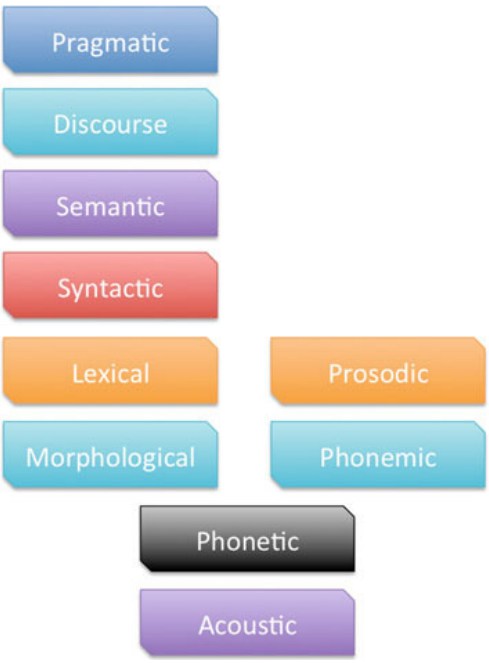
DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

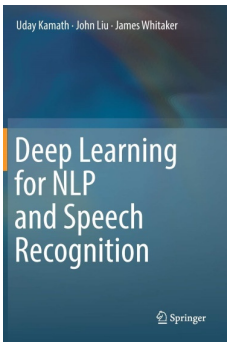
Reconocimiento de voz

Fernando Berzal, berzal@acm.org

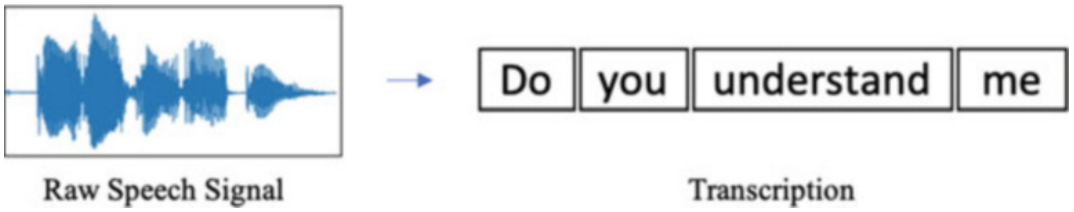
NLP

NLP

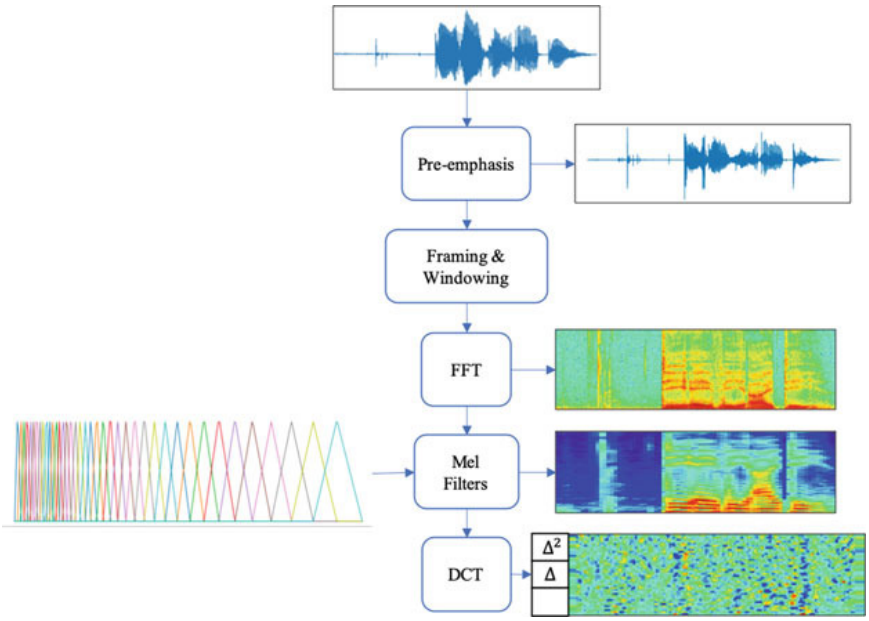




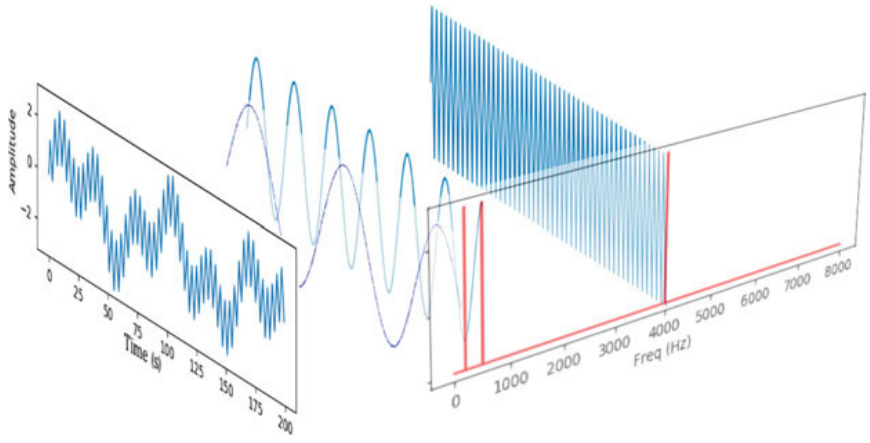
Reconocimiento de voz

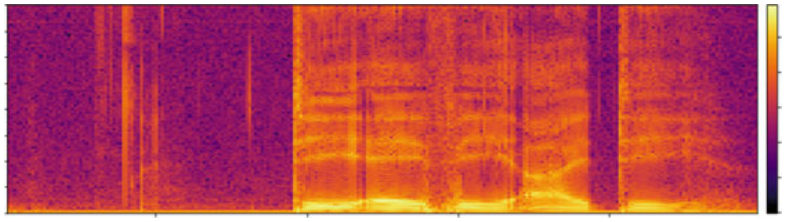


Reconocimiento de voz

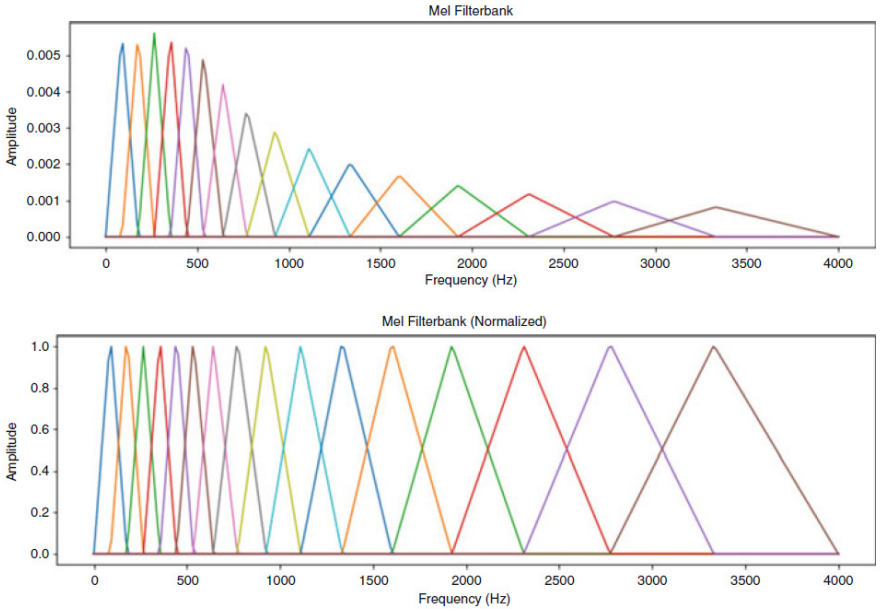


FFT: Espectro de la señal

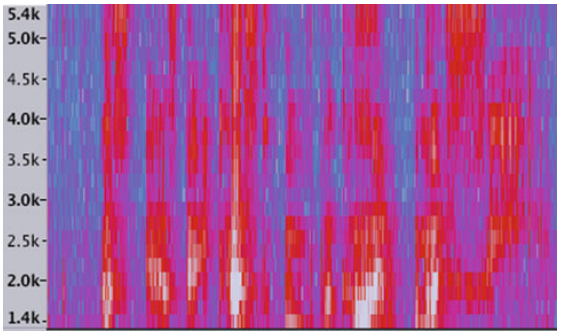




Banco de filtros Mel

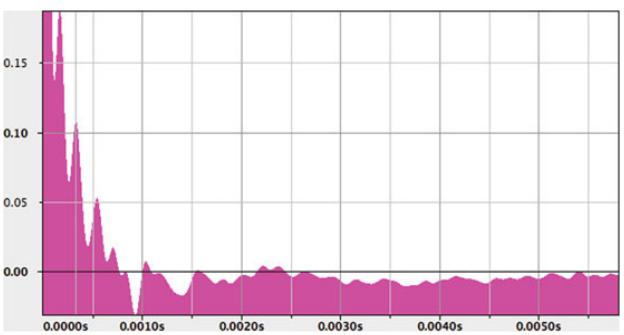


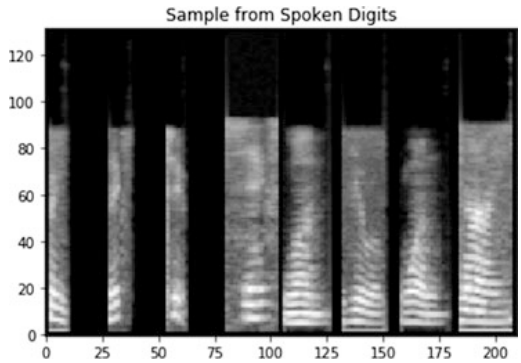
Espectrograma & cepstrum



$$C = \left| F^{-1}\left(\log F\left(f(t)\right)\right)\right|^2$$

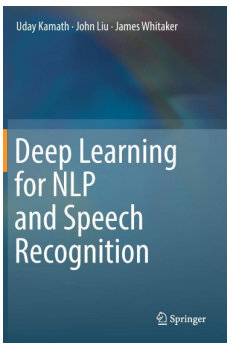
$$C = DCT\left(\log\left(MEL\left(F\left(f(t)\right)\right)\right)\right)$$

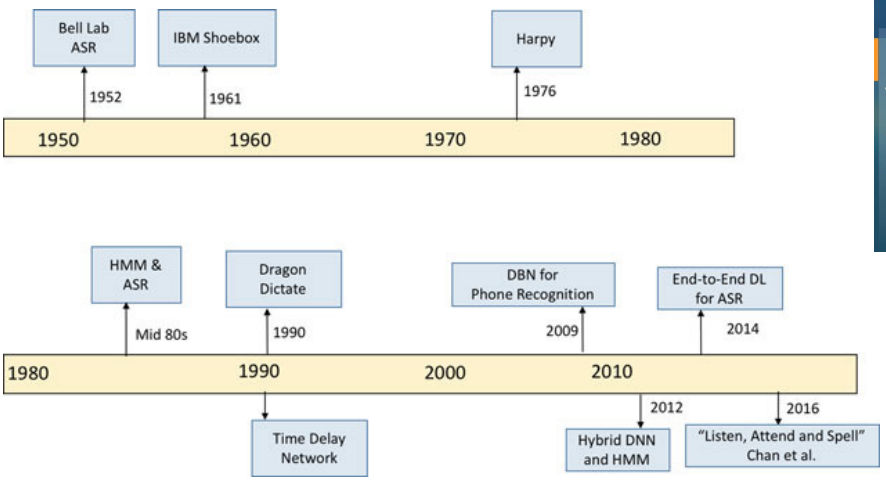




Evolución

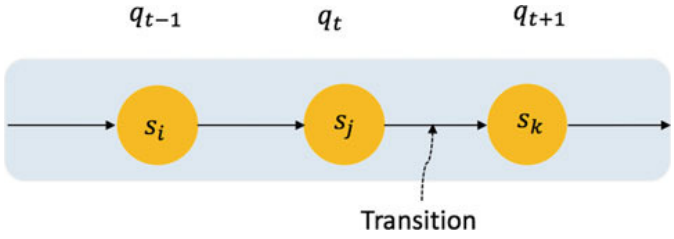
Historia de los sistemas de reconocimiento de voz



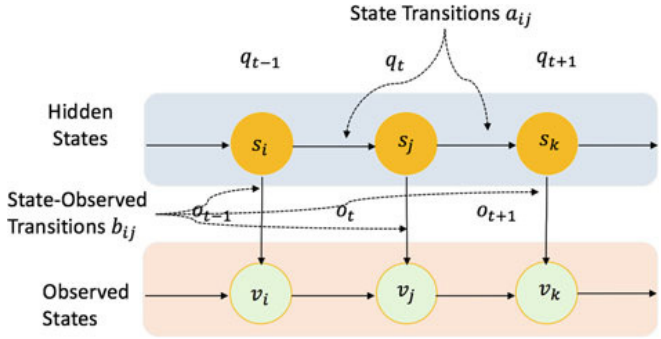


Modelos de Markov

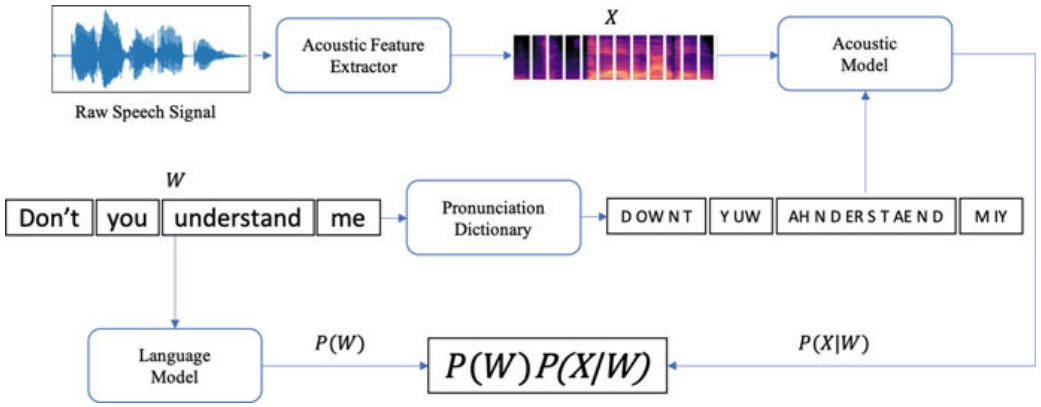
- Cadena de Markov



- Modelo oculto de Markov
  [HMM = Hidden Markov Model]



Reconocimiento de voz estadístico



$$W^* = \underset{W \in V^*}{\operatorname{argmax}} P(W|X)$$

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)}$$

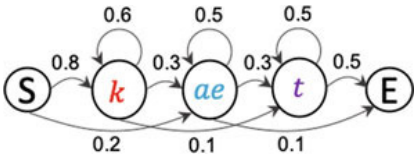
$$W^* = \underset{W \in V^*}{\operatorname{argmax}} P(X|W)P(W)$$

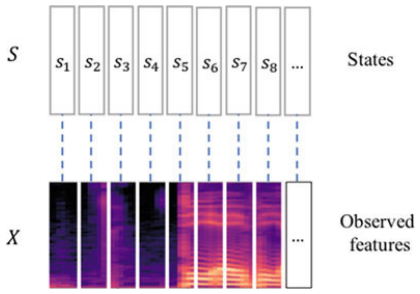
Reconocimiento de voz estadístico

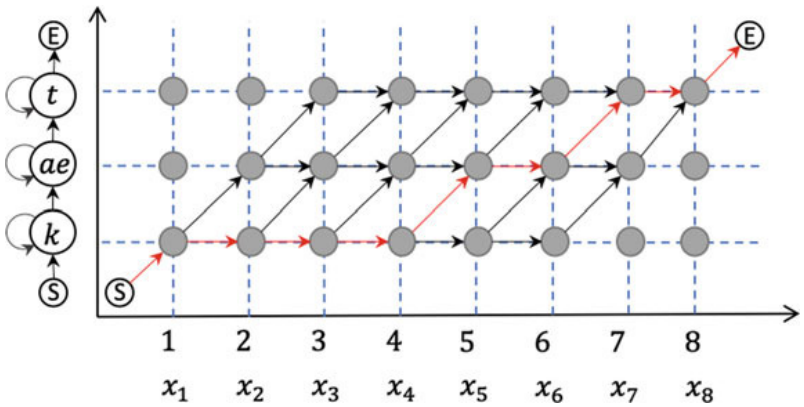
Modelo acústico

P(X|W)

HMM







Reconocimiento de voz estadístico

Modelo del lenguaje

P(W)

n-gramas

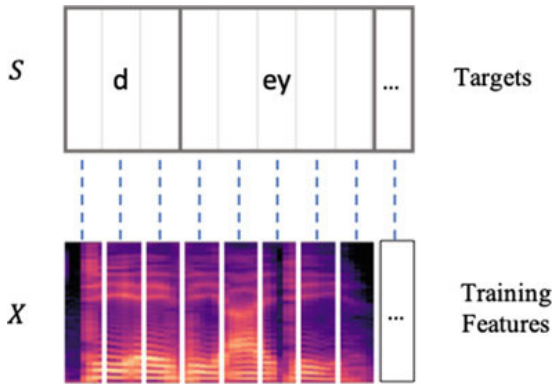
Reconocimiento de voz estadístico

Decodificación del HMM

Secuencia óptima de "palabras"

P(x|s)

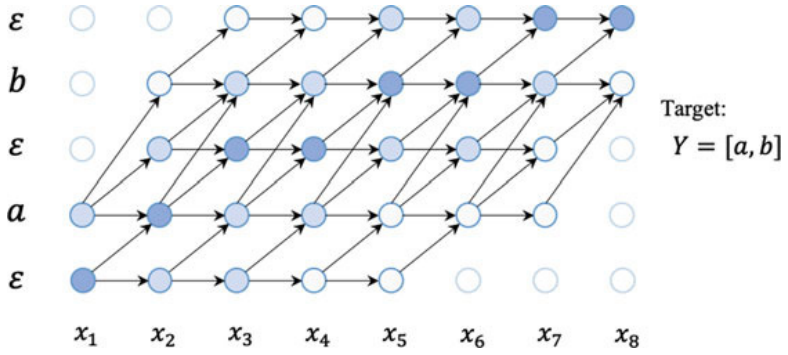
GMM $\rightarrow$ DNN



Reconocimiento de voz

Sistemas end-to-end basados en deep learning

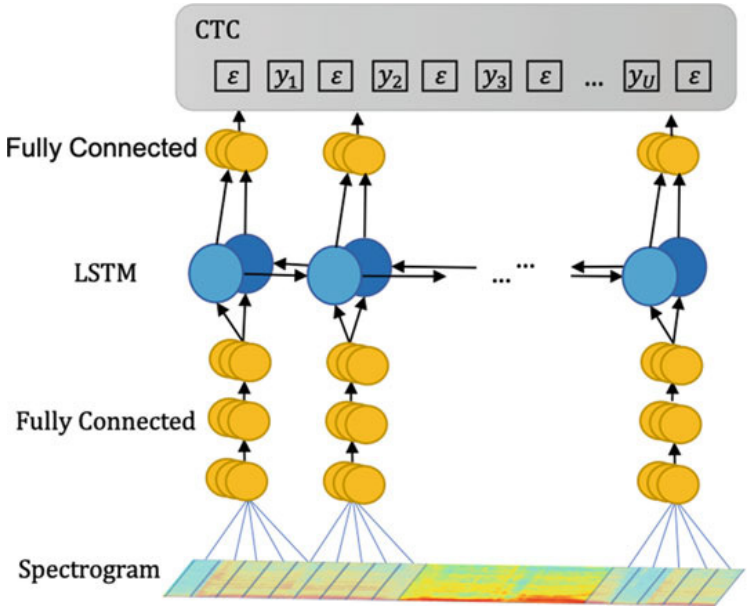
CTC [Connectionist Temporal Classification]



Reconocimiento de voz

Sistemas end-to-end basados en deep learning

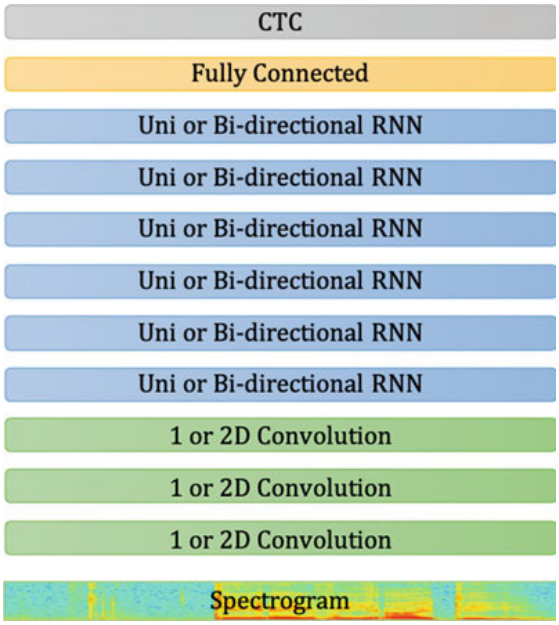
Deep Speech 1



Reconocimiento de voz

Sistemas end-to-end basados en deep learning

Deep Speech 2

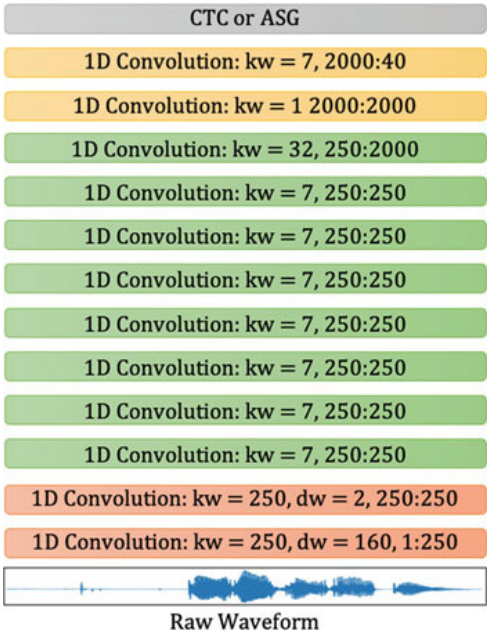


Reconocimiento de voz

Sistemas end-to-end basados en deep learning

Wav2Letter

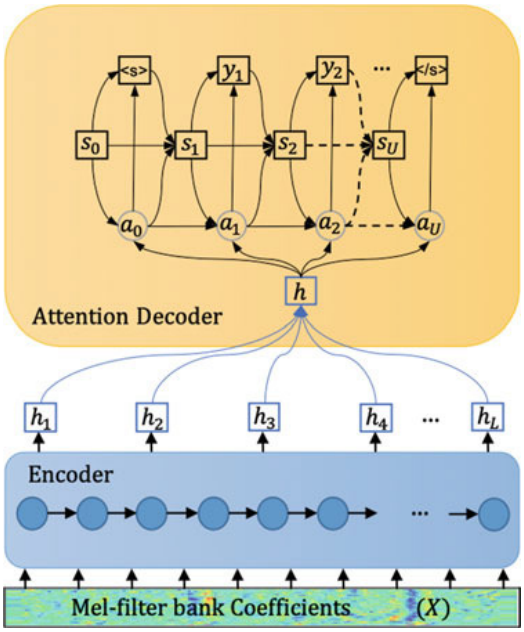
CNN



Reconocimiento de voz

Sistemas end-to-end basados en deep learning

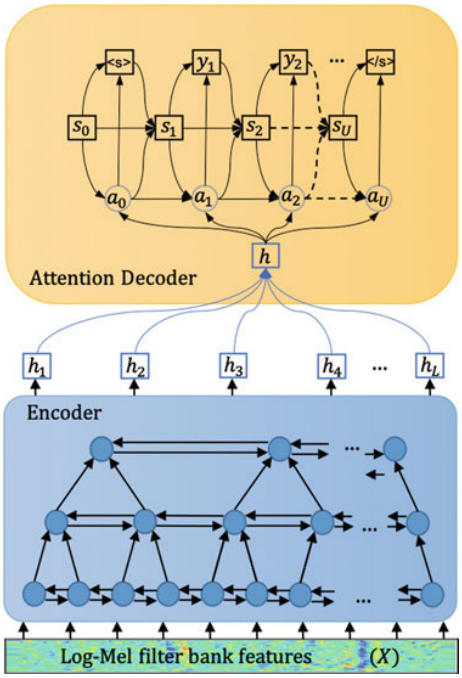
Mecanismos de atención (seq2seq)



Reconocimiento de voz

Sistemas end-to-end basados en deep learning

LAS
Listen, Attend & Spell

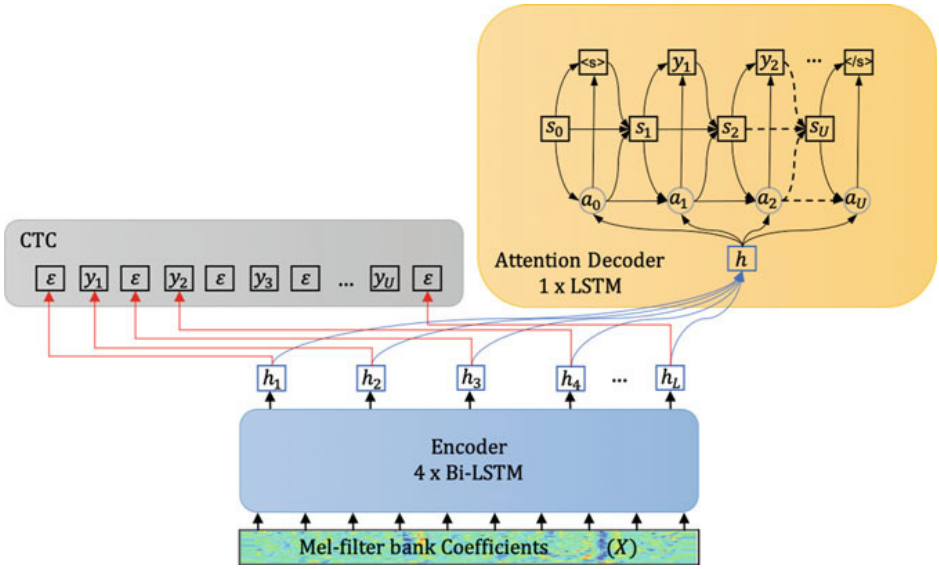


Reconocimiento de voz

Sistemas end-to-end basados en deep learning

ESPnet

CTC + Atención

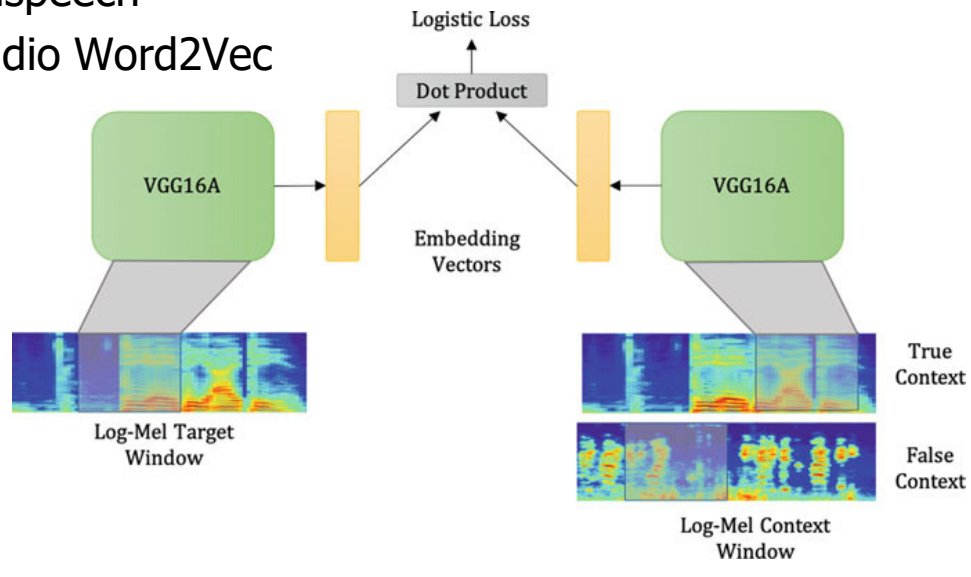


Reconocimiento de voz

Sistemas end-to-end basados en deep learning

Embeddings

- Unspeech
- Audio Word2Vec



Software

Frameworks para reconocimiento de voz

- Sphinx (Java, CMU)
  https://cmusphinx.github.io/
- Kaldi (C++)
  https://github.com/kaldi-asr/kaldi

- ESPNet (deep learning ASR, PyTorch/Chainer)
  https://espnet.github.io/espnet/

Procesamiento de audio

- SoX [Sound eXchange] (C)
  http://sox.sourceforge.net/
- LibROSA (Python)
  https://librosa.org/