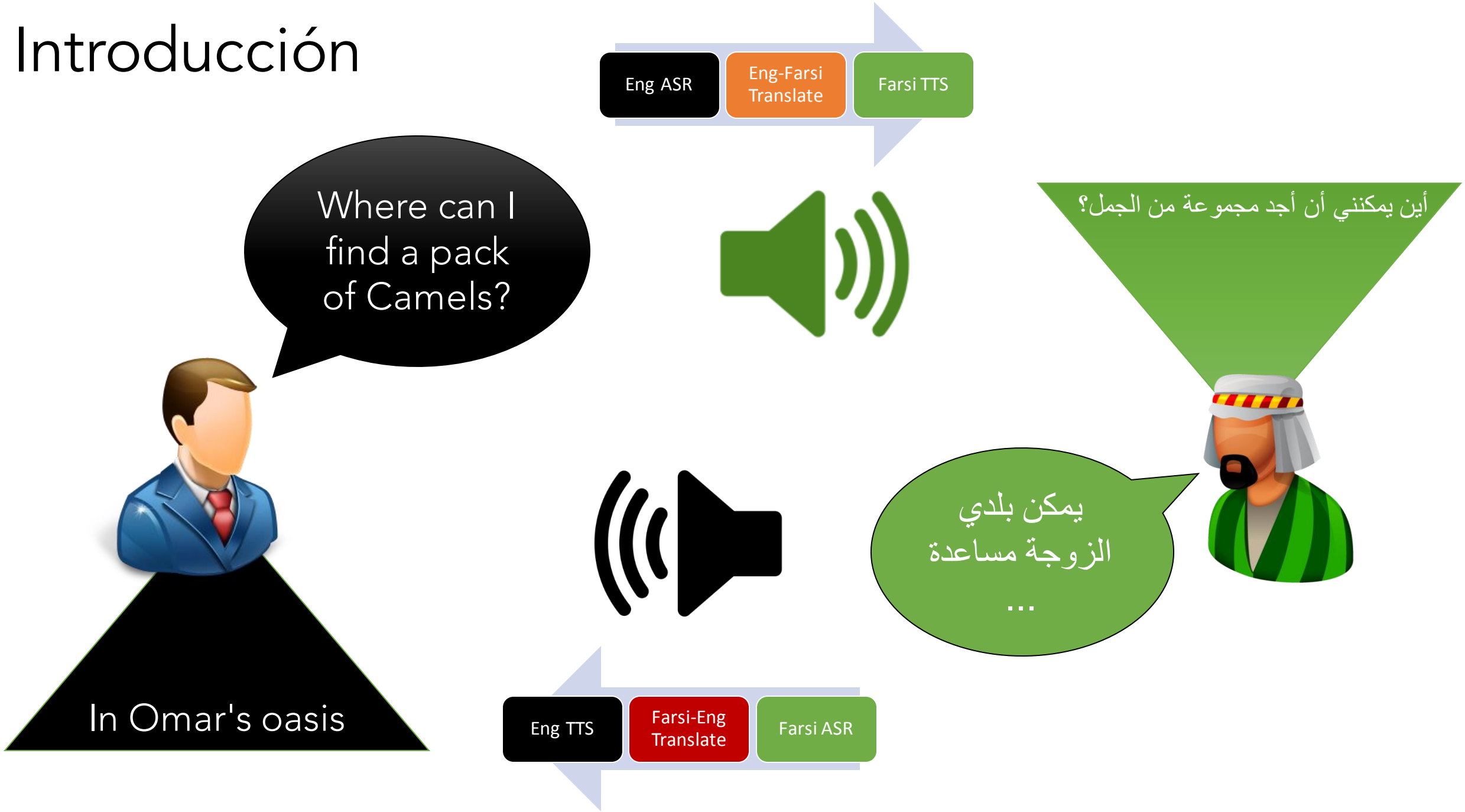
Tecnologías del Habla

Reconocimiento Automático del Habla

Contenido

- Introducción
- Tecnologías del Habla
- Reconocimiento Automático del Habla
- Redes Neuronales y RAH
- Herramientas para RAH

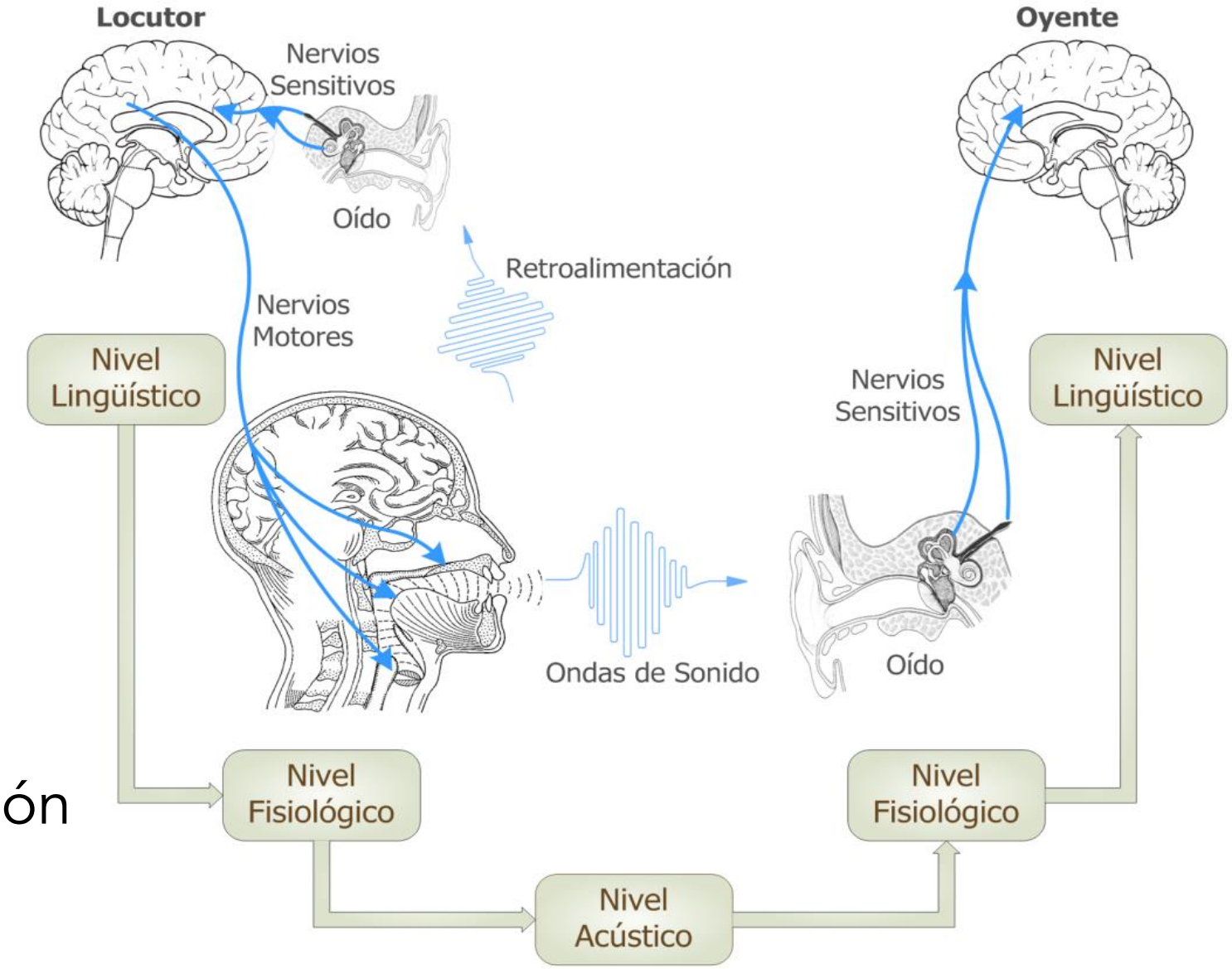
Introducción



Introducción

- Modelo de transmisión de ideas entre humanos

- Contempla Síntesis, Reconocimiento-Comprensión



Cadena del Habla

Tecnologías del Habla

→ Tecnologías con alguna interacción por voz

Áreas 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

HCI

- Agentes inteligentes
- IVR
- 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
- Enseñanza de lectura

Varios

- Traducción habla-habla
- Speech analytics
- Monitoreo de medios

Aplicaciones de Tecnologías del Habla

Reconocimiento Humano del Habla

- Sistema auditivo: registra vibraciones sonoras, las transduce a impulsos eléctricos e interpreta en términos lingüísticos
- Hablante nativo ~ 600/900 sonidos por minuto (24 fonemas esp.)
- Cerebro debe integrar esos sonidos en miles de posibles palabras, sin delimitación de inicio y fin entre cada una, y que esas palabras pueden admitir diferentes pronunciaciones
- Además discriminar voz/ruido, seguir a un hablante específico, …

Reconocimiento Humano del Habla

Fonética	• Manifestaciones físicas de producción y percepción de sonidos de una lengua (fonos)
Fonología	• Modo en que los sonidos de una lengua funcionan a nivel abstracto o mental (fonemas)
Morfología	• Formación las palabras (morfemas)
Prosodia	• Uso de melodía y acentos para dar significado
Sintaxis	• Estructura del lenguaje. Reglas para combinar palabras en oraciones
Semántica &Pragmática	• Significado de las oraciones y reglas conversacionales

Niveles Lingüísticos

Reconocimiento Automático del Habla

Tareas Relacionadas

- Comprensión del Habla (ASU)
- Detección de Habla (SAD)
- Identificación de Hablantes (SID)
- Realce de Habla

Reconocimiento Automático del Habla

- 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

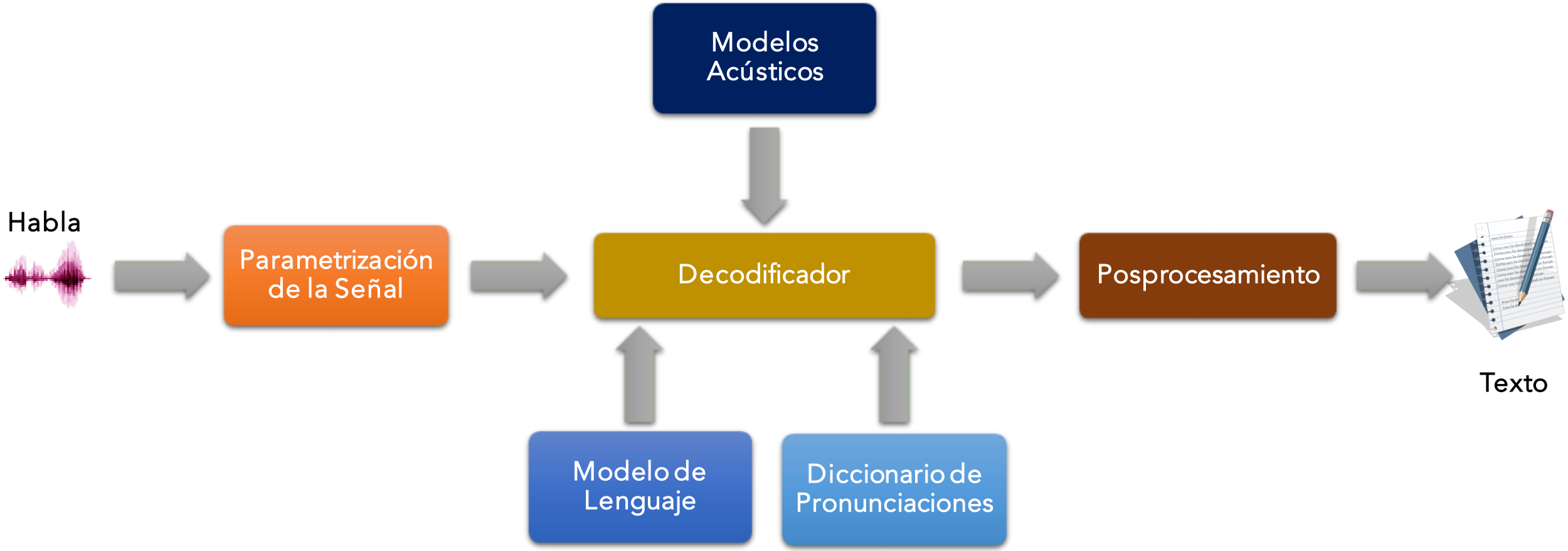
Ruidos e Interferencias

- En sensores
- Canales
- Codecs

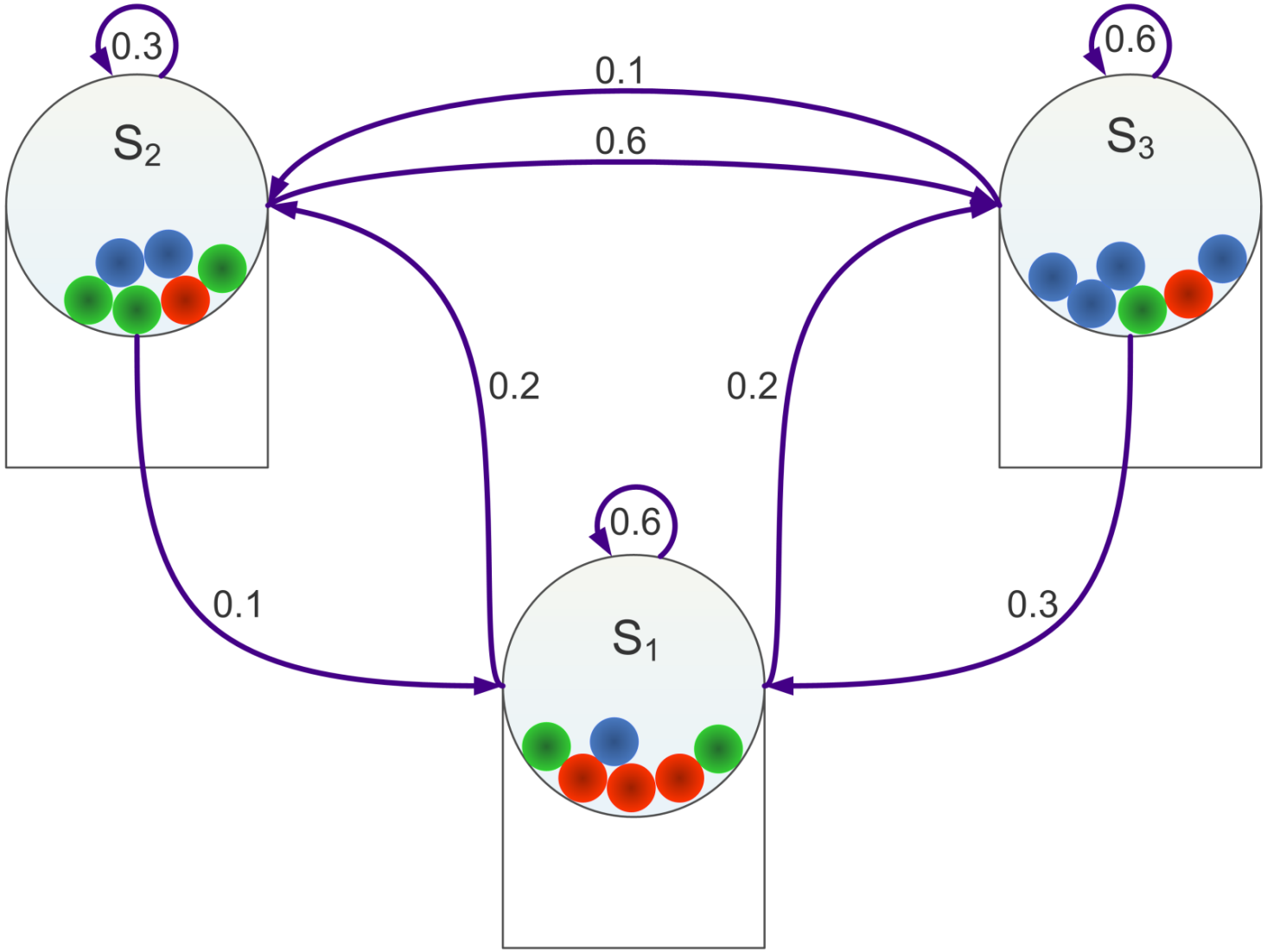
Variaciones en el Medio

Variabilidad de la Señal de Habla

Reconocimiento Automático del Habla

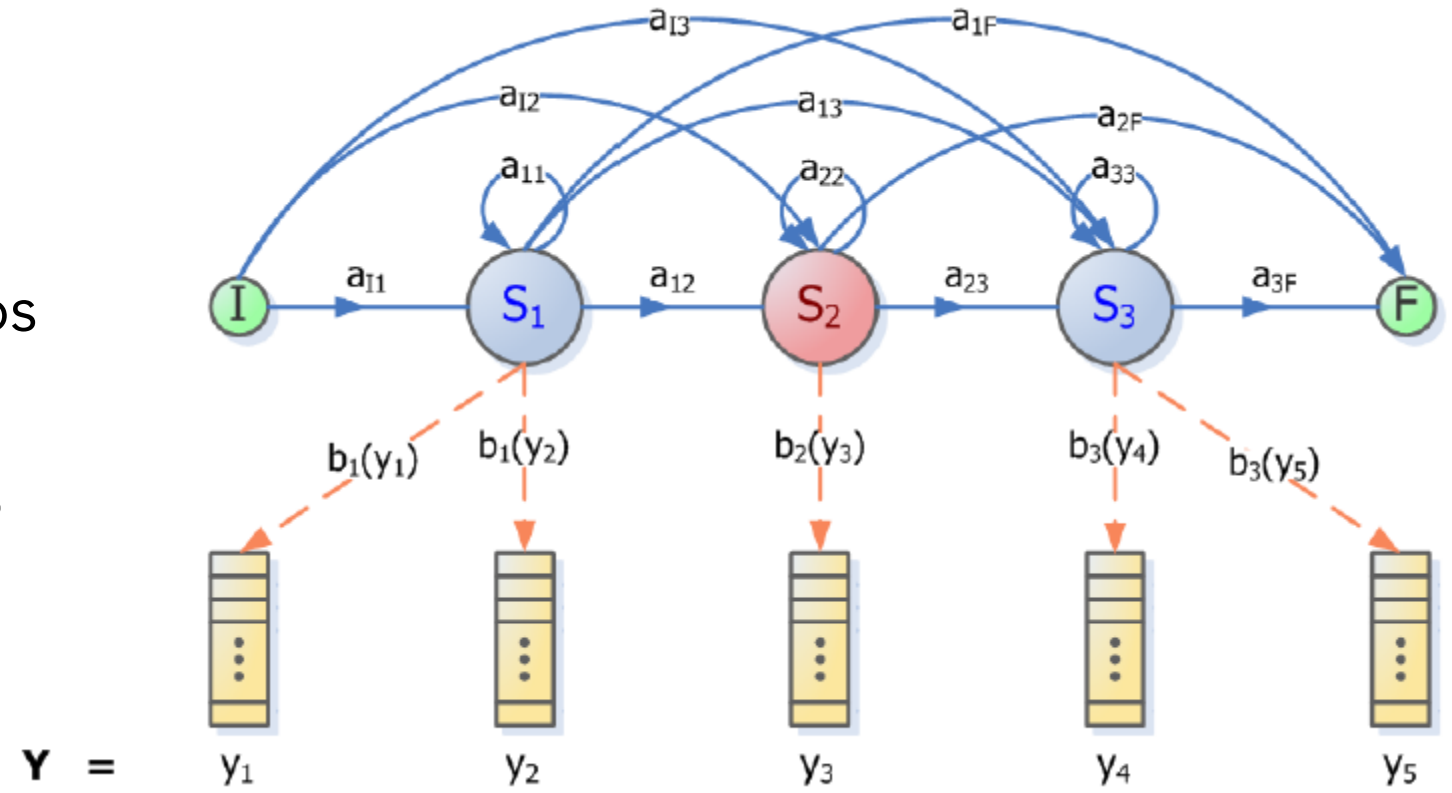


Esquema de un Reconocedor Estándar



MODELOS OCULTOS DE MARKOV

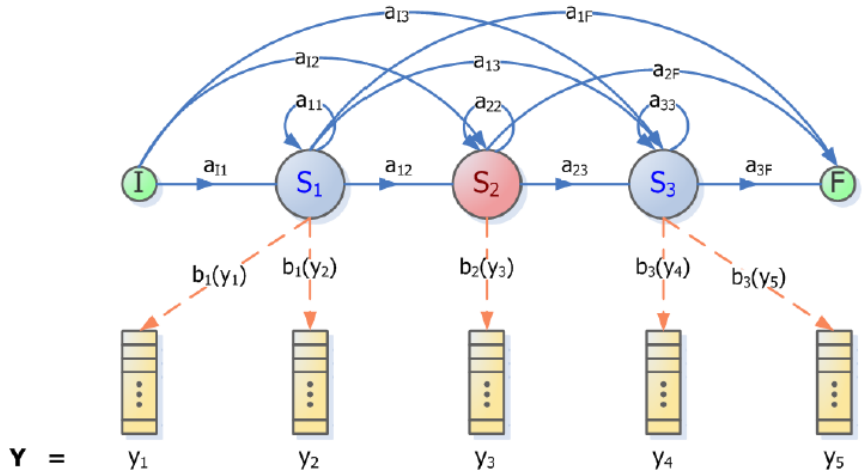
- Capaces de resolver implícitamente segmentación y clasificación de unidades
- Entrenamiento escalable, métodos eficientes para aprendizaje y decodificación, buen desempeño
- Estado del arte por más de 30 años

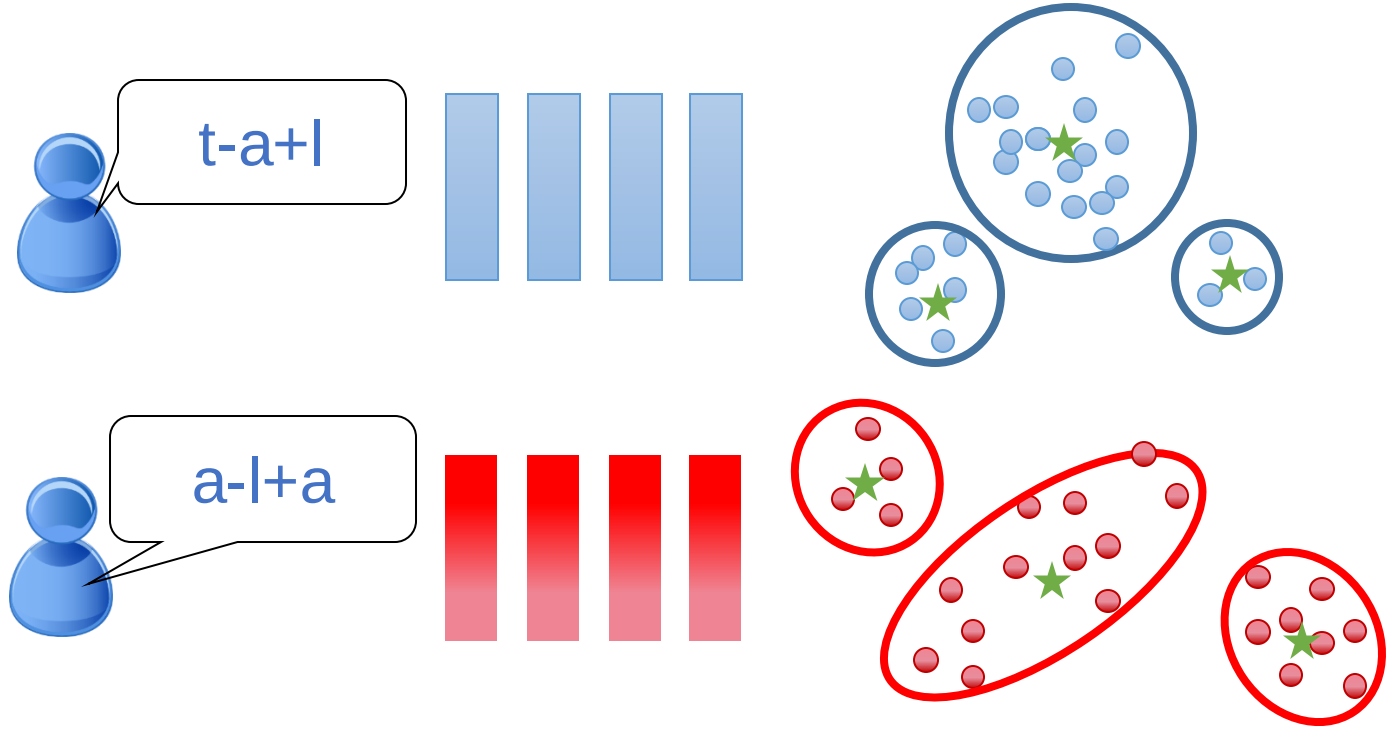


MODELOS OCULTOS DE MARKOV

GMM-HMM

Cada estado tiene una distribución estacionaria de atributos acústicos





Gaussian Mixture Model (GMM)

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

MODELOS OCULTOS DE MARKOV

ASR MODELO BAYESIANO

Sean:

O: secuencia de T atributos acústicos (MFCC, PLP, LPCC)

W: secuencia de N palabras de un vocabulario fijo y conocido

P(W|O): probabilidad condicional de la secuencia de palabras **W**, dada la secuencia observada **O**

Implementación bayesiana: $W = \underset{W}{\operatorname{argmax}}\, P(W|O)$

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

ASR MODELO BAYESIANO

Usando 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**

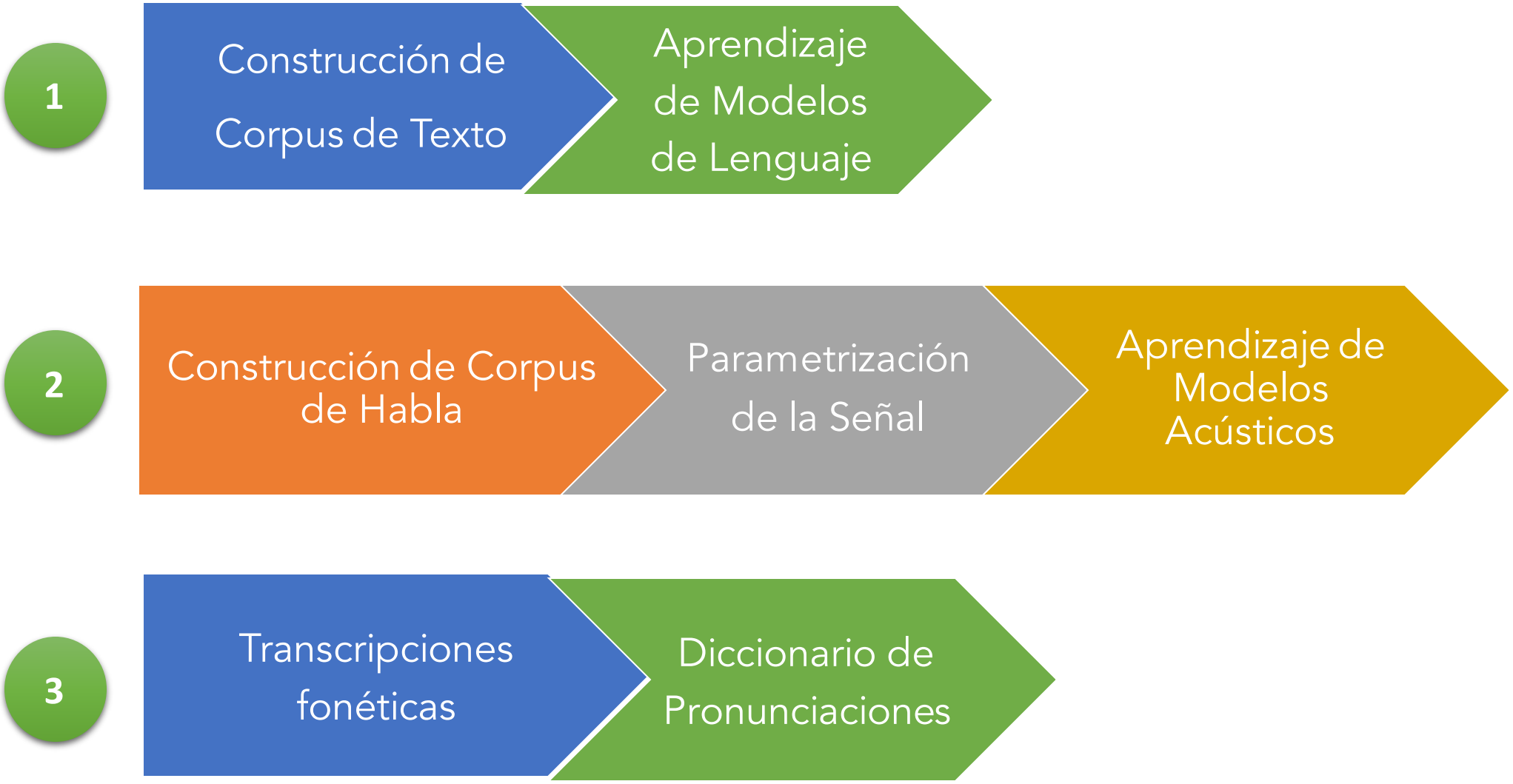
Se puede reescribir una expresión más práctica para el reconocedor:

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

Modelo Acústico

Modelo de Lenguaje

Reconocimiento Automático del Habla

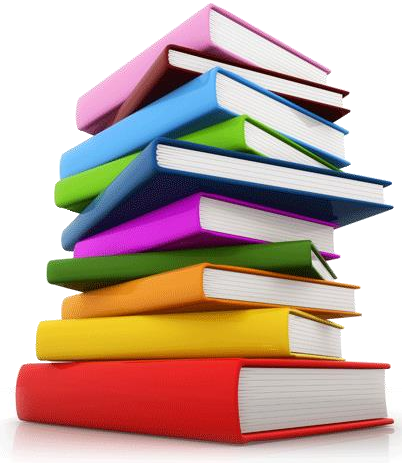


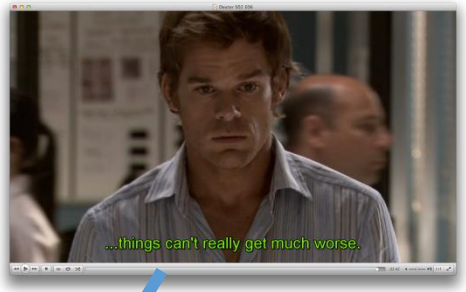
Corpus de Texto

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



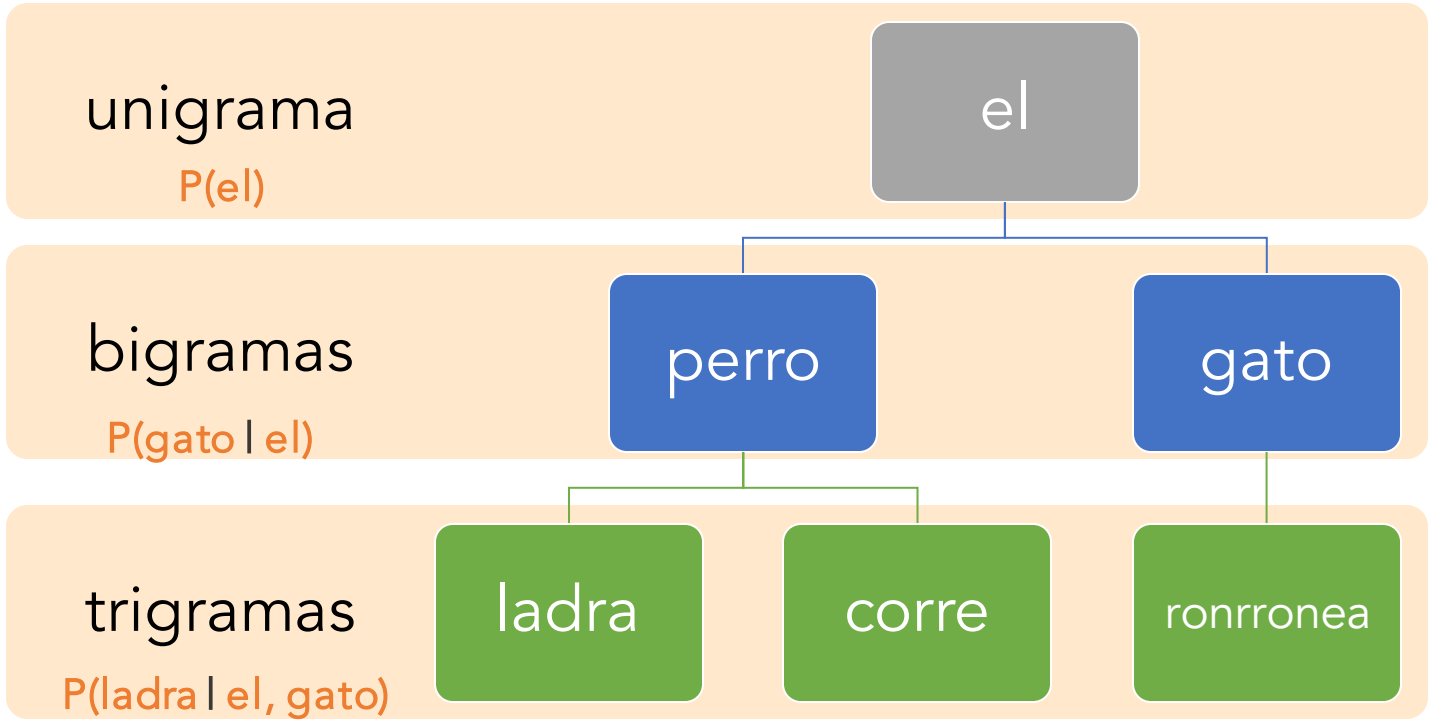


Modelo de Lenguaje

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:

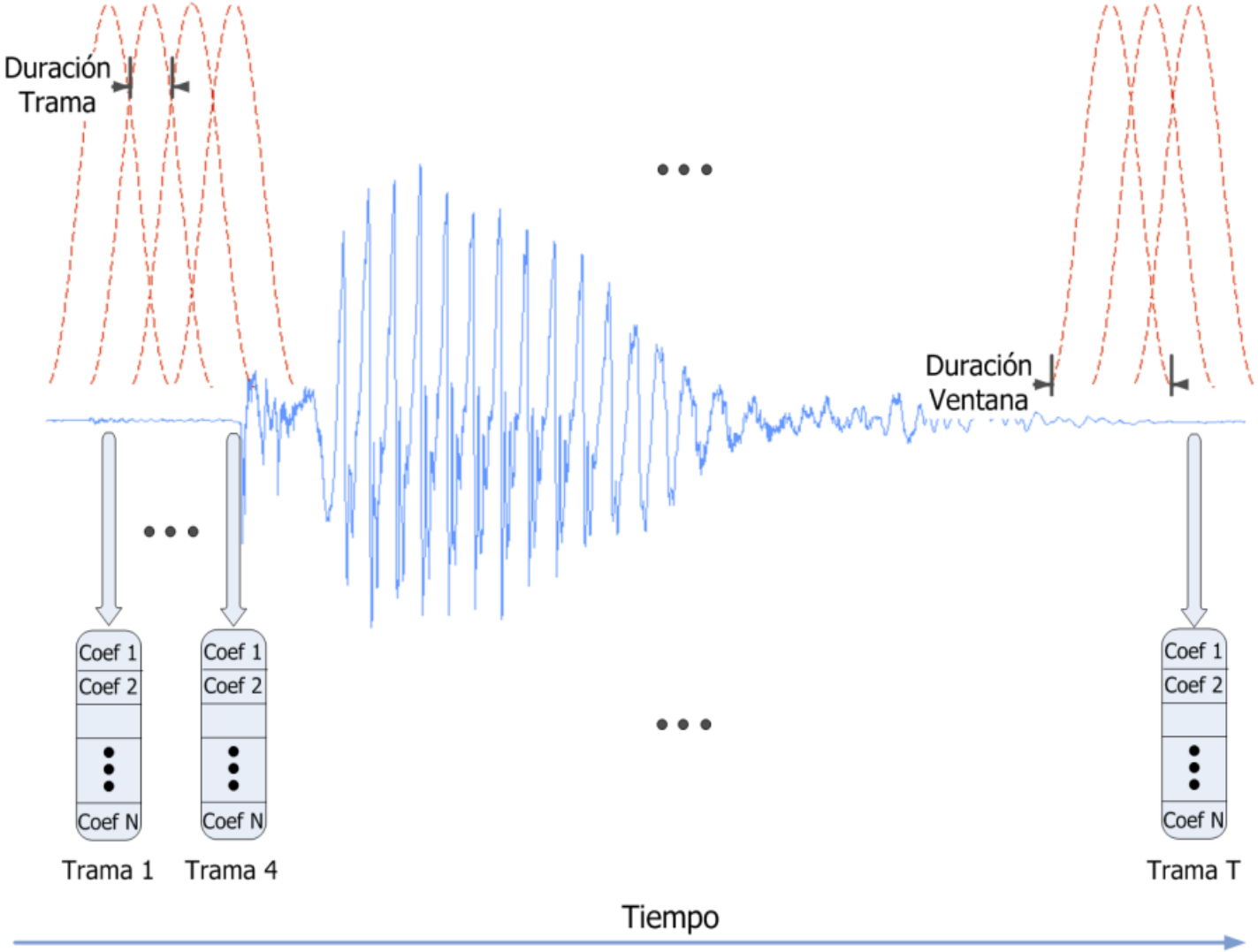
- Balanceadas fonéticamente: buena cobertura de cada combinación de sonidos
- De múltiples hablantes: para aprender un modelo acústico que sirva para todos
- Diversidad prosódica: para aprender cómo varían los sonidos con la entonación
- Diversas pronunciaciones dialectales

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.

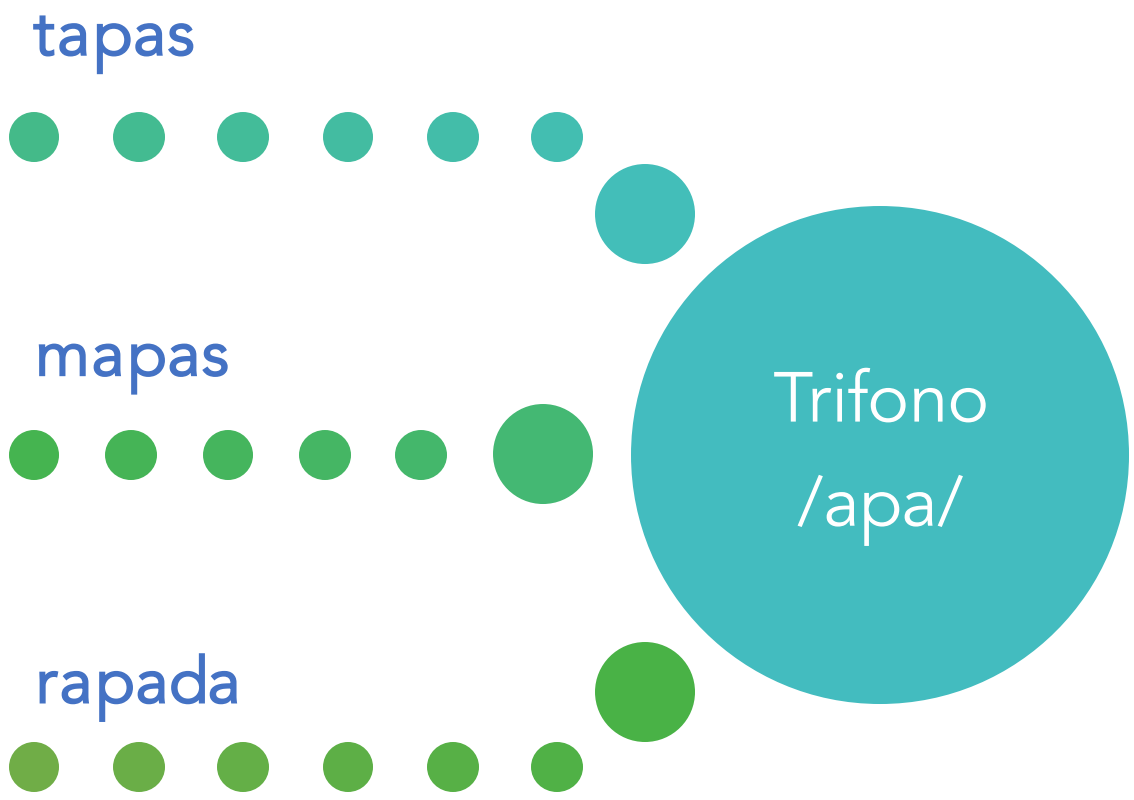


Entrenamiento Modelos Acústicos

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.

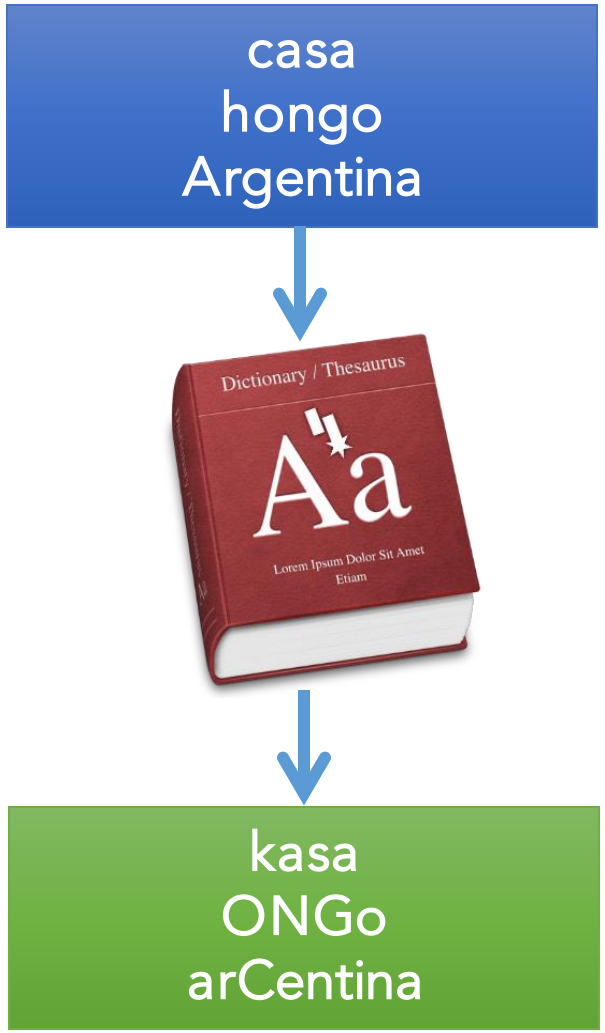


Modelo de Pronunciaciones

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

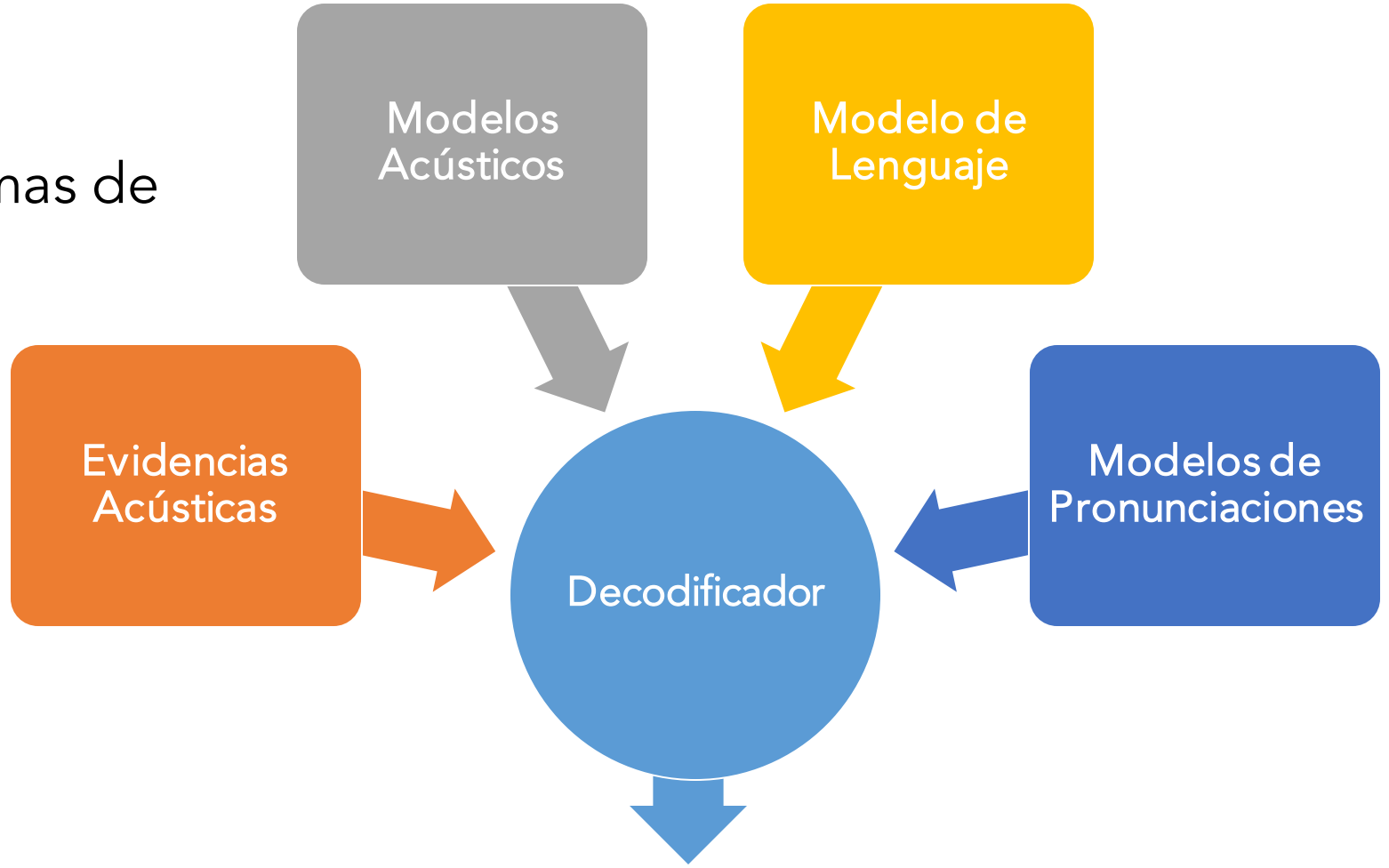


Decodificador

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

- Complejidad

Tipo de Habla	Vocabulario	Conocimiento del Usuario	Condiciones de Uso	Tipo de Aplicación
• Aislada	• Pequeño	• Dependiente	• Laboratorio	• Comando
• Conectada	• Mediano	• Independiente	• Robusto	• Palabras clave
• Continua	• Grande	• Adaptable		• Dictado
• Diálogo				• Close caption

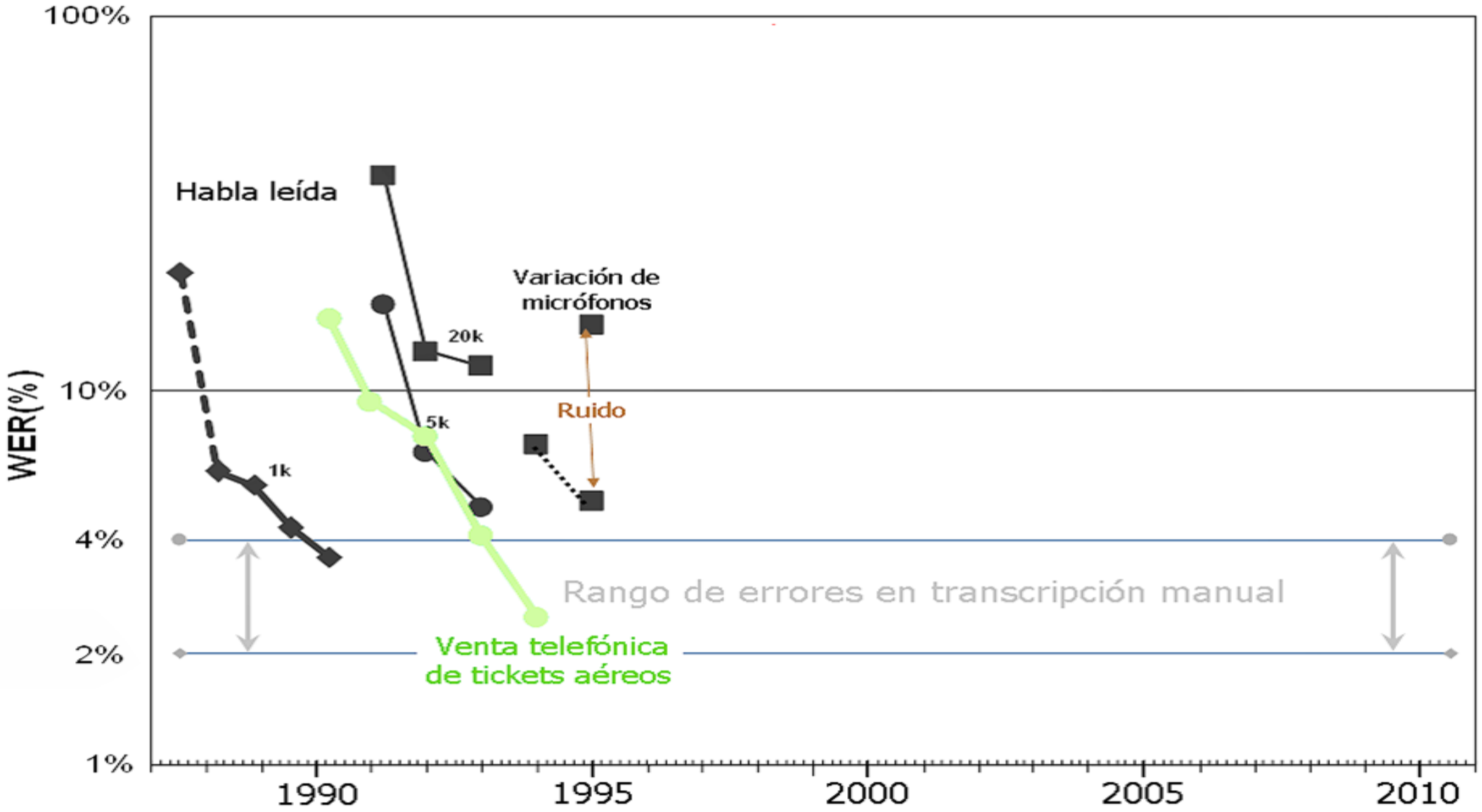
+ Complejidad

DARPA: construcción de Corpus - definición de tareas

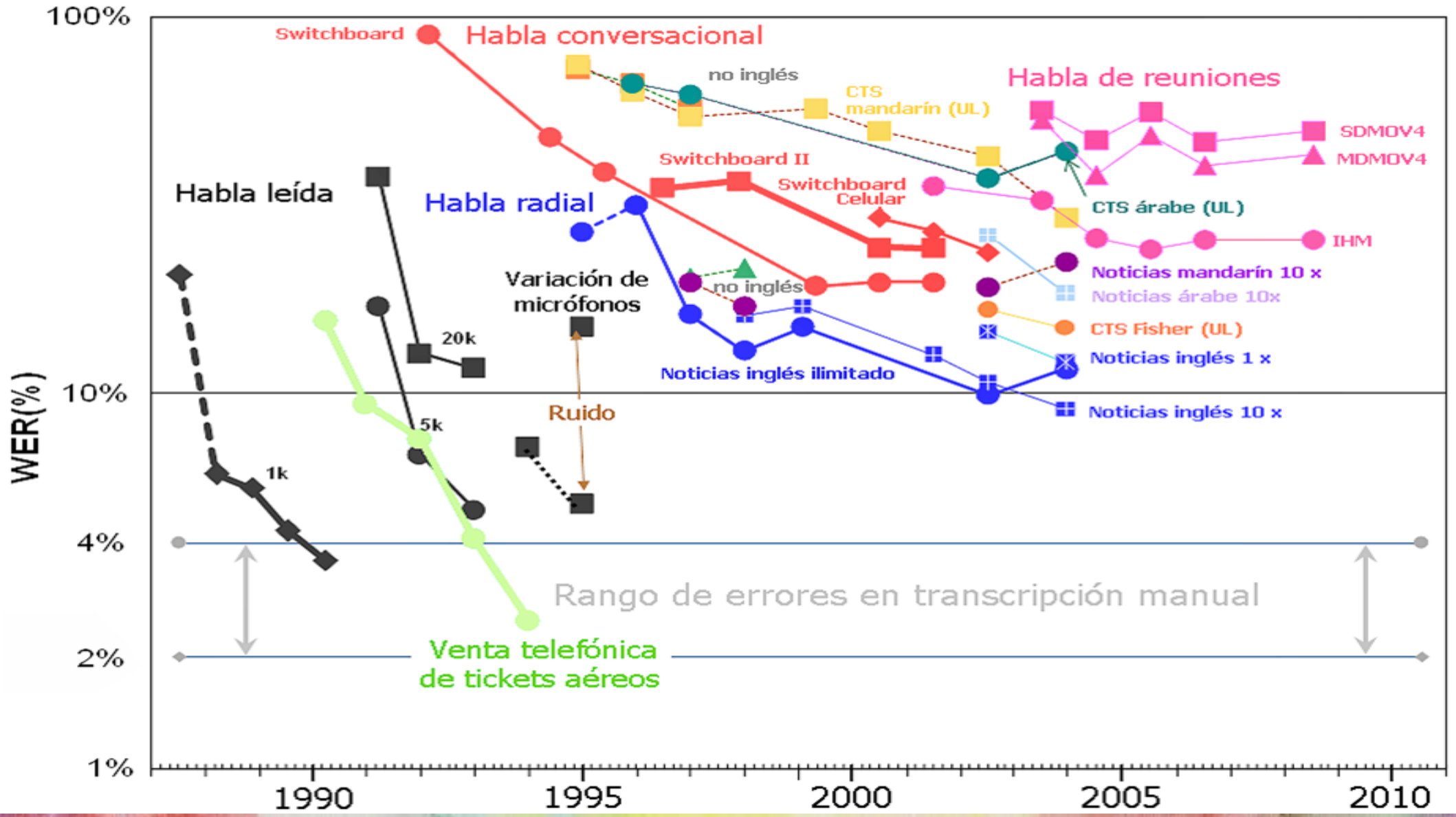
Corpus	Hs.	Lexicón	Locutores	Estilo
ATIS	10,2	< 2000	36	Habla espontánea, dominio restringido
WSJ	73+8	5000 – 20000	?	Leída, continua
TIMIT	5,3	6100	630	Leída, continua
Switchboard	240	>3M	543	Espontánea, telefónica, dominio restringido
Broadcast News	104	>1M	?	Leída, diálogos

Disponibilidad de Datos

Evolución Histórica



Evolución Histórica



Redes Neuronales en RAH

NNs en RAH

- Idea de usar NN para RAH no es nueva
- Dificultades para entrenar redes neuronales con más de 2 capas (Vanishing Gradients)
- Corpus Insuficientes
- 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.

NNs en 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

NNs en 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

NNs en 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**

NNs en RAH

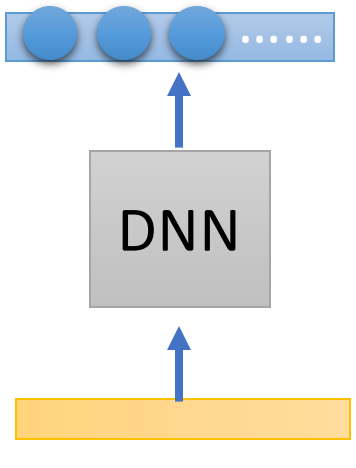
- Con grandes cantidades de datos y regularización adecuada backpropagation funciona bien para redes profundas (**DNN**)
- Pre-entrenamiento de DBN es innecesario
- Introducción de ReLUs y Drop-out
- Sistemas híbridos HMM/DNN pasan a dominar el estado del arte en ASR
- Aprendizaje de banco de filtros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

DNN-HMM

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

$$P(O|W) \approx \max_{s_1 \cdots s_T} \prod_{t=1}^{T} P(s_t|s_{t-1})\, P(x_t|s_t)$$

Usando DNN



$P(s_t|x_t)$

x_t

$$P(x_t|s_t) = \frac{P(x_t, s_t)}{P(s_t)}$$

$$= \frac{P(s_t|x_t)P(x_t)}{P(s_t)}$$

Conteo en datos de entrenamiento

DNN-HMM

- Entrenar una DNN para asociar cada estado (neuronas de salida) con un frame de observaciones acústicas (+ contexto)
- 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

NNS en RAH

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

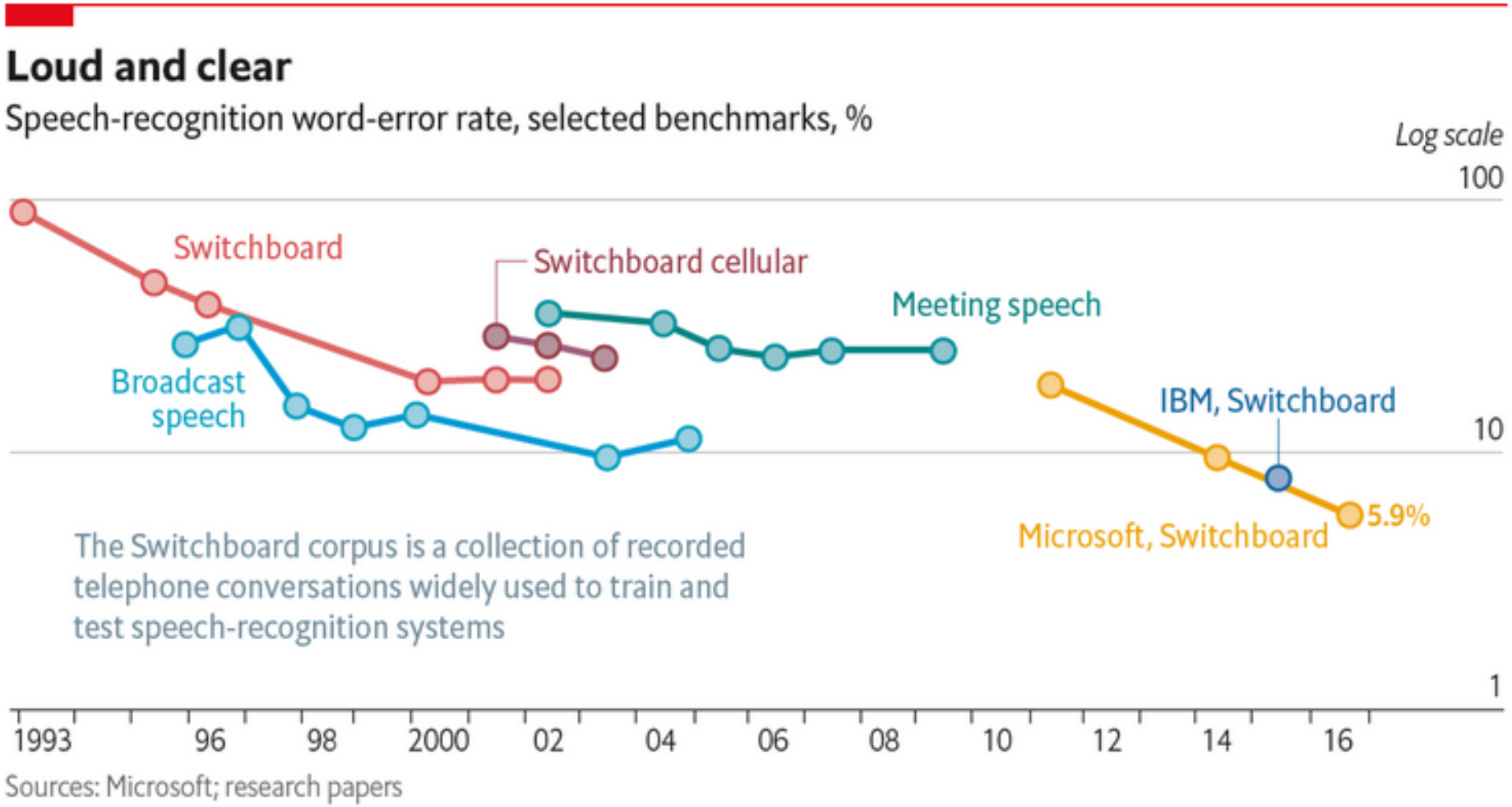
Model	Sentence Error (%)
CD-GMM-HMM ML	39.6
CD-GMM-HMM MMI	37.2
CD-GMM-HMM MPE	36.2
CD-DNN-HMM (5 hidden layers)	30.4

NNs en RAH

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

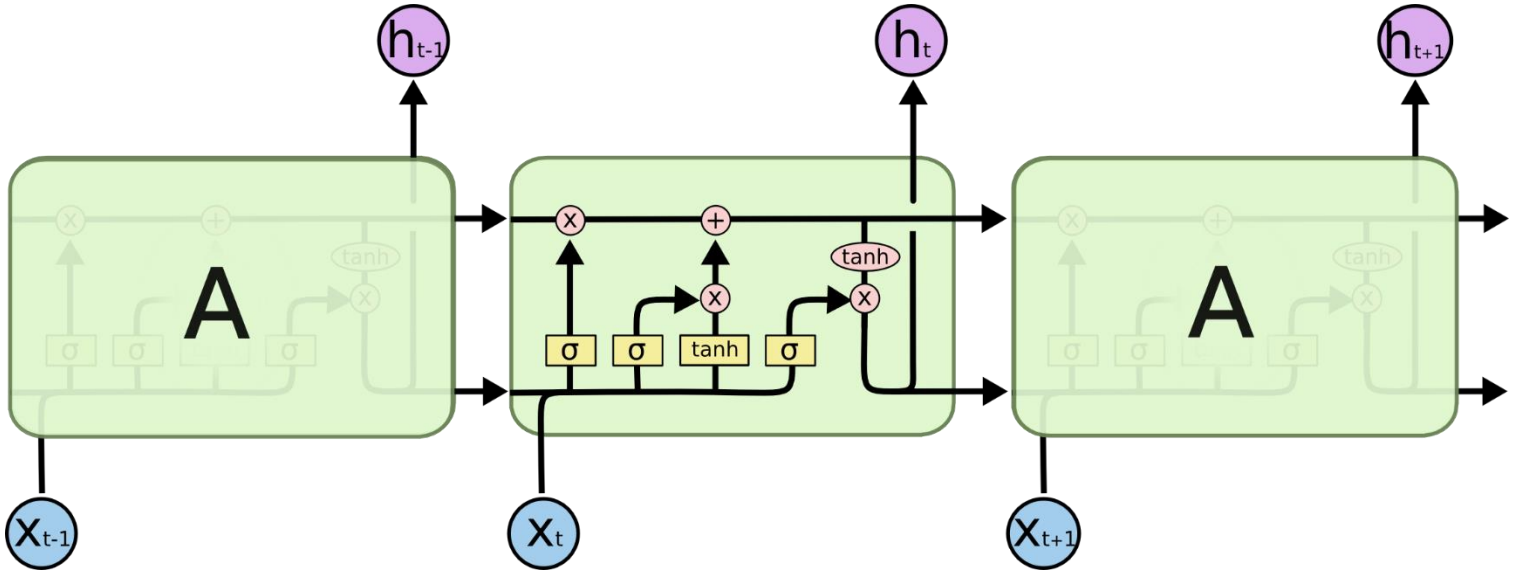
TASK	HOURS OF TRAINING DATA	DNN-HMM	GMM-HMM WITH SAME DATA	GMM-HMM WITH MORE DATA
SWITCHBOARD (TEST SET 1)	309	18.5	27.4	18.6 (2,000 H)
SWITCHBOARD (TEST SET 2)	309	16.1	23.6	17.1 (2,000 H)
ENGLISH BROADCAST NEWS	50	17.5	18.8	
BING VOICE SEARCH (SENTENCE ERROR RATES)	24	30.4	36.2	
GOOGLE VOICE INPUT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

Avances en RAH



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



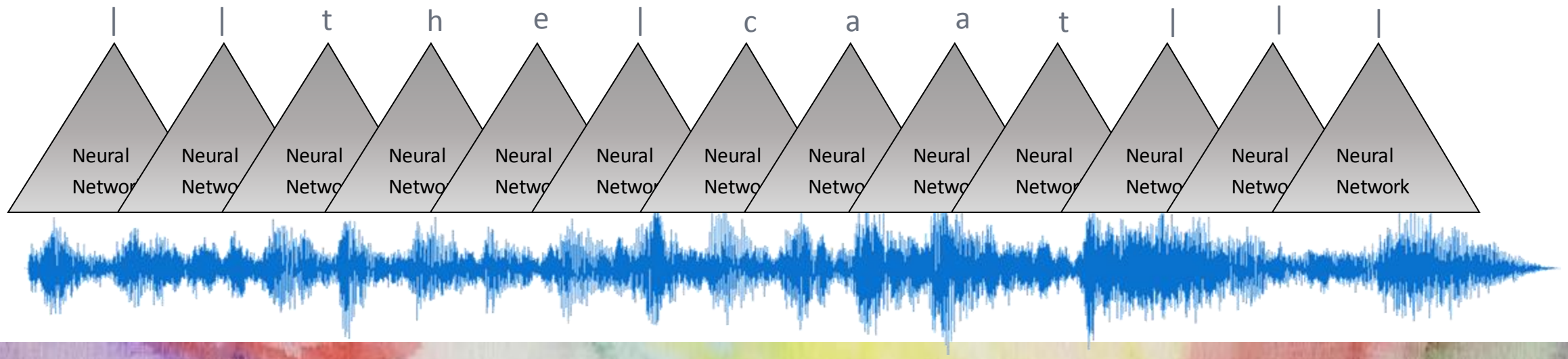
RAH Completamente Conexionista (End to End ASR)

|	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

Modelo de duraciones

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



Herramientas para RAH

Principales Soluciones Comerciales de ASR

Google (Now)
Microsoft (Cortana)
Amazon (Alexa)
Apple (Siri)
Baidu
Braina
Nuance Dragon
e-Speaking
Fusion SpeechEMR
Hound
LumenVox
One Voice Data
SmartAction Speech IVR System
Tatzi
ViaTalk
Voice Finger

Herramientas para RAH

Principales Herramientas de desarrollo para RAH	CMU Sphinx
	HTK
	Julius
	Kaldi
	RWTH ASR
	wav2letter++

Gracias

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

diegoevin@gmail.com