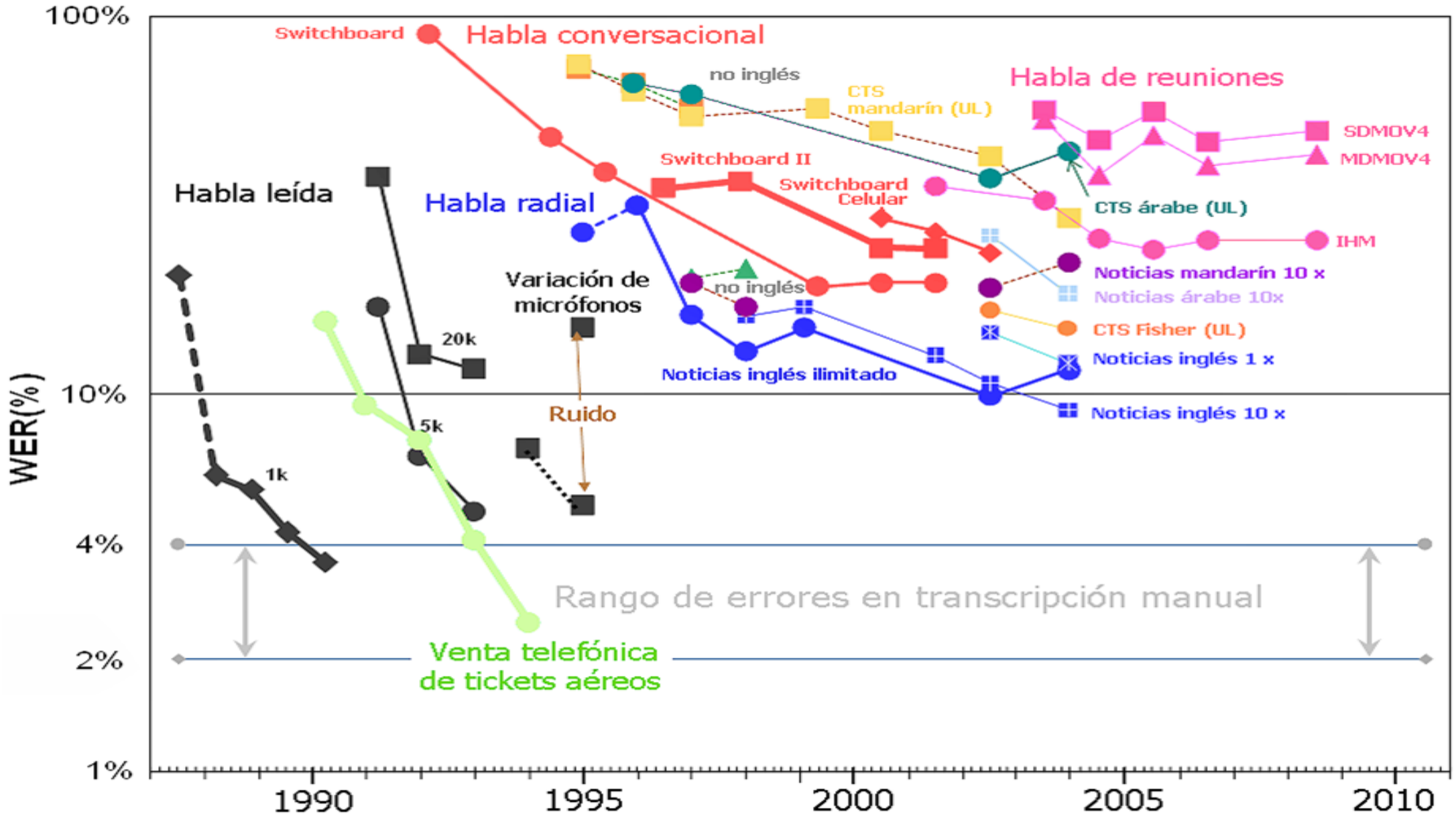
Condiciones de Uso

- Laboratorio
- Oficina
- Robusto

Tipo de Aplicación

- Comando
- Palabras clave
- Dictado
- Close caption

Evolución Histórica



Redes Neuronales y RAH

- Propuesto hace mucho
- Dificultad en entrenar redes con más de 2 capas (Vanishing Gradients)
- Corpus pequeños
- Capacidad de Cómputo Limitada

Temporal & Time-Delay (1-D Convolutional) Neural Nets

- Atlas, Homma, and Marks, “An Artificial Neural Network for Spatio-Temporal Bipolar Patterns, Application to Phoneme Classification,” NIPS, 1988.
- Waibel, Hanazawa, Hinton, Shikano, Lang. “Phoneme recognition using time-delay neural networks.” IEEE Transactions on Acoustics, Speech and Signal Processing, 1989.

Hybrid Neural Nets-HMM

- Morgan and Bourlard. “Continuous speech recognition using MLP with HMMs,” ICASSP, 1990.

Recurrent Neural Nets

- Bengio. “Artificial Neural Networks and their Application to Speech/Sequence Recognition”, Ph.D. thesis, 1991.
- Robinson. “A real-time recurrent error propagation network word recognition system,” ICASSP 1992.

Neural-Net Nonlinear Prediction

- Deng, Hassanein, Elmasry. “Analysis of correlation structure for a neural predictive model with applications to speech recognition,” *Neural Networks*, vol. 7, No. 2, 1994.

Bidirectional Recurrent Neural Nets

- Schuster, Paliwal. "Bidirectional recurrent neural networks," IEEE Trans. Signal Processing, 1997.

Neural-Net TANDEM

- Hermansky, Ellis, Sharma. "Tandem connectionist feature extraction for conventional HMM systems." ICASSP 2000.
- Morgan, Zhu, Stolcke, Sonmez, Sivadas, Shinozaki, Ostendorf, Jain, Hermansky, Ellis, Doddington, Chen, Cretin, Bourlard, Athineos, “Pushing the envelope - aside [speech recognition],” IEEE Signal Processing Magazine, vol. 22, no. 5, 2005.

Bottle-neck Features Extracted from Neural-Nets

- Grezl, Karafiat, Kontar & Cernocky. “Probabilistic and bottle-neck features for LVCSR of meetings,” ICASSP, 2007.

Redes Neuronales y RAH

Deep Learning (2009-2010)

- Nuevo paradigma inspirado en el procesamiento de información sensorial humano
- El cerebro no pre-procesa información sensorial, sino que permite su propagación por módulos que aprenden a representar las observaciones
- Emplean abstracciones jerárquicas y construcción gradual de representaciones en niveles incrementales de abstracción
- Buscan capturar dependencias espacio-temporales en base a regularidades en las observaciones

Redes Neuronales y RAH

- Introducción de **Redes de Creencia Profunda** (DBN) y algoritmos basados en auto-codificadores permiten obtener de forma no supervisada un pre entrenamiento
- Guían el entrenamiento de los niveles de representación intermedios usando aprendizaje no supervisado a nivel local
- Redes con condiciones iniciales adecuadas, permiten entrenar modelos más profundos

Redes Neuronales y RAH

G. Hinton, et al. Deep Belief Networks for Phone Recognition.
Proc. NIPS Workshop, 2009

- Corpus TIMIT
- Modelo de Lenguaje de Bigramas sobre fonos

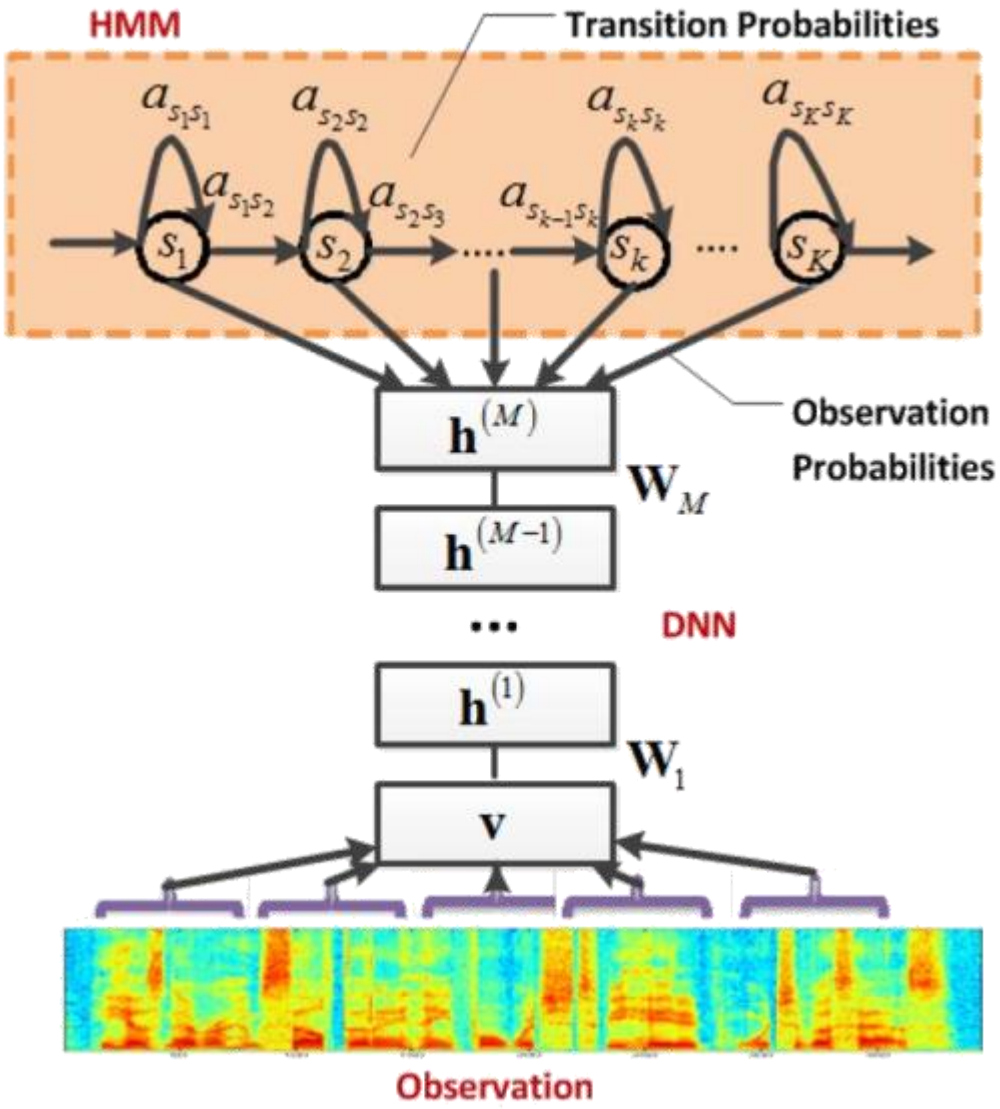
Method	PER %
Recurrent Neural Network	26,1
Bayesian Triphone HMM	25,6
Monophone HTM	24,8
Heterogeneous Classifiers	24,4
DBNs	**23,0**

RAH Basado en DNN-HMM

- Con grandes cantidades de datos y regularización adecuada se puede usar backpropagation para redes profundas (**DNN**)
- Pre-entrenamiento de DBN es innecesario
- Introducción de ReLUs y Drop-out
- Sistemas híbridos HMM/DNN pasan a dominar el estado del arte en ASR
- Aprendizaje de banco de filtros mediante NN, mejora respecto a MFCC en modelos profundos
- Introducción de Deep Learning en LM mejora el desempeño sobre N-Gramas

RAH Basado en DNN-HMM

- Entrenar una DNN para asociar cada senon (estado de HMM correspondiente a un trifono) con una neurona de salida (miles de salidas)
- Entrada: frame de observaciones acústicas
- Interpretar la salida de la red como una versión escalada del likelihood
- En el framework HMM, remplazar las GMMs usadas para estimar las fpds de salida con las salidas de la DNN



Sistemas HMM-DNN

Dahl , et al. Context-Dependent Pre-Trained DNN for LVSR
IEEE Trans. On Audio, Speech, And Language Processing, Vol. 20 (2012)

Model	Sentence Error (%)
CD-GMM-HMM ML	39.6
CD-GMM-HMM MMI	37.2
CD-GMM-HMM MPE	36.2
CD-DNN-HMM (5 hidden layers)	**30.4**

Redes Neuronales y RAH

G. Hinton, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition. Signal Processing Magazine (2012).

TASK	HOURS OF TRAINING DATA	DNN-HMM	GMM-HMM WITH SAME DATA	GMM-HMM WITH MORE DATA
SWITCHBOARD (TEST SET 1)	309	18.5	27.4	18.6 (2,000 H)
SWITCHBOARD (TEST SET 2)	309	16.1	23.6	17.1 (2,000 H)
ENGLISH BROADCAST NEWS	50	17.5	18.8	
BING VOICE SEARCH (SENTENCE ERROR RATES)	24	30.4	36.2	
GOOGLE VOICE INPUT	5,870	12.3		16.0 (>> 5,870 H)
YOUTUBE	1,400	47.6	52.3	

Redes Neuronales y RAH

Arquitecturas alternativas

- En algunos problemas las redes recurrentes (RNN) y convolutivas (CNN) tienen mejor desempeño que las DNN
- Mejoras notables al usar ensambles de RNN/LSTM, DNN y CNN

Redes bi-direccionales RNN/LSTM

- Sak, Senior, Beaufays. "LSTM Recurrent Neural Network architectures for large scale acoustic modeling" Interspeech,2014.

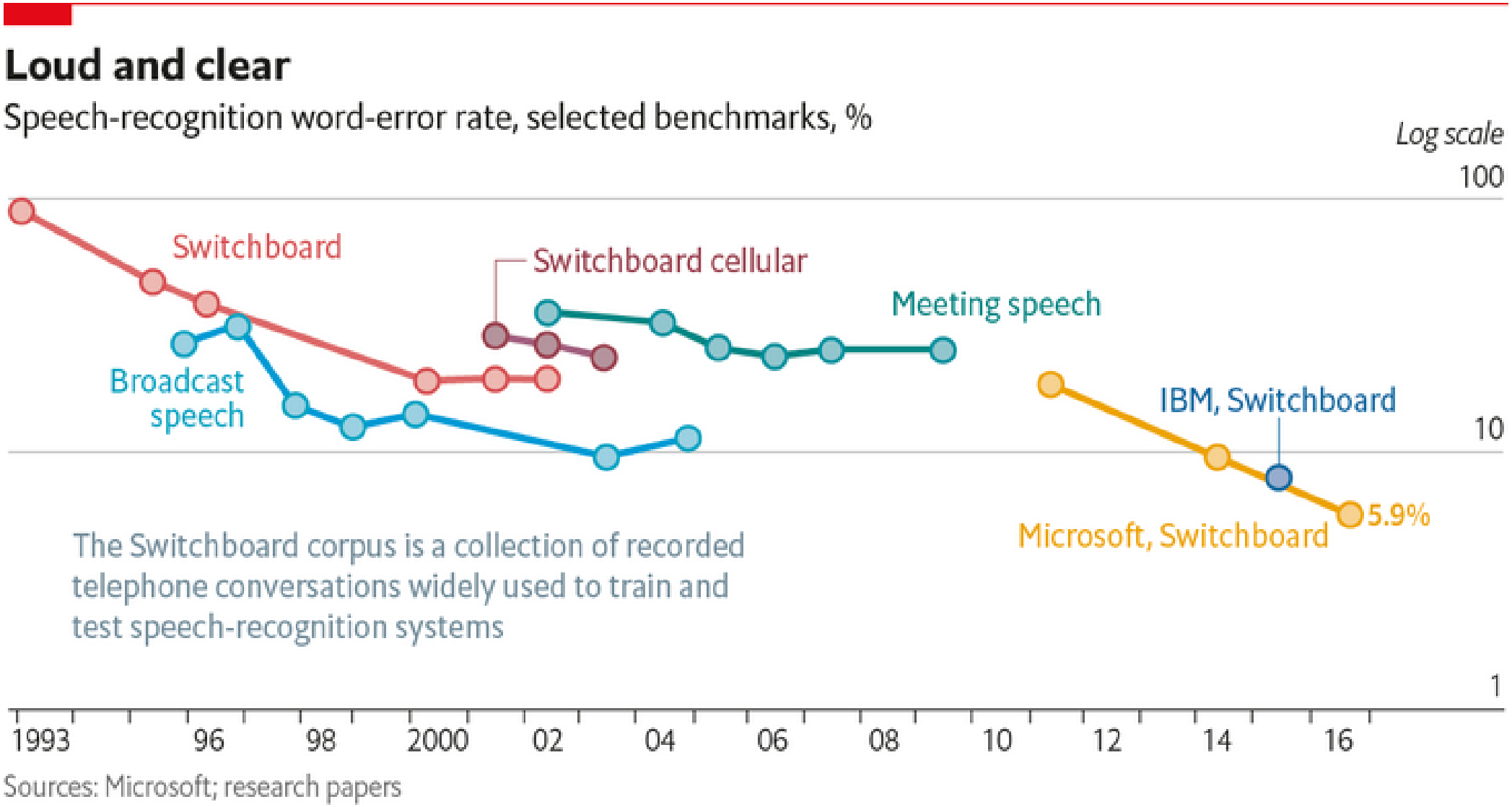
Convolutional Neural Nets

- Soltau, Saon, Sainath. "Joint Training of Convolutional and Non-Convolutional Neural Networks" ICASSP, 2014.

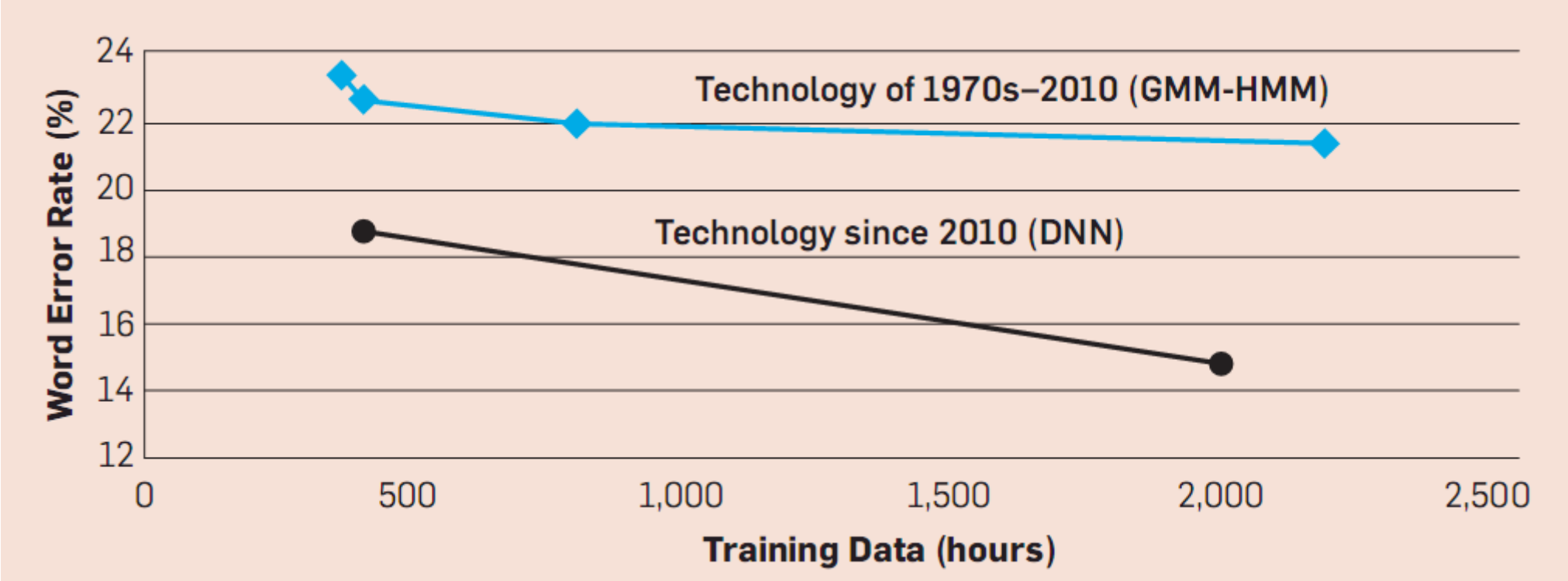
Ensamble Networks

- T. Sainath, O. Vinyals, A. Senior, H. Sak. "Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks" ICASSP 2015.
- L. Deng and John Platt, "Ensemble Deep Learning for Speech Recognition" Interspeech, 2014.

Avances en RAH



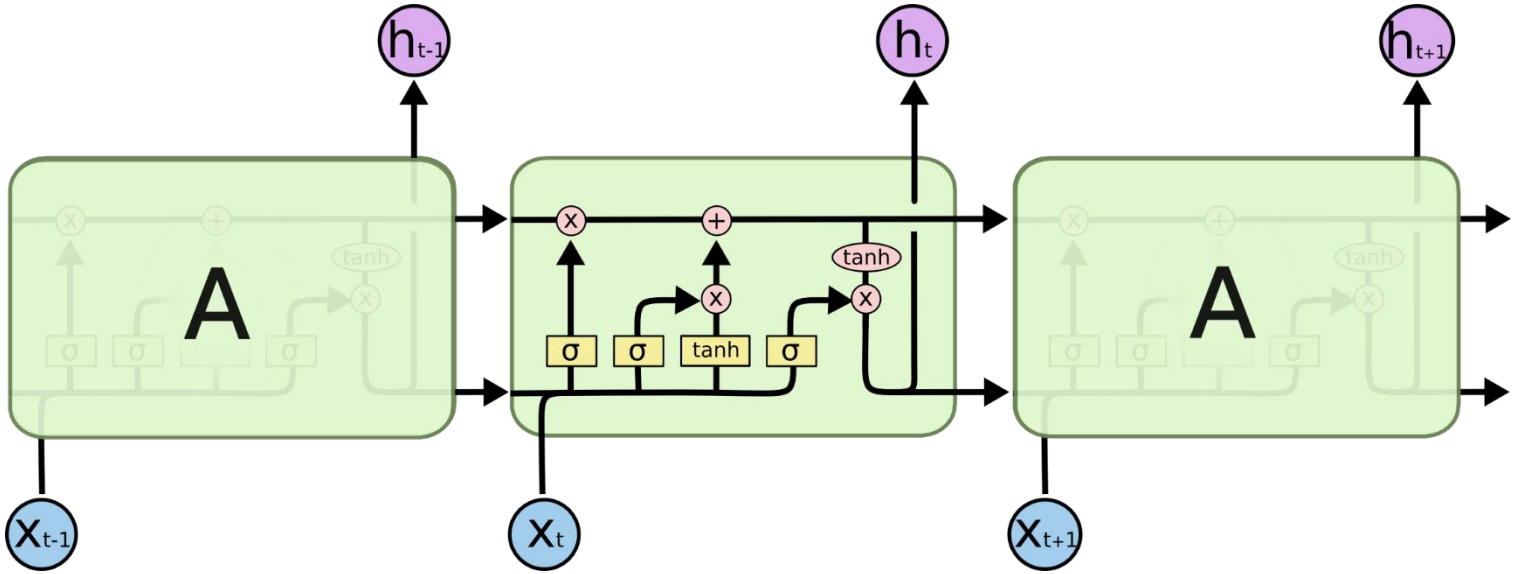
Redes Neuronales y RAH



Huang, X., Baker, J., & Reddy, R. (2014). A historical perspective of speech recognition. *Communications of the ACM*, *57*(1), 94-103.

Modelos de lenguaje empleando RNN

- Entrenar redes recurrentes para predecir la palabra/caracter siguiente dada una historia previa
- Long Short Term Memory (LSTM)
- Permite decidir qué olvidar y qué agregar



RAH Completamente Conexionistas (End to End ASR)

- Línea de investigación más reciente (Google lo adoptó en sus servicios)
- Redes neuronales mapean directamente la señal de habla a texto
- Modelos más simples, no tiene módulos separados para modelo acústico, de lenguaje y pronunciaciones
- No necesitan alineamientos para asociar frames de señal con estados internos o neuronas de salida (como DNNs)
- Es más difícil de explicar y controlar lo que hace internamente

RAH Completamente Conexionistas (E2E)

- Se necesita varios cientos, incluso miles de horas de habla etiquetada para entrenar
- Se necesitan clusters de GPU grandes para entrenarlos en un tiempo razonable
- Diferentes modelos : Basados en Atención, Clasificación Temporal Conexionista (CTC), basados en CNN

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Chen, J. (2016) “Deep speech 2: End-to-end speech recognition in english and mandarin” In International conference on machine learning (pp. 173-182).

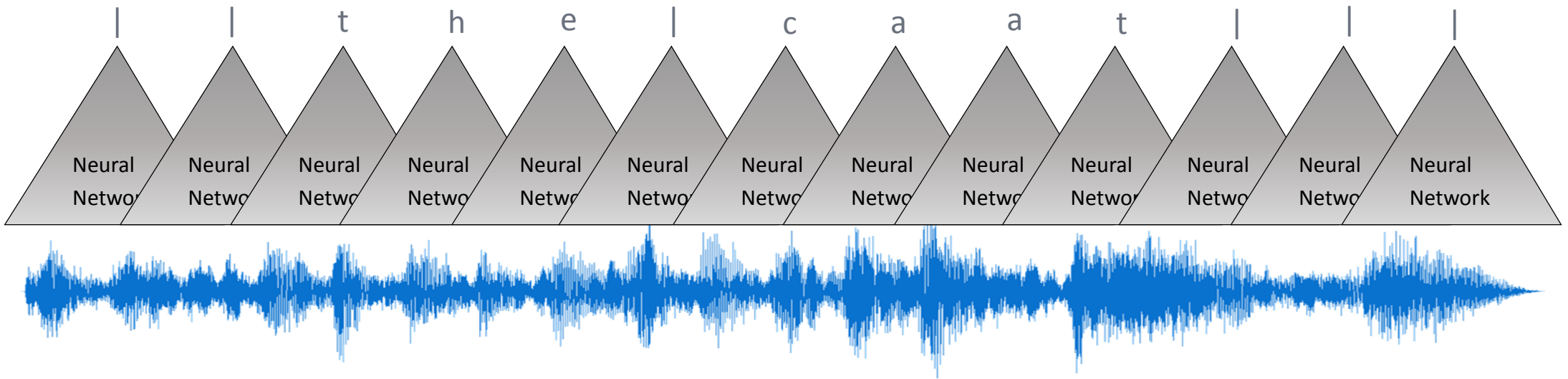
RAH Completamente Conexionista (E2E)

|	10.356
a	0.31
b	-1.254
c	0.001
…	…
y	10.211
z	-10.21

- Entrada frame de observaciones acústicas + contexto
- Salida caracteres y espacios
- Módulo para obtener palabras reconocidas

Modelo de duraciones

|||the|caat|||ssaattt| → |the|cat|sat|



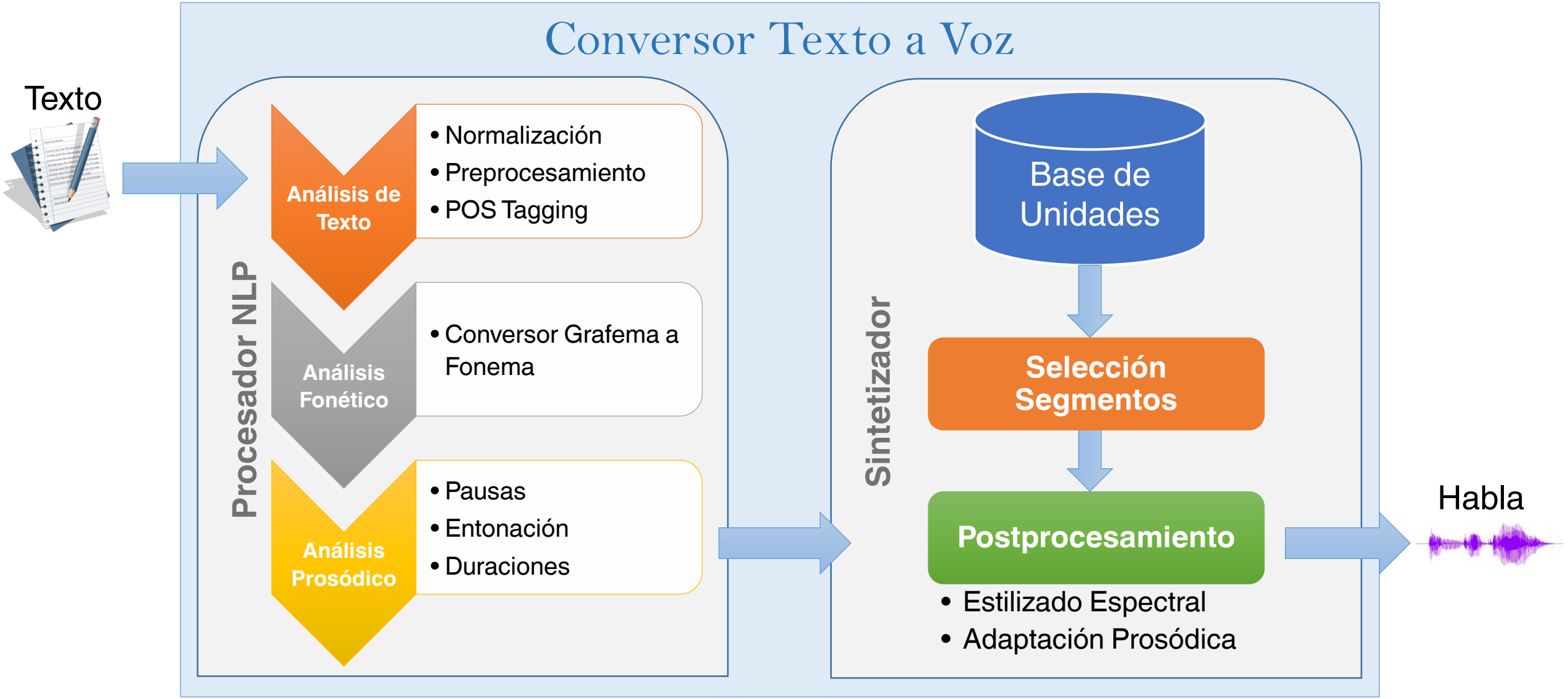
Síntesis del Habla (TTS)

- Generación artificial de habla humana a partir de un texto de entrada, con el objetivo de que suene inteligible y natural.
- Estado de las implementaciones comerciales:
- INTELIGIBILIDAD próxima a la del habla humana.
- NATURALIDAD deficiente debido a una inadecuada predicción de la prosodia, y/o problemas en la concatenación debido a la selección de unidades.

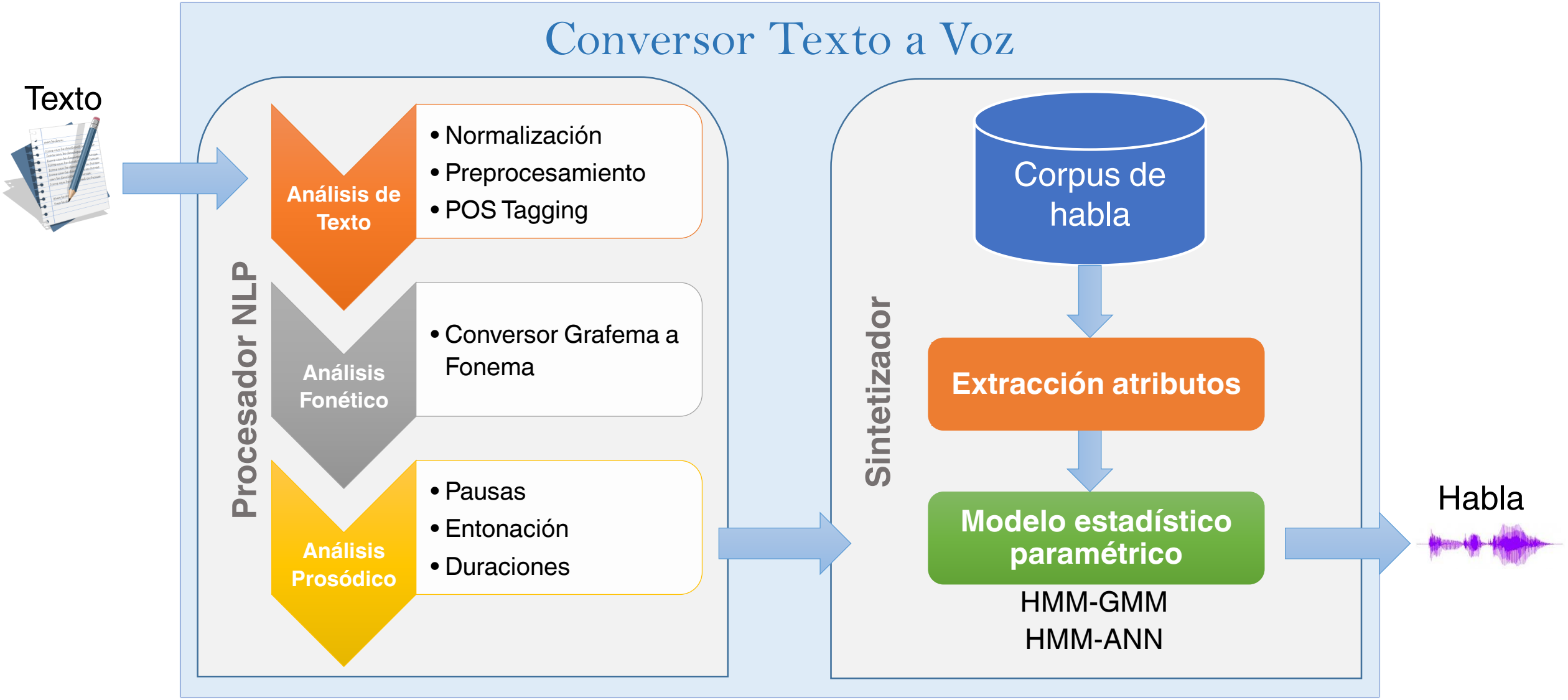
Síntesis del Habla (TTS)

- Actualmente hay dos paradigmas de TTS:
- Síntesis concatenativa: se construye una BD de unidades acústicas, que se seleccionan y concatenan para generar habla.
- Síntesis Paramétrica-Estadística: usa HMMs como modelos generativos para producir habla. Usan modelos acústicos dependientes del contexto.

Síntesis Concatenativa



Síntesis Paramétrica



Procesamiento de Texto

- Normalización del Texto
  - Detecta símbolos ortográficos y resuelve ambigüedades
  - Tokeniza
  - Clasifica palabras según sus rasgos
- Preprocesamiento
  - Expande abreviaturas, deletrea acrónimos, transforma números, separa en sílabas
  - Determina la tonicidad de las palabras
- Analizador Morfo-Sintáctico (POS Tagging)
  - Localiza y etiqueta estructuras sintácticas presentes en las frases de entrada

Síntesis del Habla y Redes Neuronales

- Wavenet (DeepMind 2016) presentan sintetizadores basados en redes neuronales y revolucionan el campo de TTS
- Basados en CNN. Toma una forma de onda de audio digitalizada como entrada, la hace atravesar las capas convolucionales y genera la forma de onda de salida
- Naturalidad e inteligibilidad casi perfectas, procesamiento muy lento
- Tacotron (Google), aprende a sintetizar de pares texto-audio. Toma caracteres del texto como entradas, los pasa a través de diferentes submódulos de redes neuronales y genera el espectrograma del audio como salida

Gracias

http://www.investigacion.frc.utn.edu.ar/cintra/

diegoevin@gmail.com

RAH (Radio Rex 1920)