



Reconocimiento del habla en Python

Javier Jorge Cano - @javierjorgecano - PyConES 2019



¿Qué aprenderás en esta charla?

- Conocer la tarea del reconocimiento del habla.

¿Qué aprenderás en esta charla?

- Conocer la tarea del reconocimiento del habla.
- Conocer los desafíos que plantea este problema.

¿Qué aprenderás en esta charla?

- Conocer la tarea del reconocimiento del habla.
- Conocer los desafíos que plantea este problema.
- Conocer las partes y los conceptos intuitivos de un sistema de reconocimiento del habla.

¿Qué aprenderás en esta charla?

- Conocer la tarea del reconocimiento del habla.
- Conocer los desafíos que plantea este problema.
- Conocer las partes y los conceptos intuitivos de un sistema de reconocimiento del habla.
- Conocer herramientas para poder desarrollar un sistema de reconocimiento del habla con Python y recursos abiertos.

Sobre mi:

- Javier Jorge Cano (jjorge@dsic.upv.es).
- Estudiante de Doctorado @ UPV.
- Miembro del grupo *Machine Learning and Language Processing* (MLLP).
- Transcripción, traducción y síntesis del habla.
- *transLectures-UPV Platform* (TLP).

The screenshot shows a video player on the left and a transcription interface on the right. The video player displays a man speaking and text overlays: "aromatic grape varieties, is produced". The transcription interface shows the following text blocks:

Text (French)	Text (English)	Rate (cps)
qui est bien compensée par l'acidité. Les arômes en rétro-nasal se combinent bien avec l'acidité donnant au vin assez de longueur.	which is balanced by its acidity. Along the retro-nasal pathway, the aromas combine well with the acidity of the wine, giving it good length.	14.4 cps
Ce Muscat d'Alsace dont les notes fruitées sont caractéristiques des cépages dits aromatiques est obtenu	This Muscat from Alsace, with fruity notes that are characteristic of aromatic grape varieties, is produced	18.0 cps
par un élevage dans des contenants neutres	by ageing in neutral containers,	12.4 cps
de type cuves en inox. Le choix de ce contenant	like stainless steel vats. Choosing this type of container	13.9 cps
assure ainsi la préservation de cette composante aromatique en limitant les phénomènes d'oxydation	ensures the preservation of this aromatic component by limiting oxidation phenomena during this	13.6 cps
		15.7 cps
		13.3 cps
		10.8 cps
		17.9 cps
		13.7 cps

Reconocimiento del habla

Sistema de reconocimiento del habla

Preámbulo

Extracción de características

Entrenamiento

Reconocimiento

Conclusiones

Reconocimiento del habla

Sistema de reconocimiento del habla

Preámbulo

Extracción de características

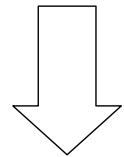
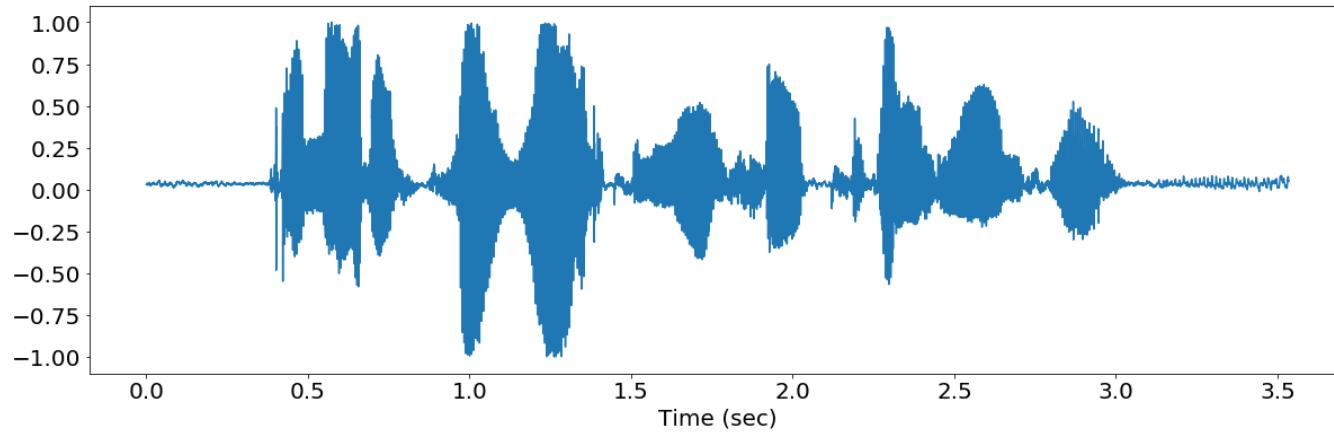
Entrenamiento

Reconocimiento

Conclusiones

Reconocimiento del habla

¿Qué es el reconocimiento del habla?
Automatic **S**peech **R**ecognition (ASR)



“I noticed how white and well shaped his own hands were”

Reconocimiento del habla

- Aplicaciones:

Reconocimiento del habla

- Aplicaciones:
 - Uso de comandos de voz y control.

Reconocimiento del habla

- Aplicaciones:
 - Uso de comandos de voz y control.
 - Dictado a texto.

Reconocimiento del habla

- Aplicaciones:
 - Uso de comandos de voz y control.
 - Dictado a texto.
 - Transcripción de contenido audiovisual.

Reconocimiento del habla

- Aplicaciones:
 - Uso de comandos de voz y control.
 - Dictado a texto.
 - Transcripción de contenido audiovisual.
 - ...

Reconocimiento del habla

- Aplicaciones:
 - Uso de comandos de voz y control.
 - Dictado a texto.
 - Transcripción de contenido audiovisual.
 - ...
- Tareas relacionadas:

Reconocimiento del habla

- Aplicaciones:
 - Uso de comandos de voz y control.
 - Dictado a texto.
 - Transcripción de contenido audiovisual.
 - ...
- Tareas relacionadas:
 - Traducción del habla.

Reconocimiento del habla

- Aplicaciones:
 - Uso de comandos de voz y control.
 - Dictado a texto.
 - Transcripción de contenido audiovisual.
 - ...
- Tareas relacionadas:
 - Traducción del habla.
 - Sistemas de diálogo.

Reconocimiento del habla

- Aplicaciones:
 - Uso de comandos de voz y control.
 - Dictado a texto.
 - Transcripción de contenido audiovisual.
 - ...
- Tareas relacionadas:
 - Traducción del habla.
 - Sistemas de diálogo.
 - Reconocimiento del interlocutor.

Reconocimiento del habla

- Aplicaciones:
 - Uso de comandos de voz y control.
 - Dictado a texto.
 - Transcripción de contenido audiovisual.
 - ...
- Tareas relacionadas:
 - Traducción del habla.
 - Sistemas de diálogo.
 - Reconocimiento del interlocutor.
 - Diarización.

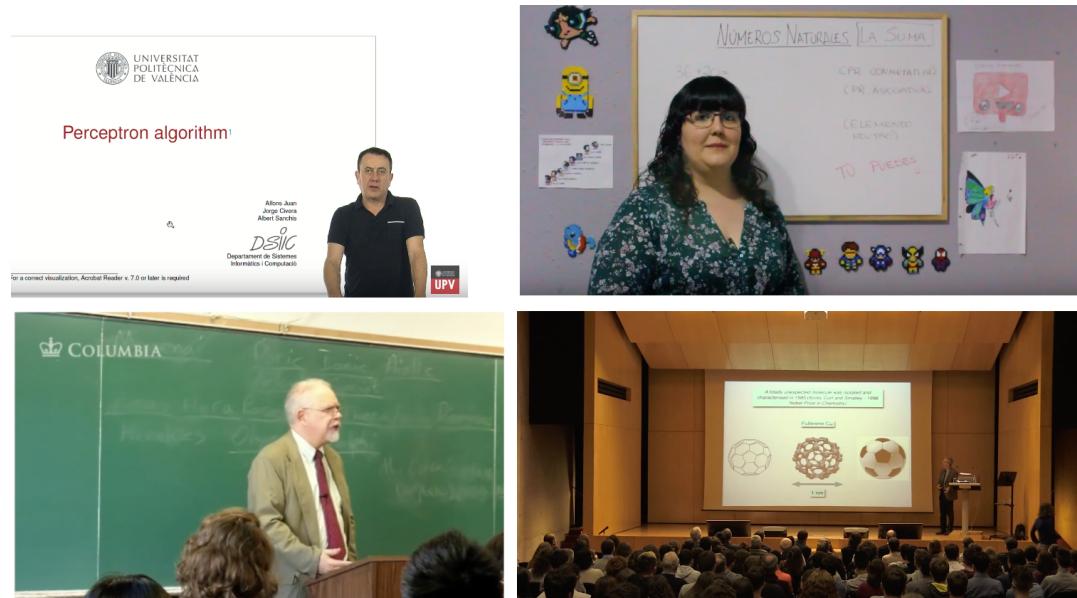
Reconocimiento del habla

- Aplicaciones:
 - Uso de comandos de voz y control.
 - Dictado a texto.
 - Transcripción de contenido audiovisual.
 - ...
- Tareas relacionadas:
 - Traducción del habla.
 - Sistemas de diálogo.
 - Reconocimiento del interlocutor.
 - Diarización.
 - ...

Reconocimiento del habla

¿Por qué es complicada esta tarea?

- Alta variabilidad.
- Tiempo de respuesta.
- Escasez de datos.



Reconocimiento del habla

Sistema de reconocimiento del habla

Preámbulo

Extracción de características

Entrenamiento

Reconocimiento

Conclusiones

Sistema de reconocimiento del habla

Conjuntos de datos

- **Speech Commands**¹
- **LibriSpeech**²

Software

- **TLK**: *TransLectures Toolkit* & **PyTLK**, ambos desarrollados en el grupo MLLP.
- Alternativas: Kaldi³ & PyKaldi⁴

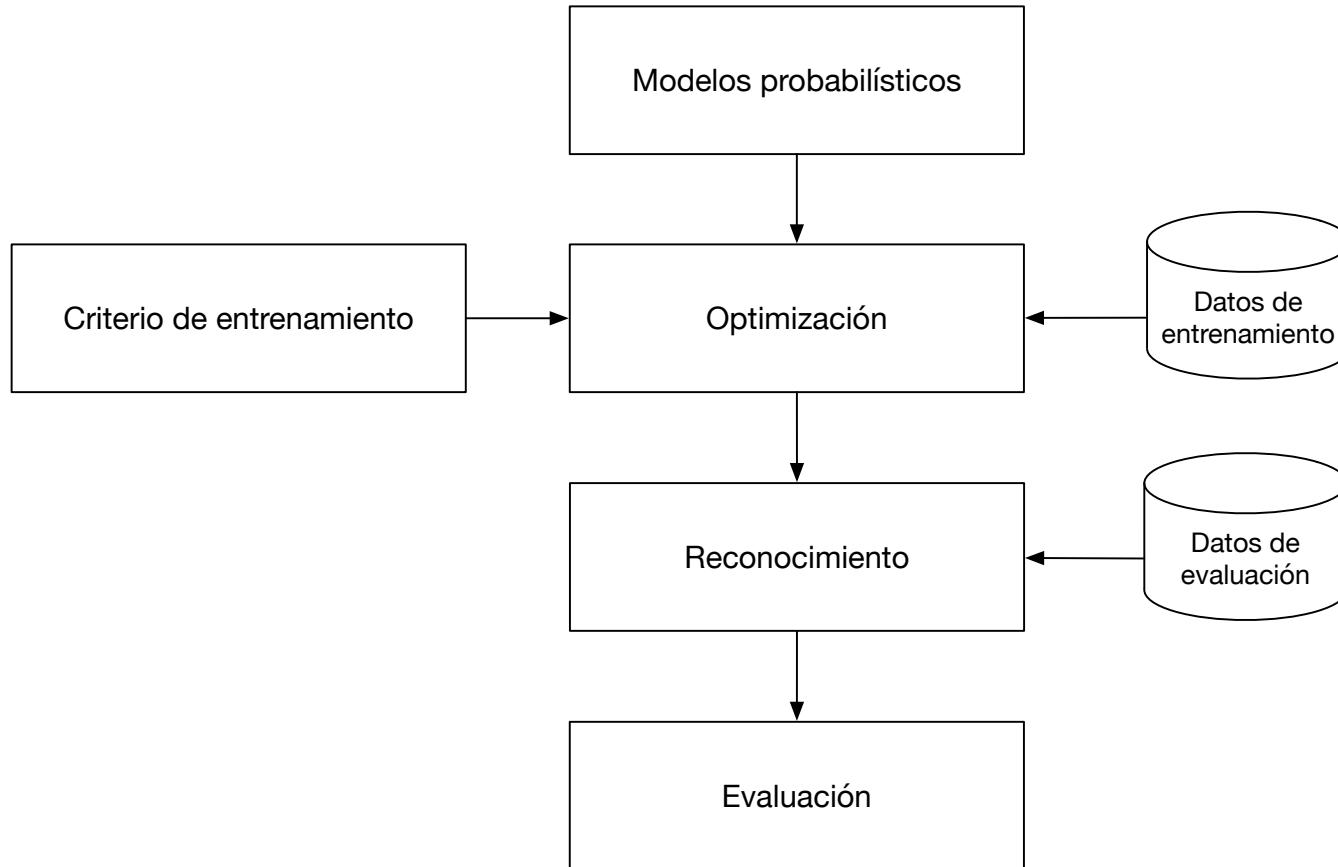
¹http://download.tensorflow.org/data/speech_commands_v0.01.tar.gz

²<http://www.openslr.org/12>

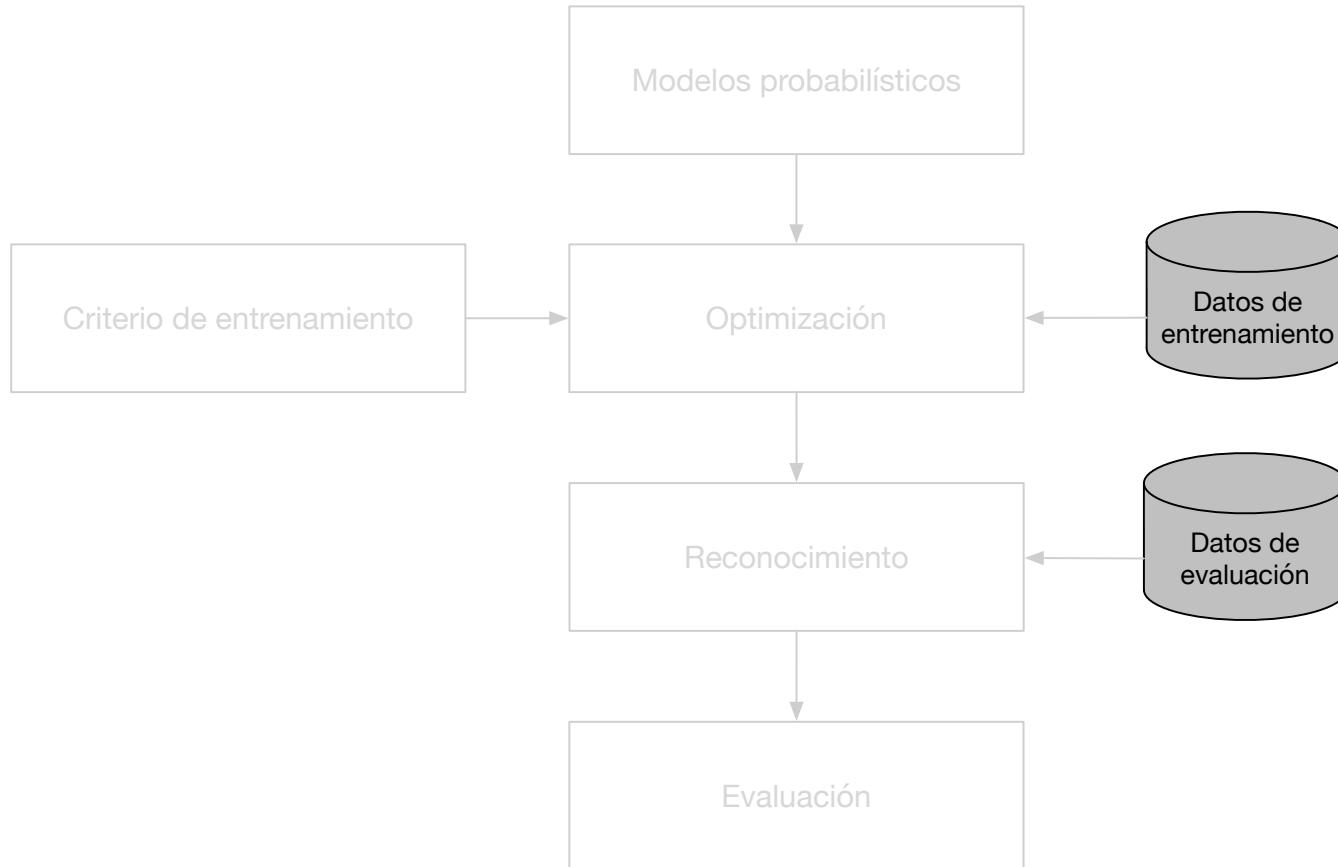
³<https://kaldi-asr.org>

⁴<https://github.com/pykaldi/pykaldi>

Sistema de reconocimiento del habla

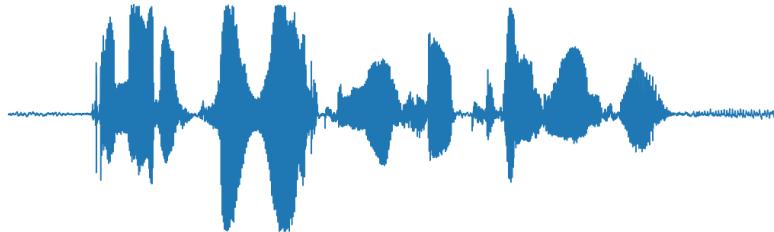


Sistema de reconocimiento del habla



Sistema de reconocimiento del habla

Extracción de características



Sistema de reconocimiento del habla

Extracción de características



“I noticed how white and well shaped his own hands were”

Sistema de reconocimiento del habla

Extracción de características



“I noticed how white and well shaped his own hands were”

I	ay
Noticed	n əw t əh s
How	hə aw
White	w əy t
And	əh n d
Well	w eh l
Shaped	sh ey p t
His	hh ih z
Own	ow n
Hands	hh ae n d z
Were	w er

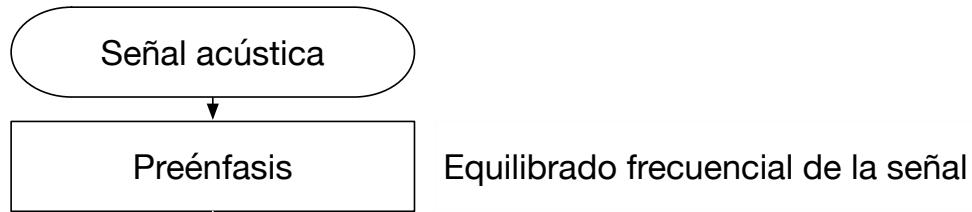
Sistema de reconocimiento del habla

Extracción de características

Señal acústica

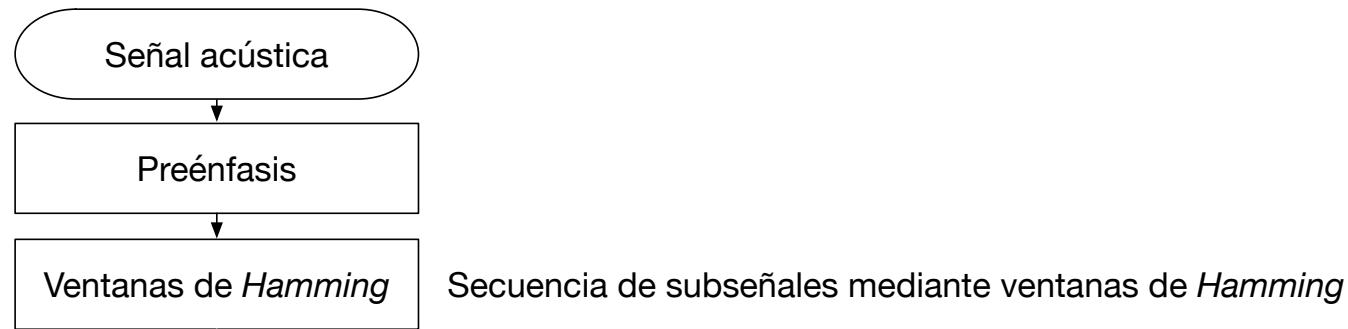
Sistema de reconocimiento del habla

Extracción de características



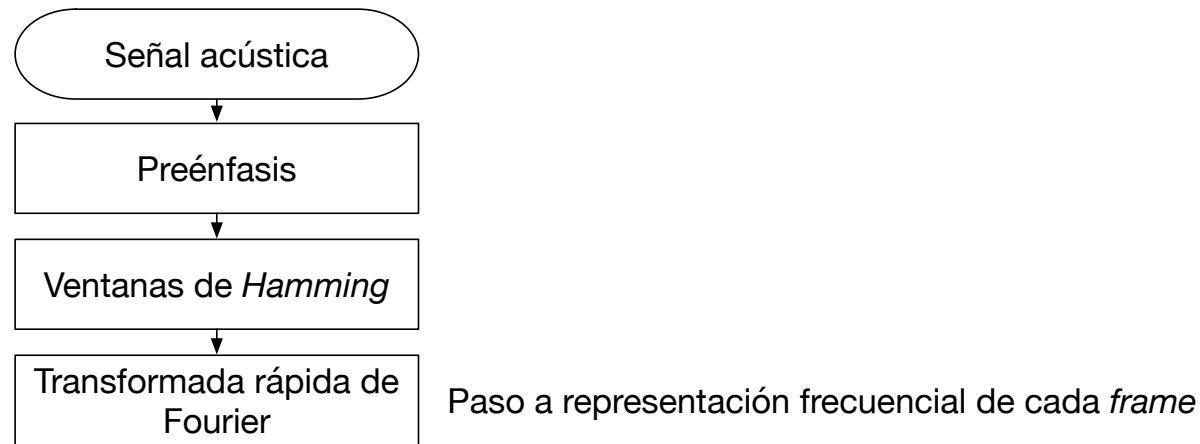
Sistema de reconocimiento del habla

Extracción de características



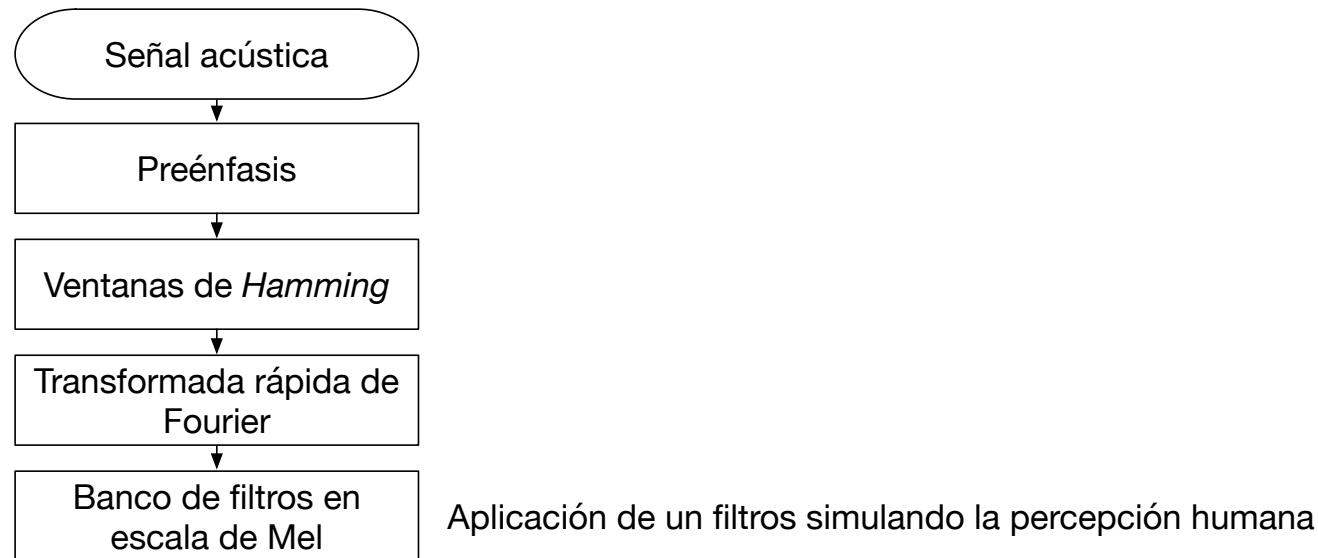
Sistema de reconocimiento del habla

Extracción de características



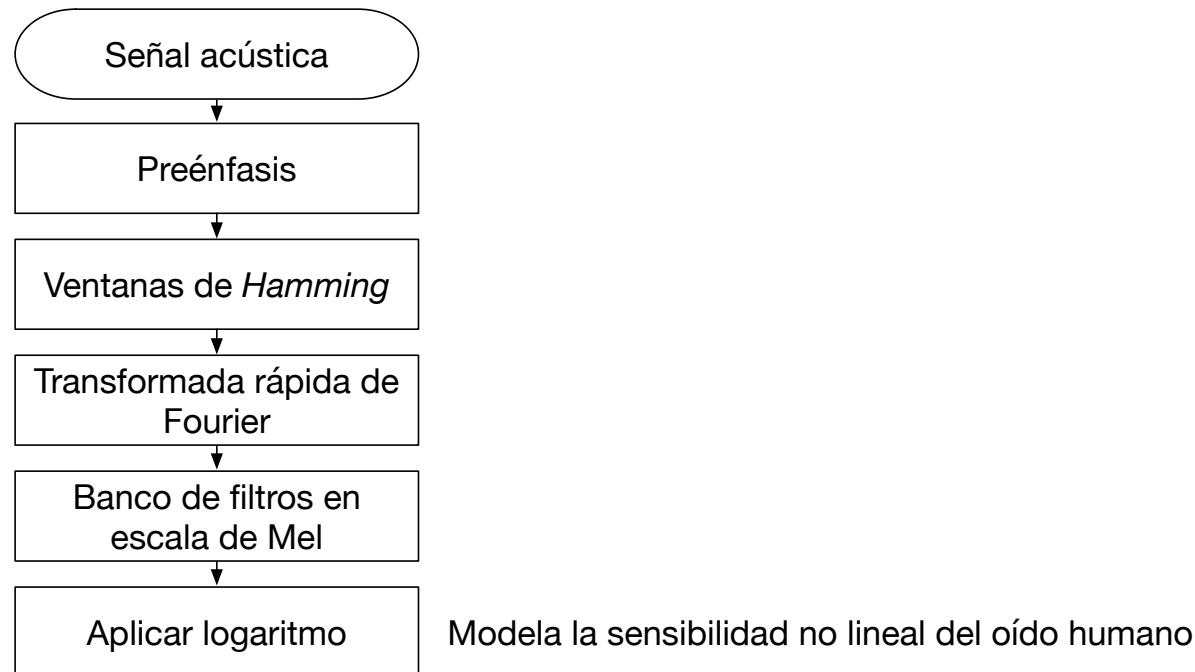
Sistema de reconocimiento del habla

Extracción de características



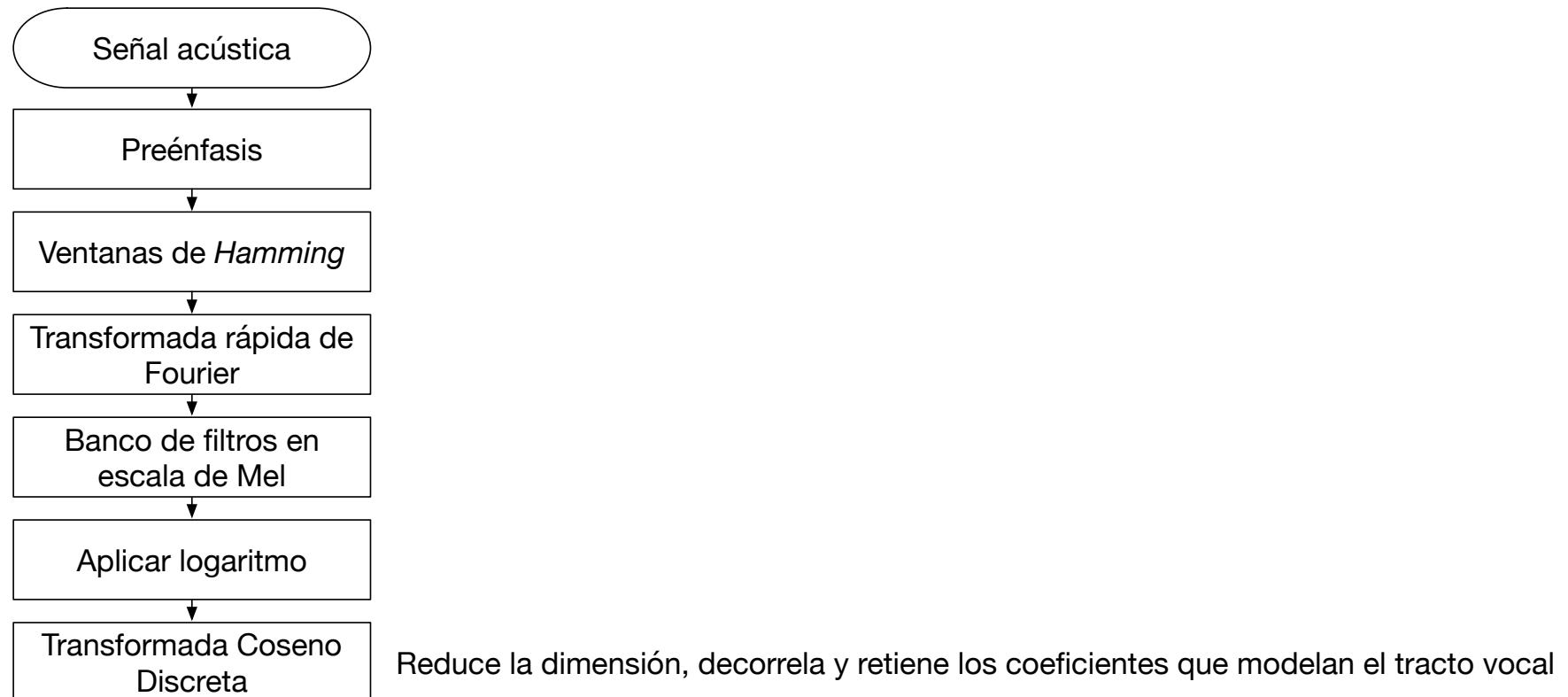
Sistema de reconocimiento del habla

Extracción de características



Sistema de reconocimiento del habla

Extracción de características



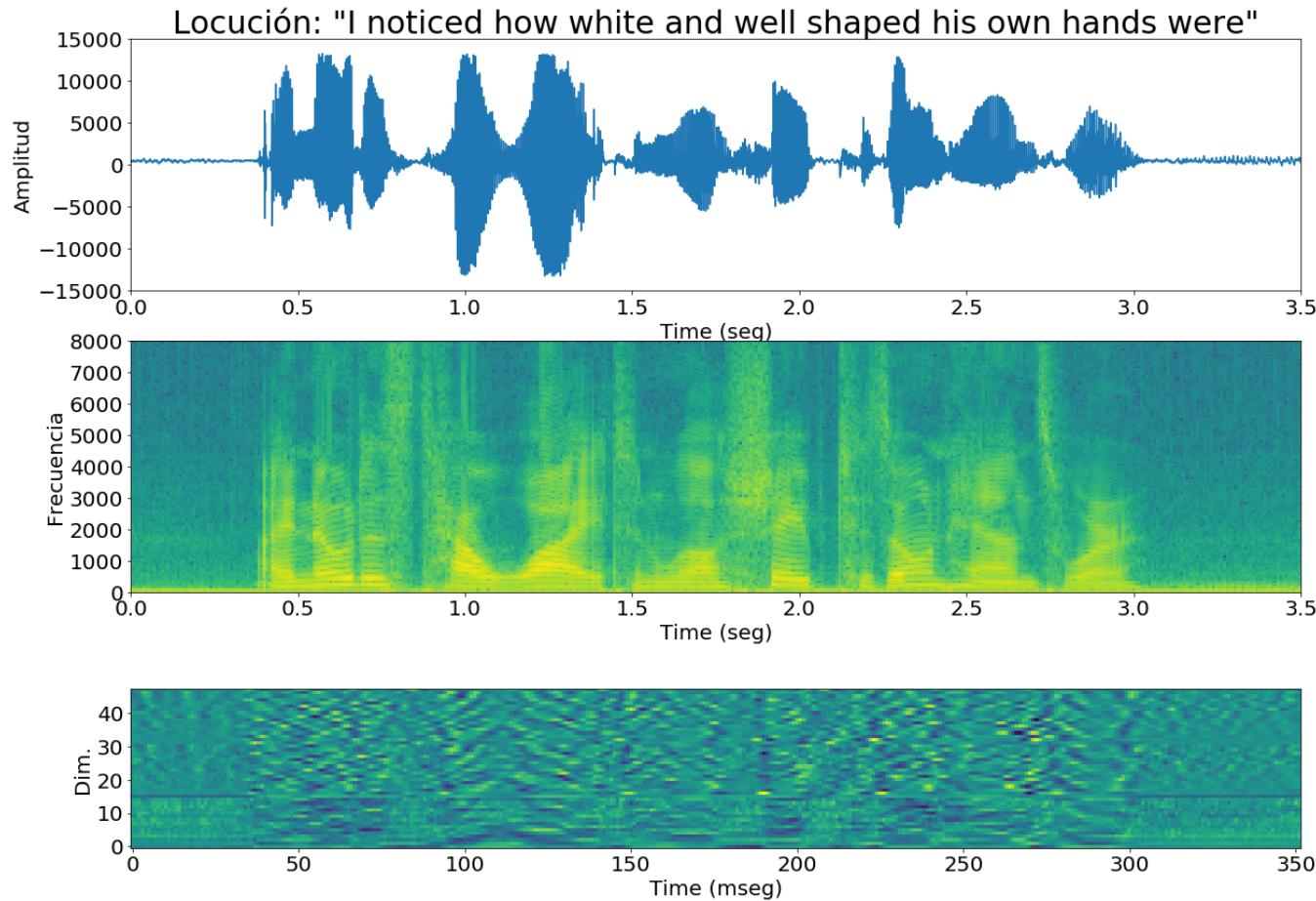
Sistema de reconocimiento del habla

Extracción de características



Sistema de reconocimiento del habla

Extracción de características

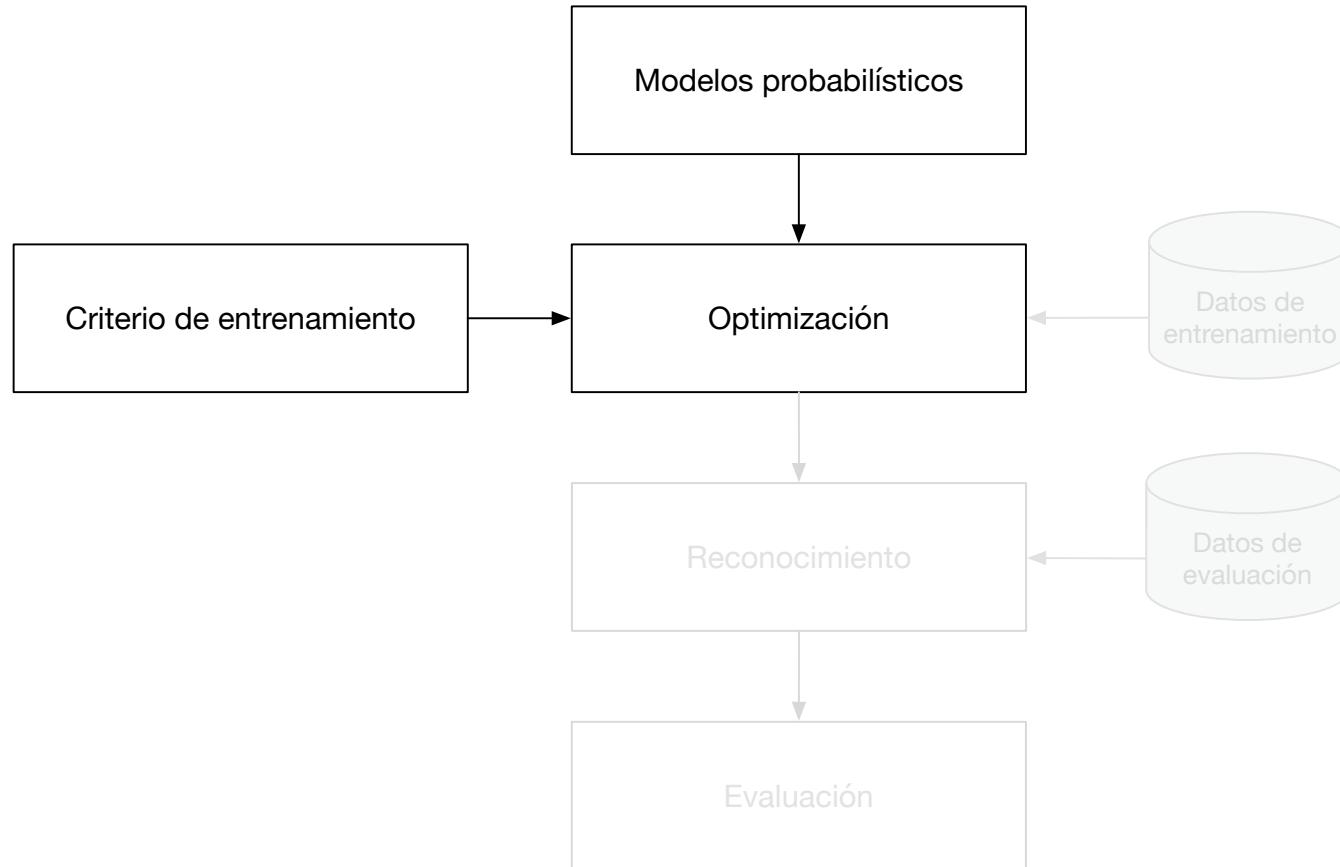


Sistema de reconocimiento del habla

Extracción de características

Extracción de características con PyTLK - Demo

Sistema de reconocimiento del habla



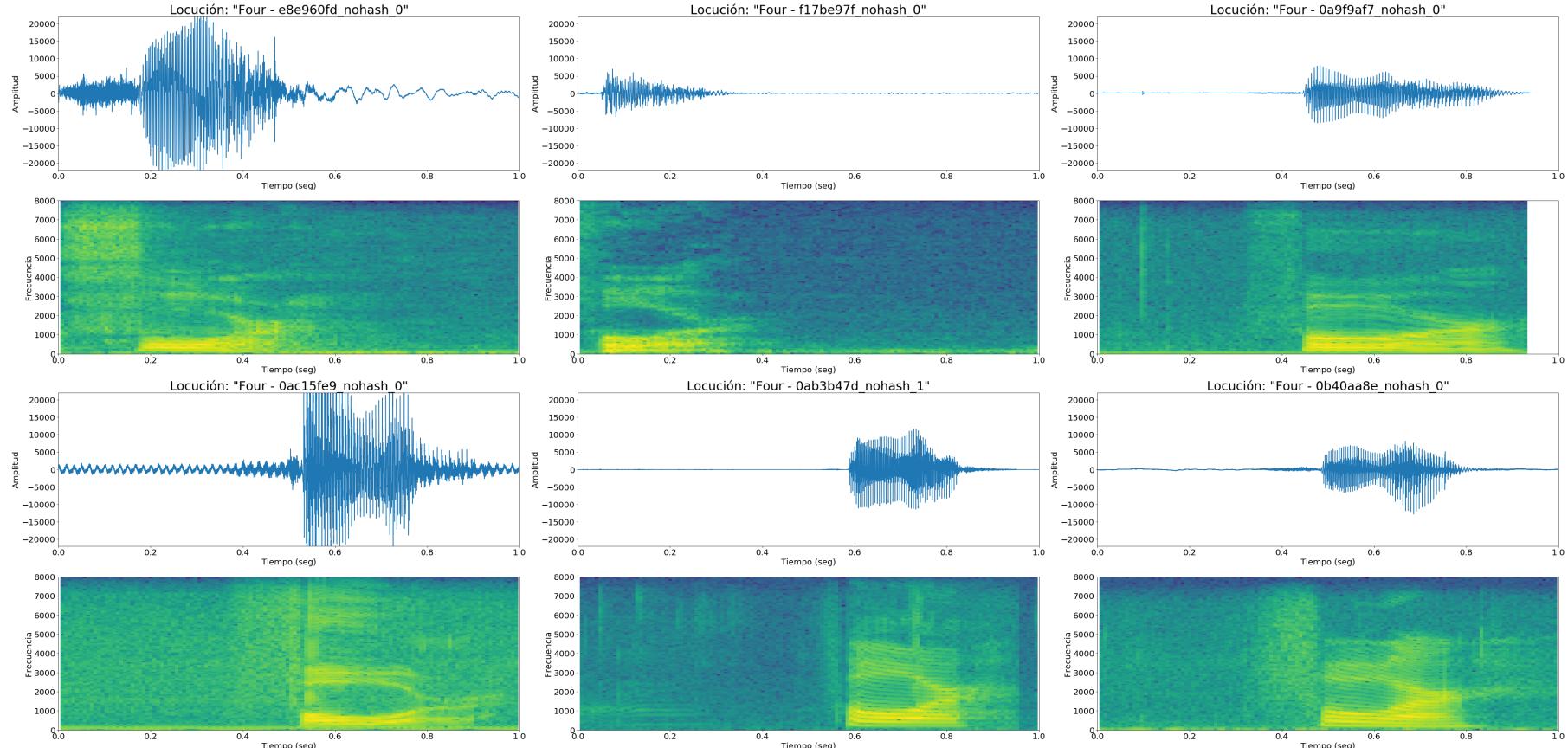
Sistema de reconocimiento del habla

Modelado

- Modelado acústico.
- Modelado léxico.
- Modelado del lenguaje.

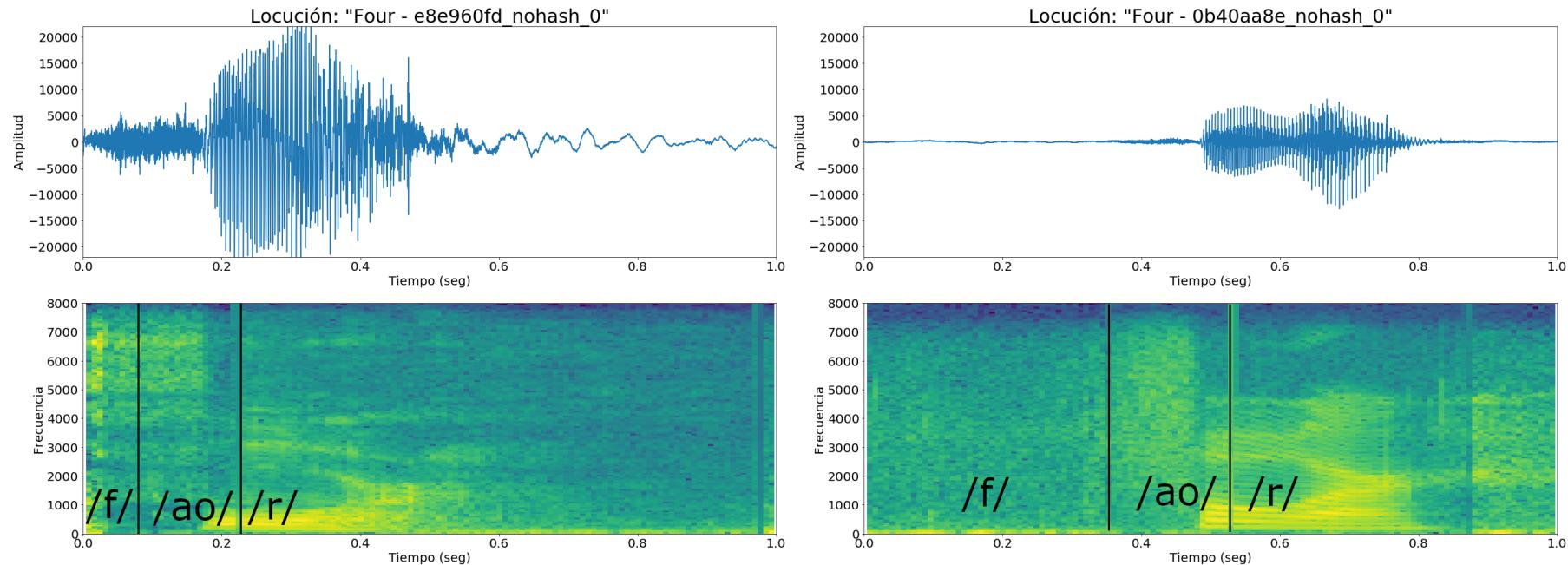
Sistema de reconocimiento del habla

Modelo Acústico



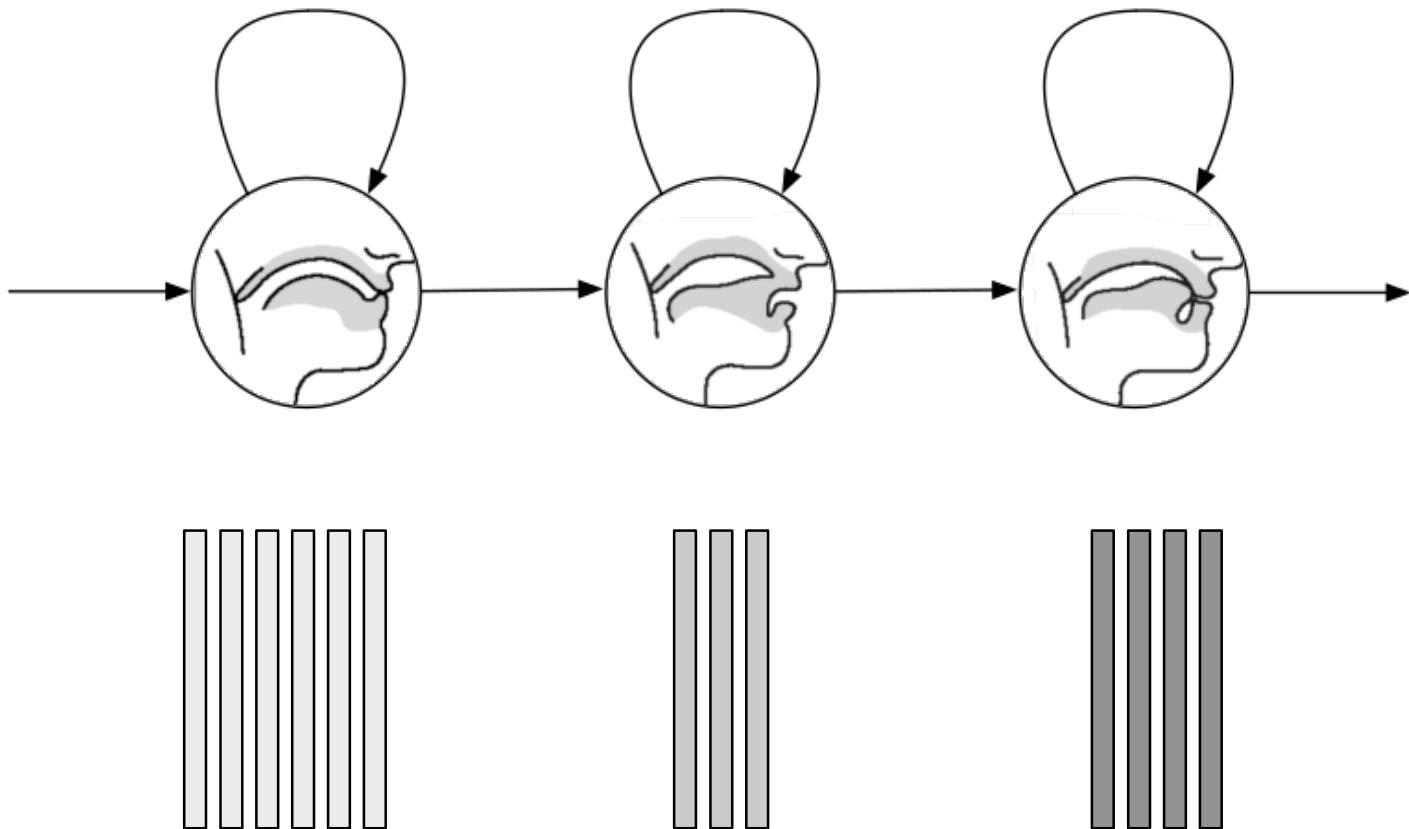
Sistema de reconocimiento del habla

Modelo Acústico



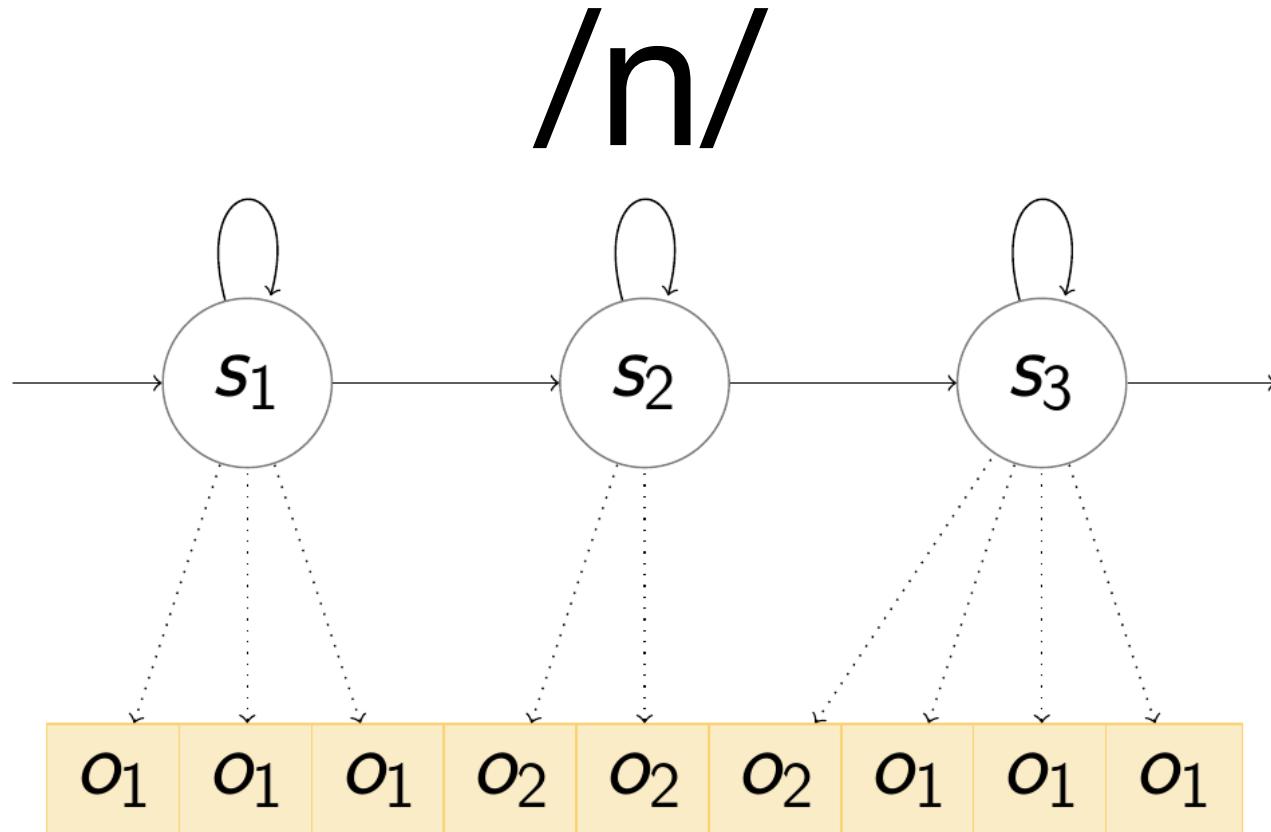
Sistema de reconocimiento del habla

Modelo Acústico



Sistema de reconocimiento del habla

Modelos Ocultos de Markov (Hidden Markov Models - HMM)



Sistema de reconocimiento del habla

Modelo acústico - HMM

Con secuencias y un modelo inicial, existen algoritmos que permiten obtener:

- Parámetros del modelo que mejor explican el conjunto de secuencias de vectores acústicos.

Sistema de reconocimiento del habla

Modelo acústico - HMM

Con secuencias y un modelo inicial, existen algoritmos que permiten obtener:

- Parámetros del modelo que mejor explican el conjunto de secuencias de vectores acústicos.
- Probabilidad de que modelo haya generado una secuencia de vectores acústicos.

Sistema de reconocimiento del habla

Modelo acústico - HMM

Con secuencias y un modelo inicial, existen algoritmos que permiten obtener:

- Parámetros del modelo que mejor explican el conjunto de secuencias de vectores acústicos.
- Probabilidad de que modelo haya generado una secuencia de vectores acústicos.
- Dado un modelo y una secuencia de vectores acústicos, encontrar la secuencia de fonemas.

Sistema de reconocimiento del habla

Modelo acústico - HMM

Con secuencias y un modelo inicial, existen algoritmos que permiten obtener:

- Parámetros del modelo que mejor explican el conjunto de secuencias de vectores acústicos.
- Probabilidad de que modelo haya generado una secuencia de vectores acústicos.
- Dado un modelo y una secuencia de vectores acústicos, encontrar la secuencia de fonemas.

Algoritmos eficientes basados en programación dinámica para estas tareas:

- Algoritmo de Viterbi⁵
- Algoritmo de *Forward-backward*⁶

⁵<https://media.upv.es/#/portal/video/60cdad80-7246-11e9-b1db-e795b40ece52>

⁶<https://media.upv.es/#/portal/video/7ac7cf10-70b4-11e9-a7d3-3df1cef1857d>

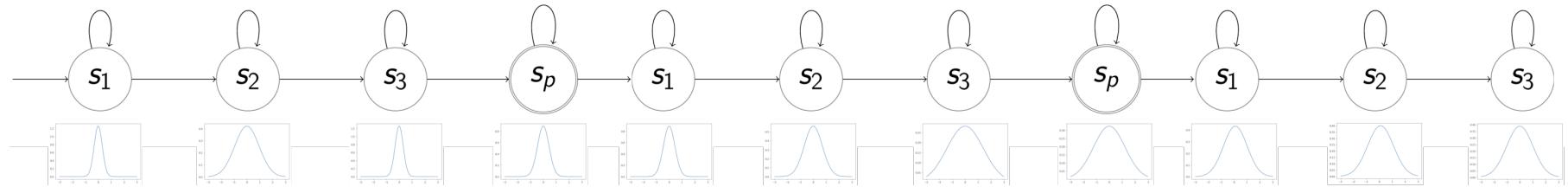
Sistema de reconocimiento del habla

Modelo acústico - HMM

/f/

/ao/

/r/

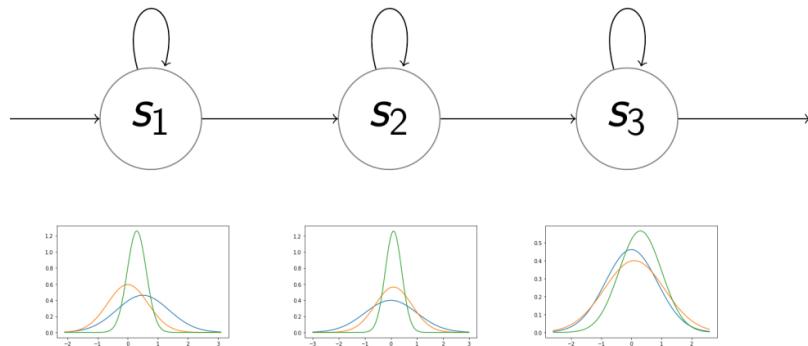


Sistema de reconocimiento del habla

Modelo acústico - HMM

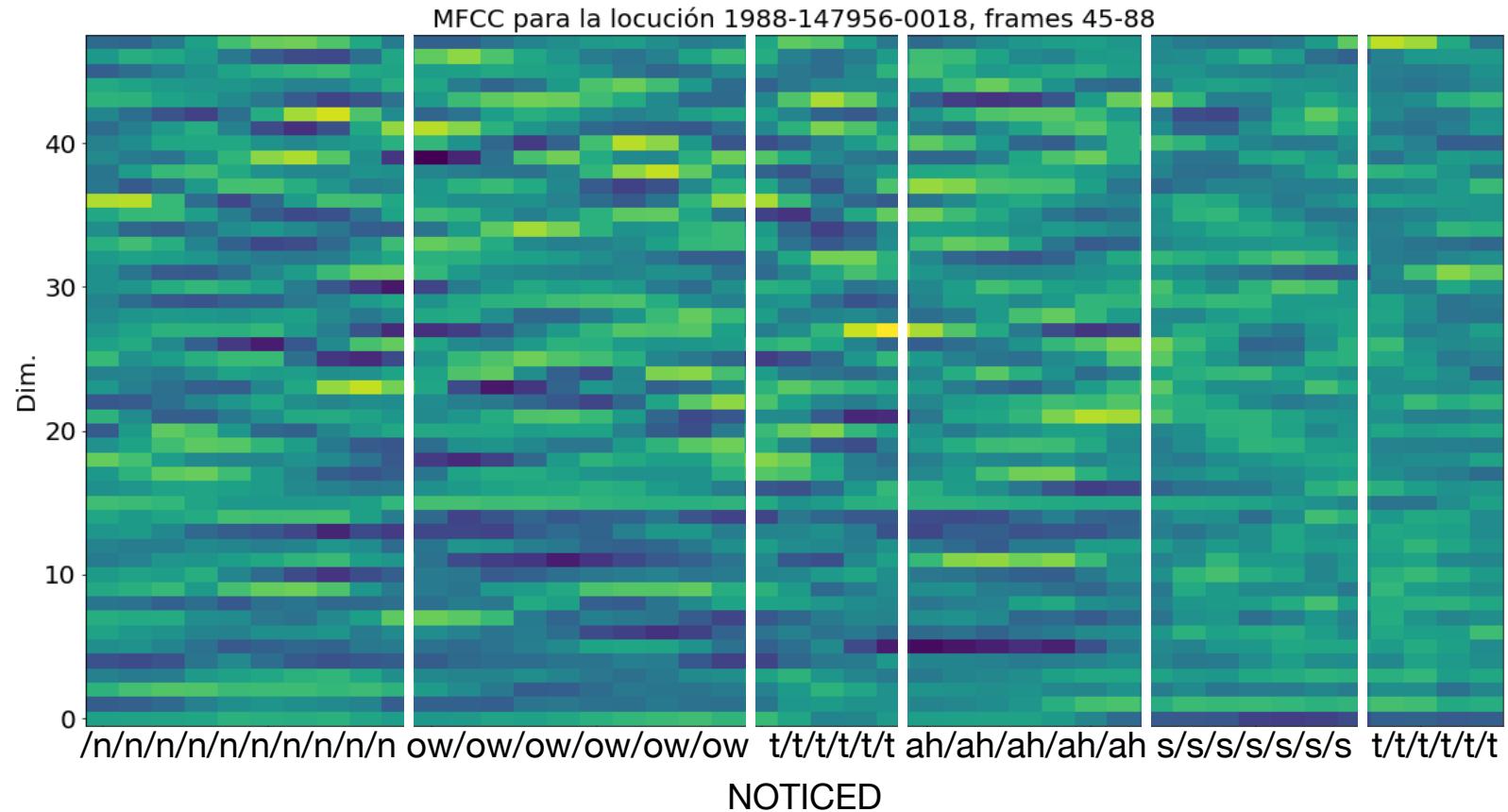
Podemos utilizar mixturas de Gaussianas (**Gaussian Mixture Models - GMM**):

/ao/



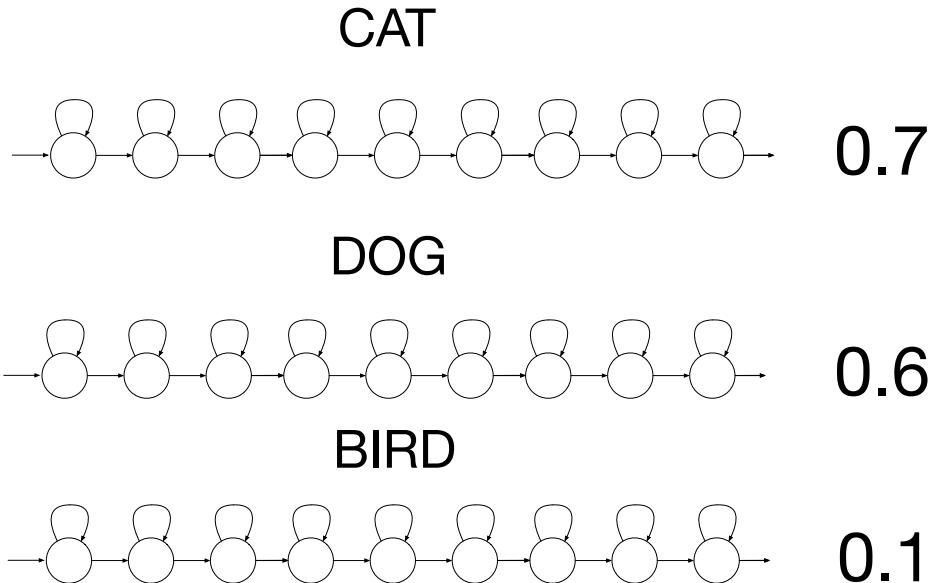
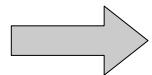
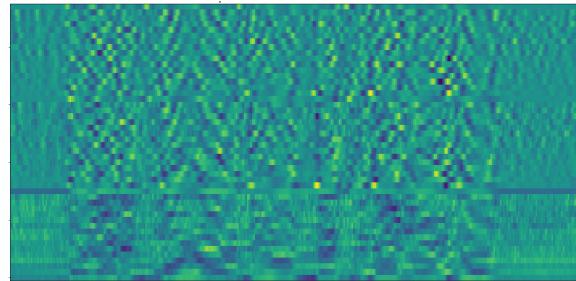
Sistema de reconocimiento del habla

Modelo acústico - HMM



Sistema de reconocimiento del habla

Modelo acústico - HMM



Sistema de reconocimiento del habla

Modelo acústico - HMM

Necesitamos:

- Vectores acústicos, MFCCs.

Con PyTLK podemos:

Sistema de reconocimiento del habla

Modelo acústico - HMM

Necesitamos:

- Vectores acústicos, MFCCs.
- Léxico.

Con PyTLK podemos:

Sistema de reconocimiento del habla

Modelo acústico - HMM

Necesitamos:

- Vectores acústicos, MFCCs.
- Léxico.
- Transcripciones.

Con PyTLK podemos:

Sistema de reconocimiento del habla

Modelo acústico - HMM

Necesitamos:

- Vectores acústicos, MFCCs.
- Léxico.
- Transcripciones.
- HMM iniciales para cada fonema.

Con PyTLK podemos:

Sistema de reconocimiento del habla

Modelo acústico - HMM

Necesitamos:

- Vectores acústicos, MFCCs.
- Léxico.
- Transcripciones.
- HMM iniciales para cada fonema.

Con PyTLK podemos:

- Crear modelo inicial.

Sistema de reconocimiento del habla

Modelo acústico - HMM

Necesitamos:

- Vectores acústicos, MFCCs.
- Léxico.
- Transcripciones.
- HMM iniciales para cada fonema.

Con PyTLK podemos:

- Crear modelo inicial.
- Entrenamiento de HMM basados en Gaussianas.

Sistema de reconocimiento del habla

Modelo acústico - HMM

Necesitamos:

- Vectores acústicos, MFCCs.
- Léxico.
- Transcripciones.
- HMM iniciales para cada fonema.

Con PyTLK podemos:

- Crear modelo inicial.
- Entrenamiento de HMM basados en Gaussianas.
- Clasificar con GMM-HMM.

Sistema de reconocimiento del habla

Modelo acústico - HMM

Necesitamos:

- Vectores acústicos, MFCCs.
- Léxico.
- Transcripciones.
- HMM iniciales para cada fonema.

Con PyTLK podemos:

- Crear modelo inicial.
- Entrenamiento de HMM basados en Gaussianas.
- Clasificar con GMM-HMM.
- Convertir Monofonema → Trifonemas → Trifonemas ligados.

Sistema de reconocimiento del habla

Modelo de lenguaje

Necesitamos un modelo que evalúe secuencias de palabras, ejemplo:

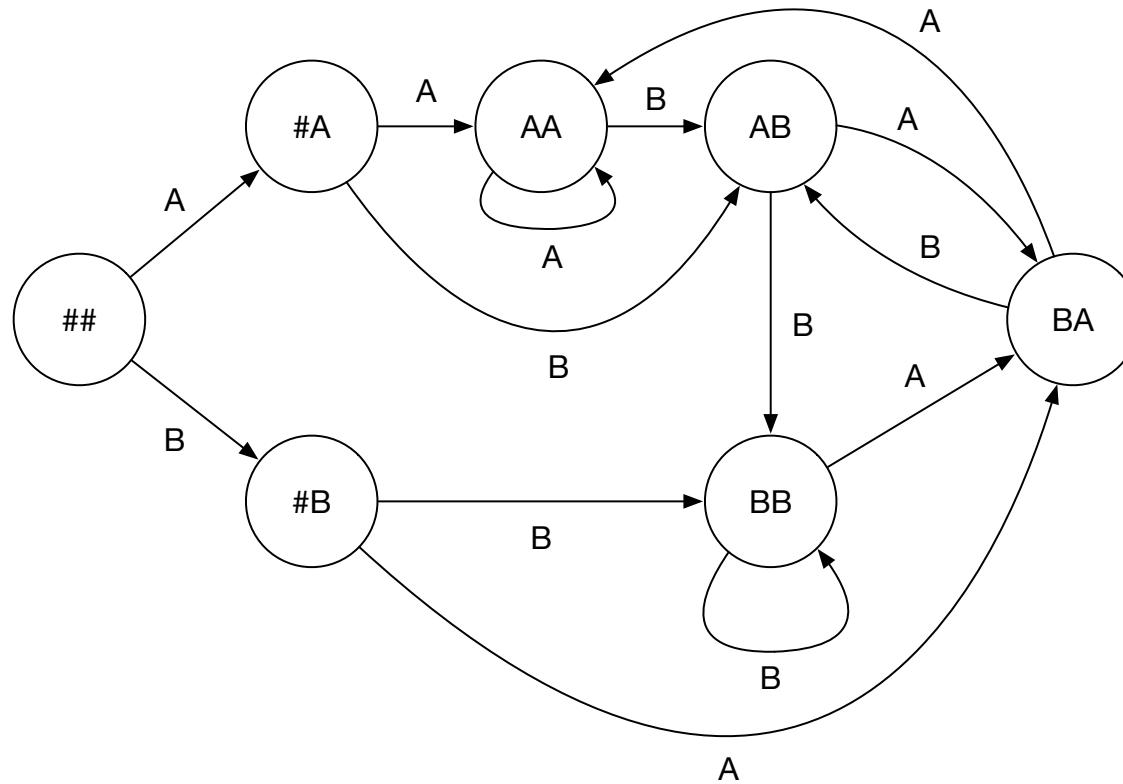
- *El perro marrón* → probable.
- *El perro rojo* → poco probable.

Modelo de n-gramas: Proporciona la probabilidad de una palabra considerando:

- Unigramas.
- Bigramas.
- Trigramas.
- ...

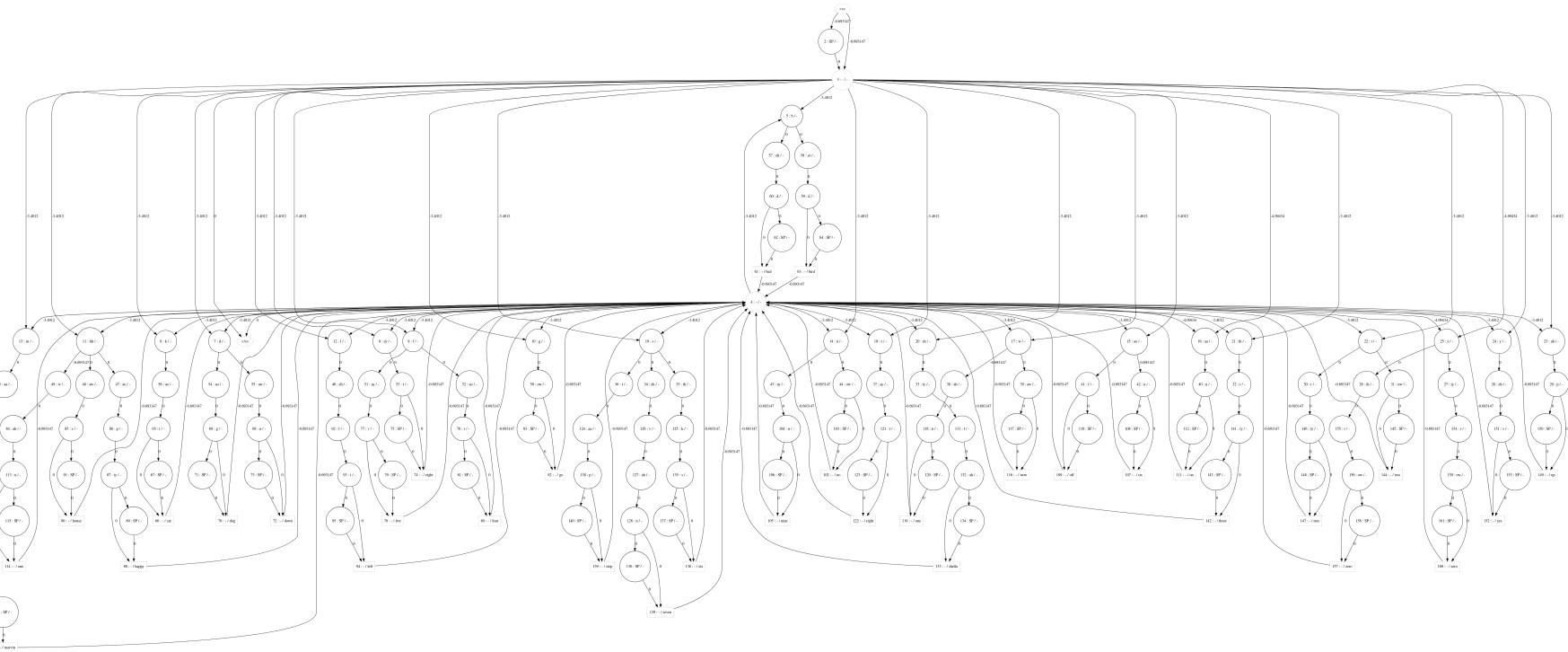
Sistema de reconocimiento del habla

Modelo de lenguaje - n-gramas



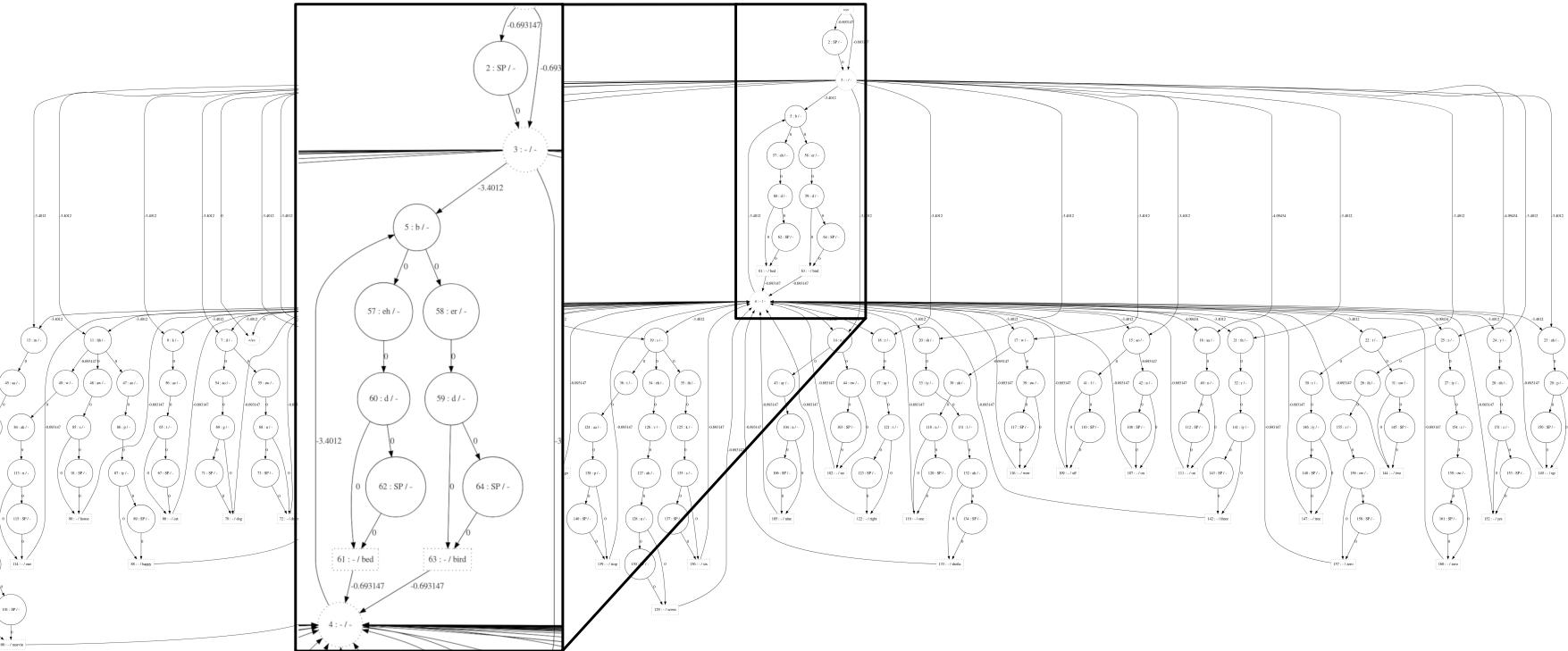
Sistema de reconocimiento del habla

Modelo de lenguaje - n-gramas



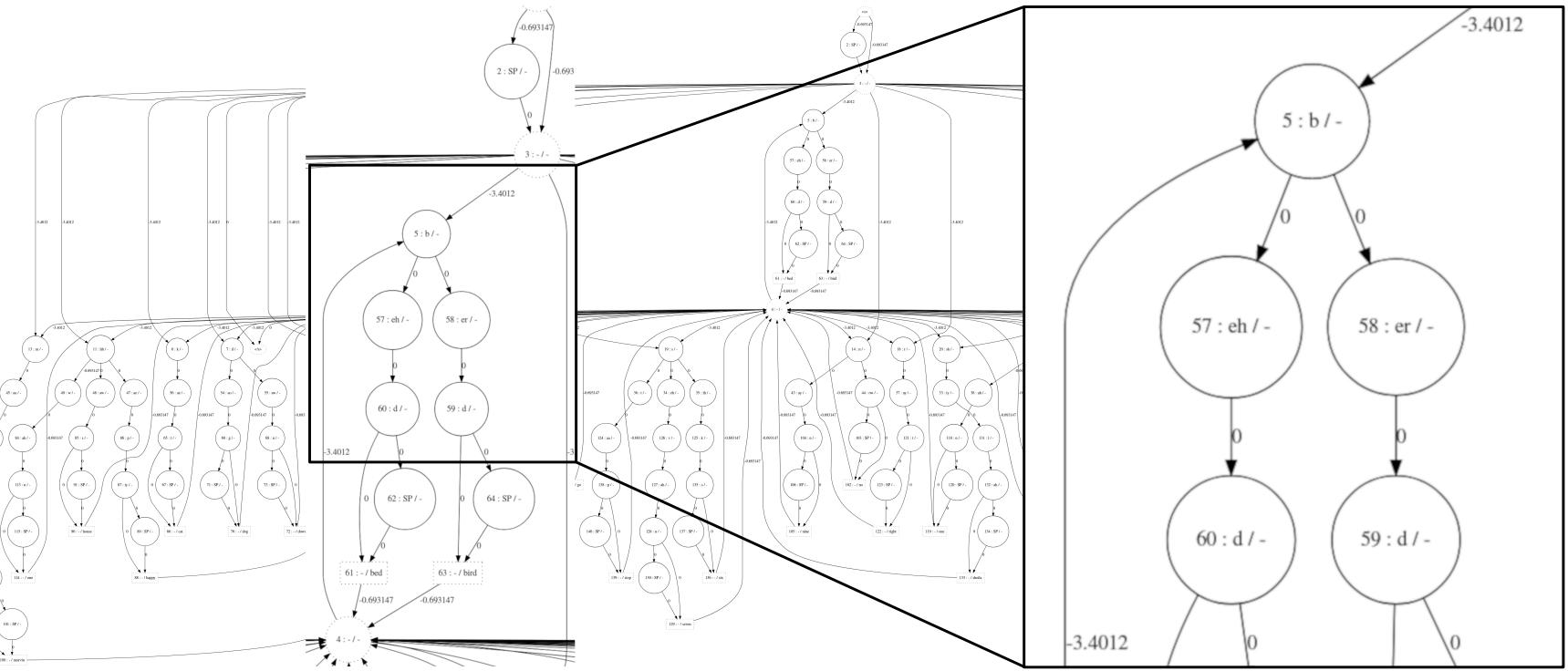
Sistema de reconocimiento del habla

Modelo de lenguaje - n-gramas



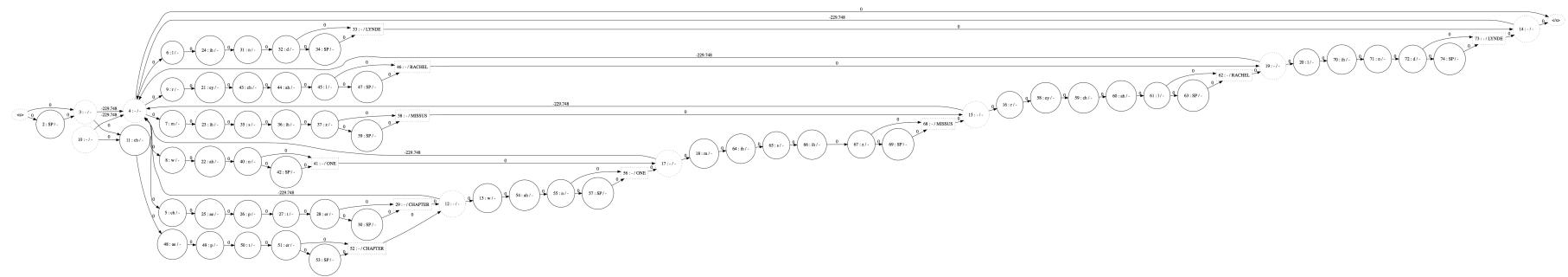
Sistema de reconocimiento del habla

Modelo de lenguaje - n-gramas



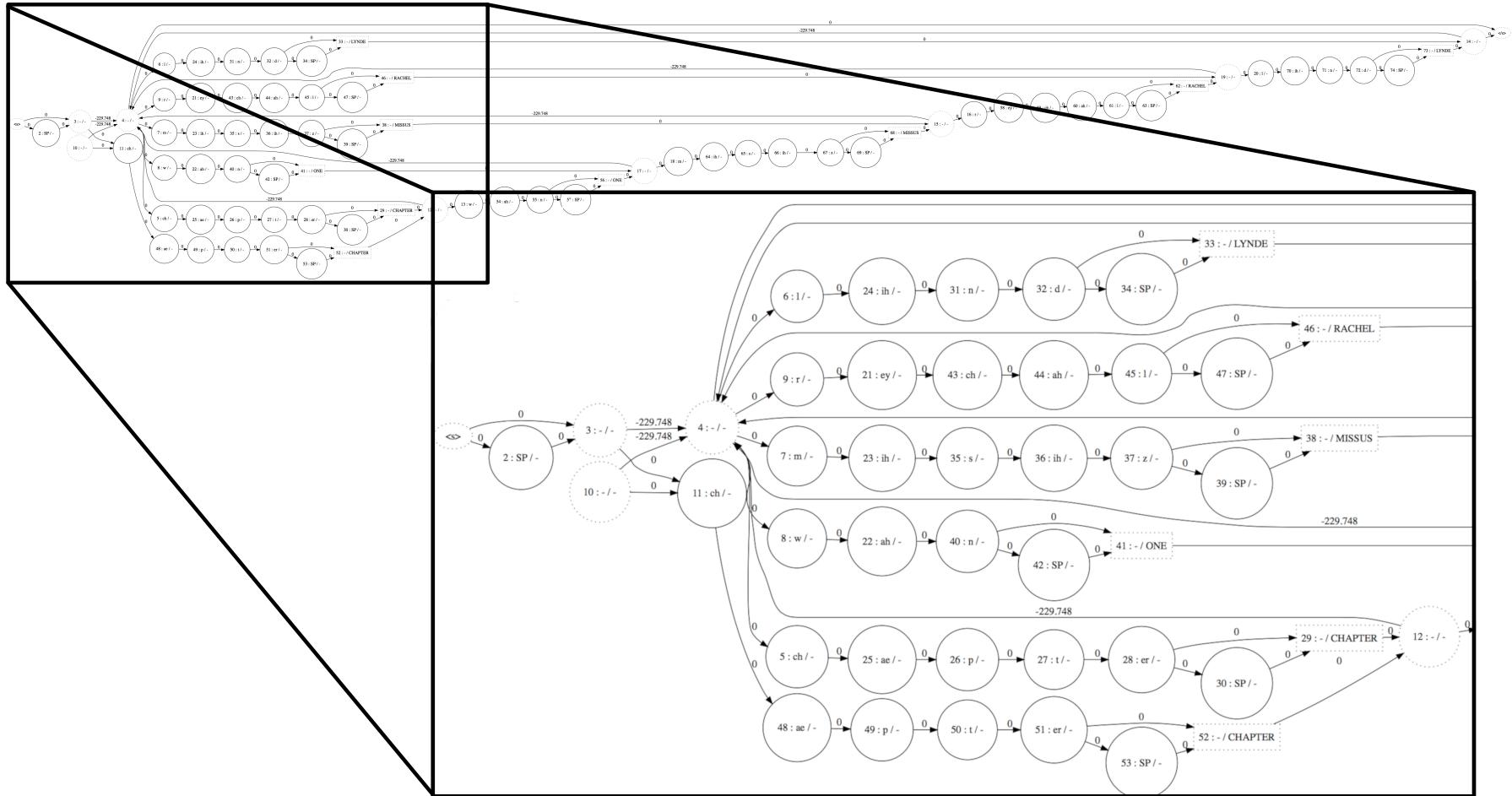
Sistema de reconocimiento del habla

Modelo de lenguaje - n-gramas



Sistema de reconocimiento del habla

Modelo de lenguaje - n-gramas



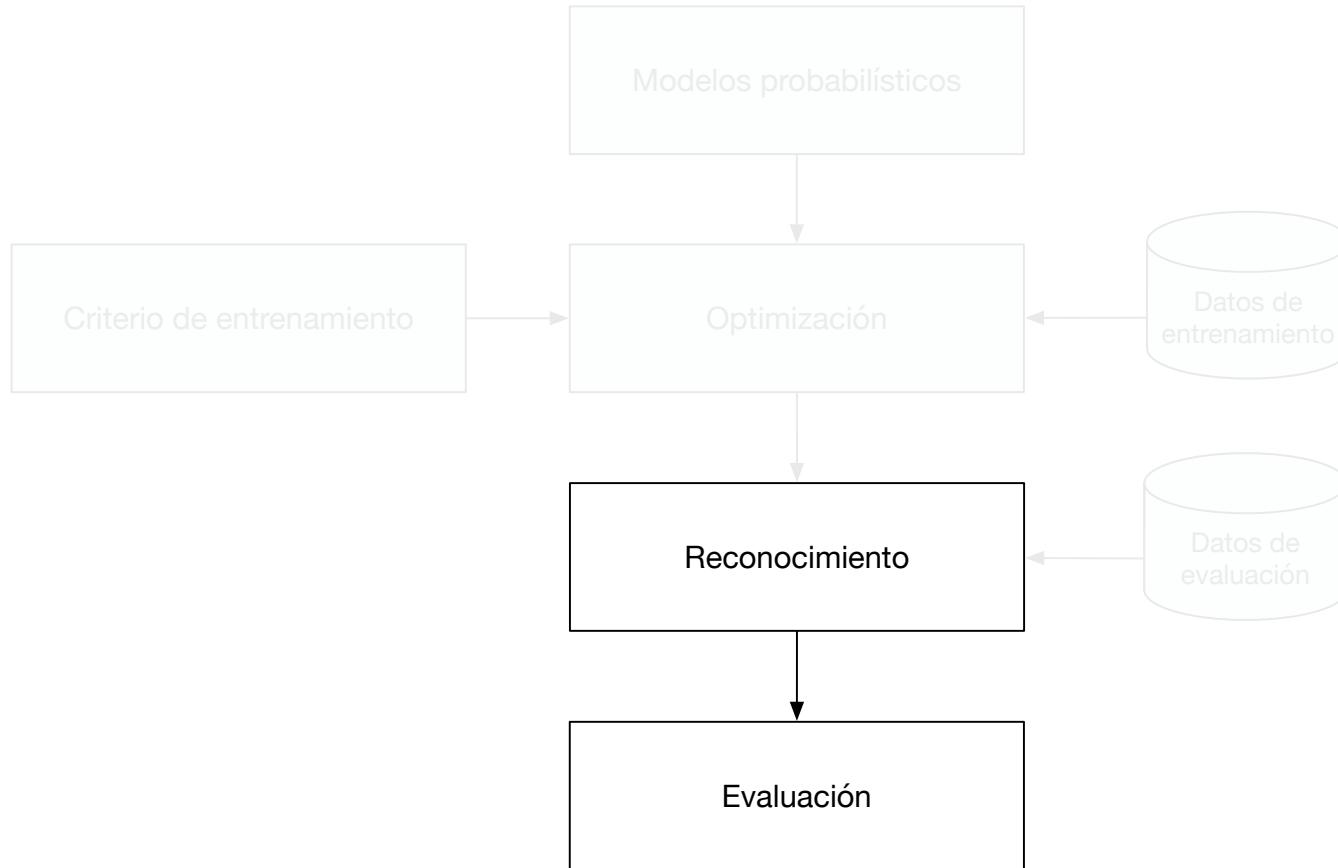
Sistema de reconocimiento del habla

Modelo de lenguaje

Pasos en PyTLK:

- Obtener modelo de los datos (SRILM, ...).
- Convertir a formato TLK.
- Convertir a Grafo TLK.

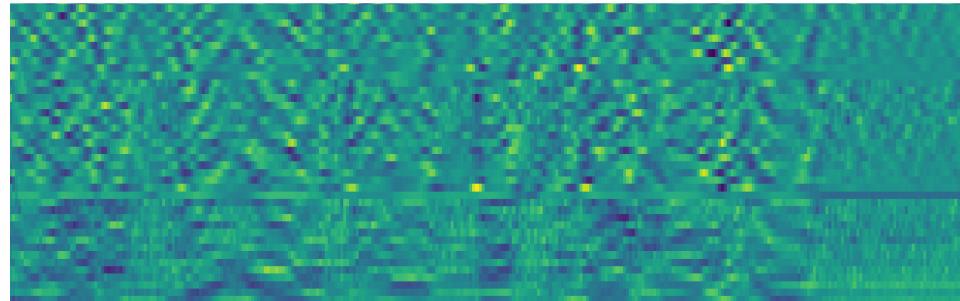
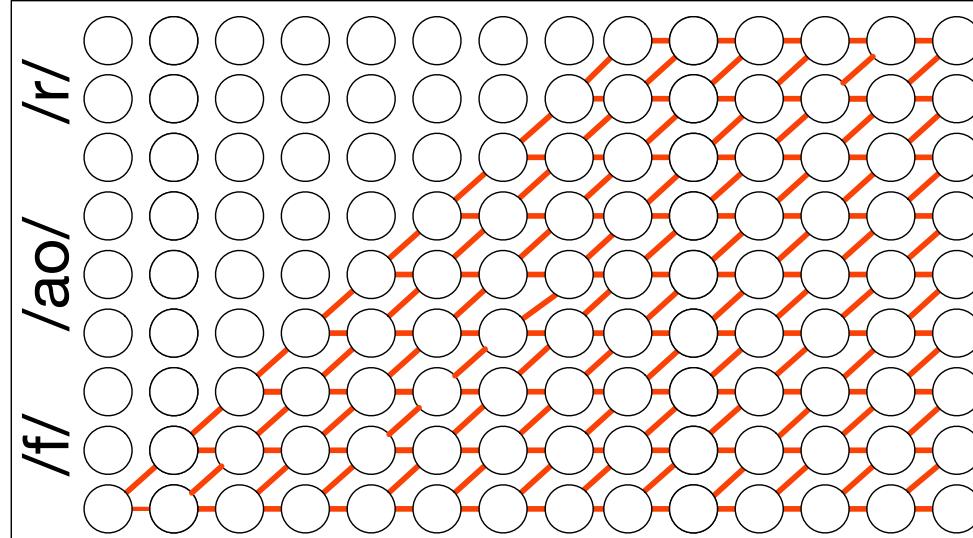
Sistema de reconocimiento del habla



Sistema de reconocimiento del habla

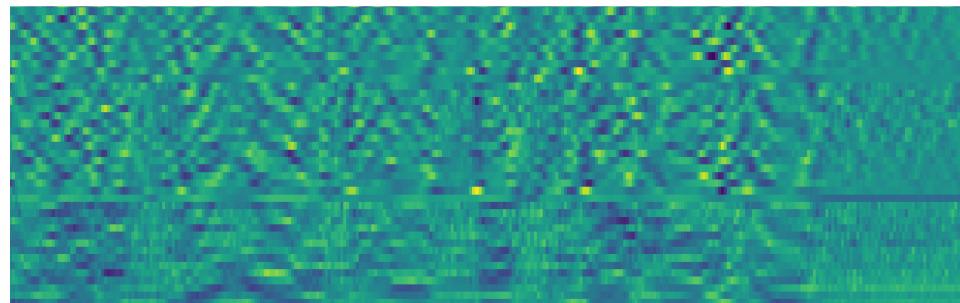
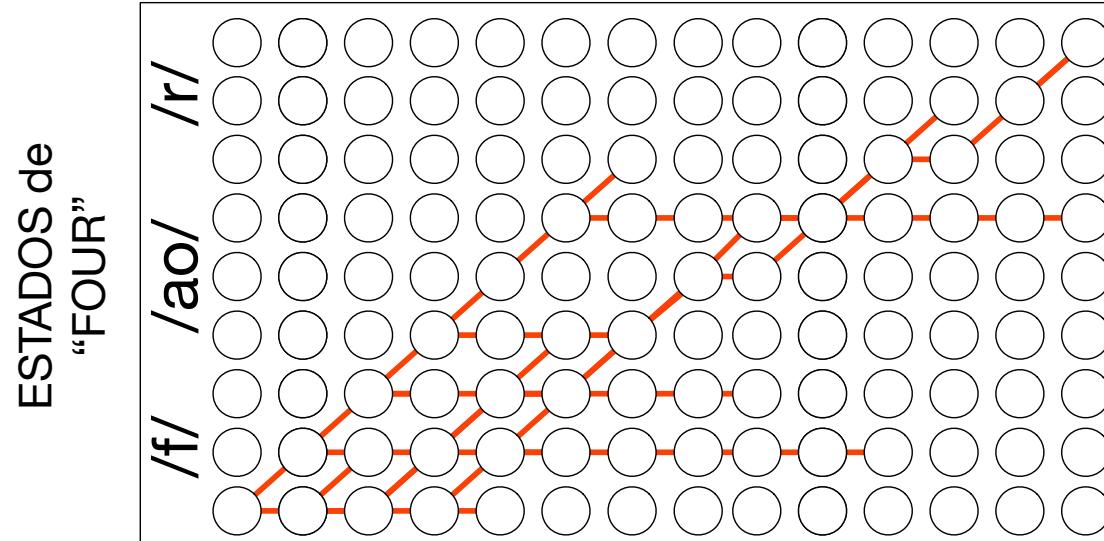
Reconocimiento

ESTADOS de
“FOUR”



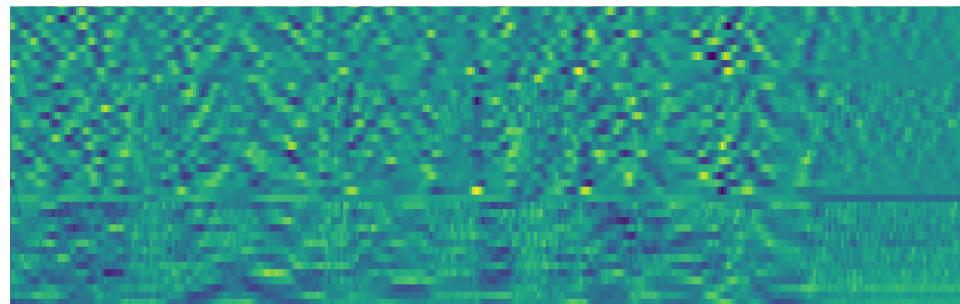
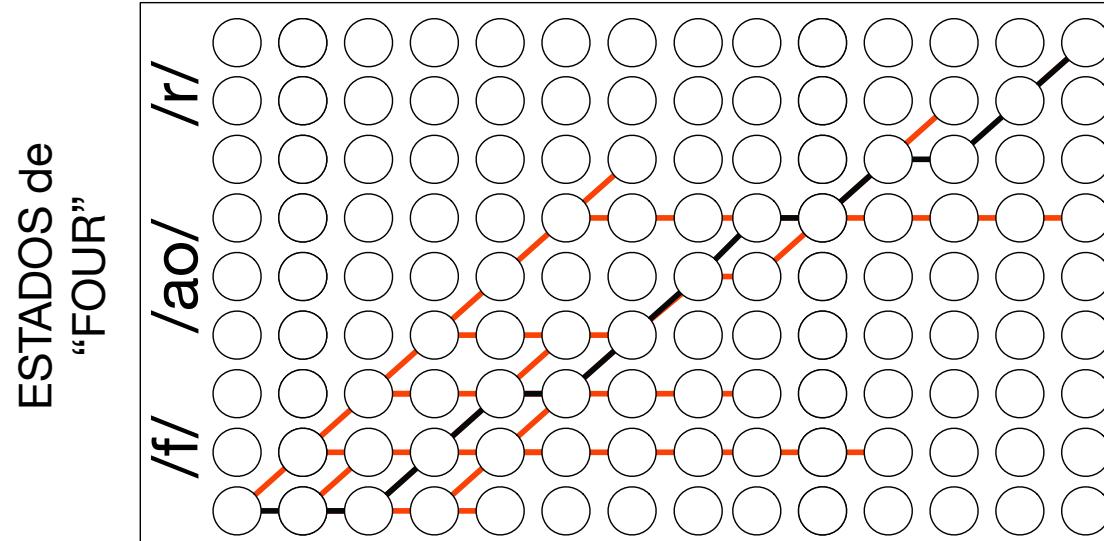
Sistema de reconocimiento del habla

Reconocimiento



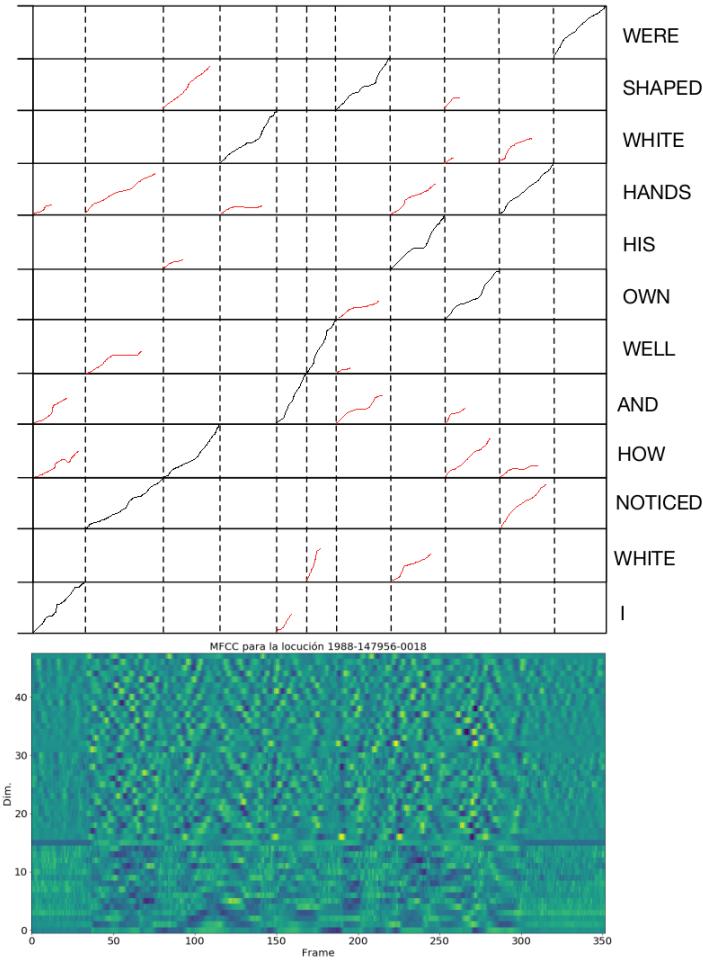
Sistema de reconocimiento del habla

Reconocimiento



Sistema de reconocimiento del habla

Reconocimiento



Sistema de reconocimiento del habla

Reconocimiento

Reconocimiento con GMM-HMM - Demo PyTLK

- Modelos HMM.
- LM o Grafo.
- Modelo léxico.

Sistema de reconocimiento del habla

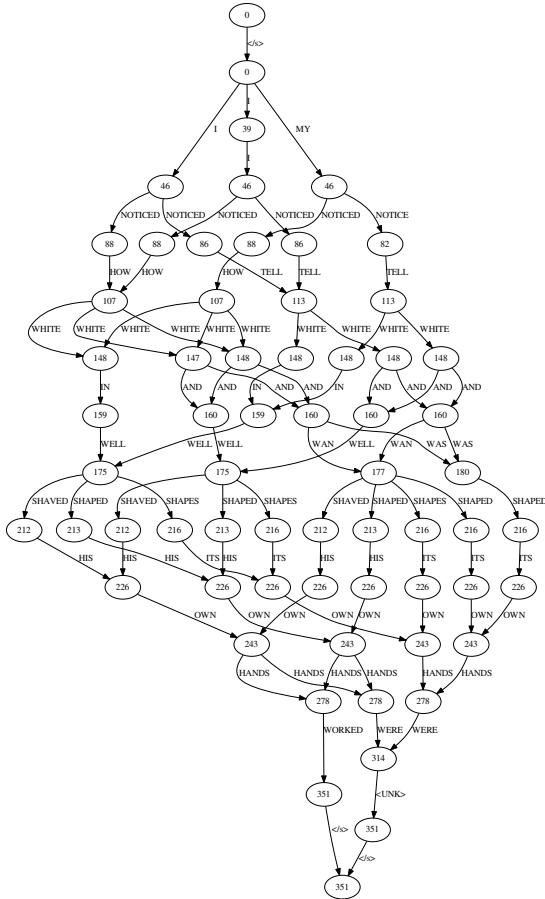
Reconocimiento

GMM-HMM → Hybrid-HMM - Demo PyTLK

- Sacamos alineamientos (*frame-trifonema*).
- Entrenamos una red neuronal (BLSTM) para clasificar *frame-trifonema*.
- Utilizamos la red en lugar de las GMM para la emisión.

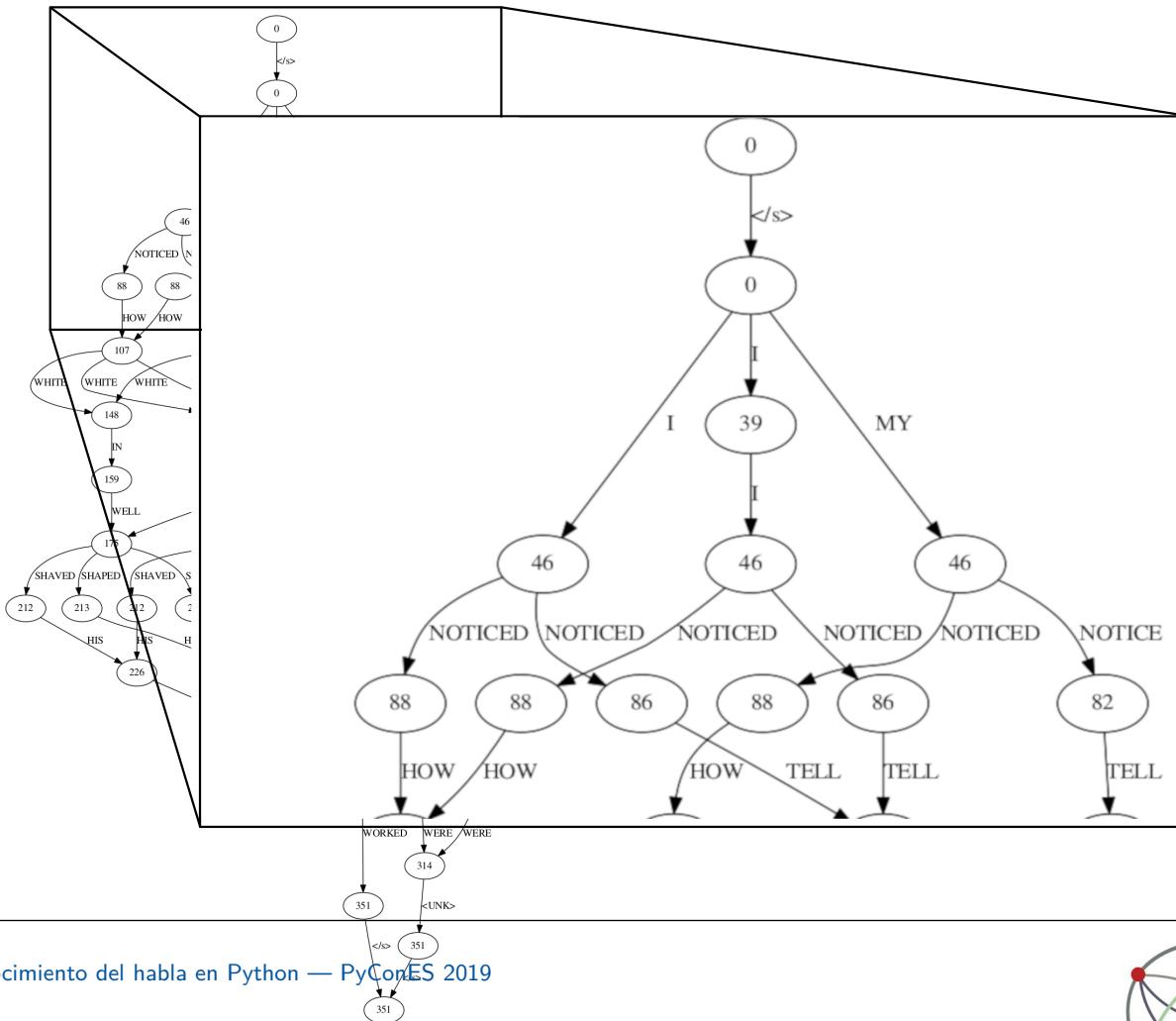
Sistema de reconocimiento del habla

Técnicas avanzadas:



Sistema de reconocimiento del habla

Técnicas avanzadas:



MLLP

Sistema de reconocimiento del habla

Técnicas avanzadas:

- Adaptación.
- Modelos de lenguaje basados en redes neuronales.
- Reconocimiento multi-pasada.
- ...

Conclusiones

Hemos visto:

- En qué consiste la tarea del reconocimiento del habla.

Conclusiones

Hemos visto:

- En qué consiste la tarea del reconocimiento del habla.
- Los desafíos que plantea este problema.

Conclusiones

Hemos visto:

- En qué consiste la tarea del reconocimiento del habla.
- Los desafíos que plantea este problema.
- Las partes y los conceptos intuitivos de un sistema de reconocimiento del habla.

Conclusiones

Hemos visto:

- En qué consiste la tarea del reconocimiento del habla.
- Los desafíos que plantea este problema.
- Las partes y los conceptos intuitivos de un sistema de reconocimiento del habla.
- Cómo desarrollar un sistema de reconocimiento del habla con Python y recursos abiertos.

Gracias por la atención

Contacto: jjorge@dsic.upv.es

