# Análisis de modelos End to End en Reconocimiento Automático del Habla

Jaime Ferrando Huertas, Javier Martínez Bernia

February 2021

## Contents

# 1 Introducción

El Reconocimiento Automático del Habla ha sido un campo con mucha investigación en los últimos años. Tradicionalmente ha sido un campo dominado por modelos híbridos donde el modelo acústico se modelaba mediante un modelo oculto de Markov con estimación de probabilidades mediante mixturas de gaussianas (HMM-GMM) y el modelo de lenguaje era basado en n-gramas. En los últimos años ha habido una transición de estos modelos híbridos hacia modelos llamados "end-to-end" donde se encapsula en un solo modelo neuronal todo el proceso de reconocimiento del habla. Durante esta transición también estamos viendo distintas variaciones de modelos híbridos que también añadían estas redes a su modelo, ya sea sustituyendo las mixturas gaussianas del modelo acústico por una DNN o el modelo de n-gramas por una red para modelar el lenguaje.

En este trabajo vamos a estudiar distintas arquitecturas end-to-end que han marcado un antes y un después en el campo de investigación. Se hará también un breve comentario al ranking de modelos con mejor rendimiento en el benchmark de LibriSpeech[8].

# 2 Modelos híbridos

Los modelos híbridos basados en HMM han sido los más efectivos en el Reconocimiento del Habla durante muchos años. Estos modelos hacen una clara distinción en dos partes: modelo acústico y modelo de lenguaje. Cada uno de estos modelos es entrenado de manera independiente y se juntan para formar el modelo completo. El modelo acústico reconoce la secuencia de fonemas que más se acerca a la señal acústica y la pasa al modelo de lenguaje para que este identifique que cadena de palabras perteneciente al lenguaje es más probable dada esta secuencia de fonemas.

Los modelos acústicos suelen consistir de HMM. Las probabilidades de transición de estos HMM han sido modeladas por mixturas de gaussianas la mayoría de las veces, pero vemos que ha habido una transición muy rápida en los últimos años hacia modelos de redes neuronales. Estos últmos se conocen como HMM-DNN. Un punto importante de estos modelos es que necesitan de datos de entrenamiento alineados, es decir que el texto correspondiente al audio este alineado con las secciones a las que pertenece. Esto es un inconveniente, ya que añade un paso extra en el preparado de datos para los modelos

Hemos visto suceder lo mismo en los modelos de lenguaje, modelos que tradicionalmente se entrenaban con n-gramas ahora son reemplazados por modelos más complejos como RNN o Transformers.

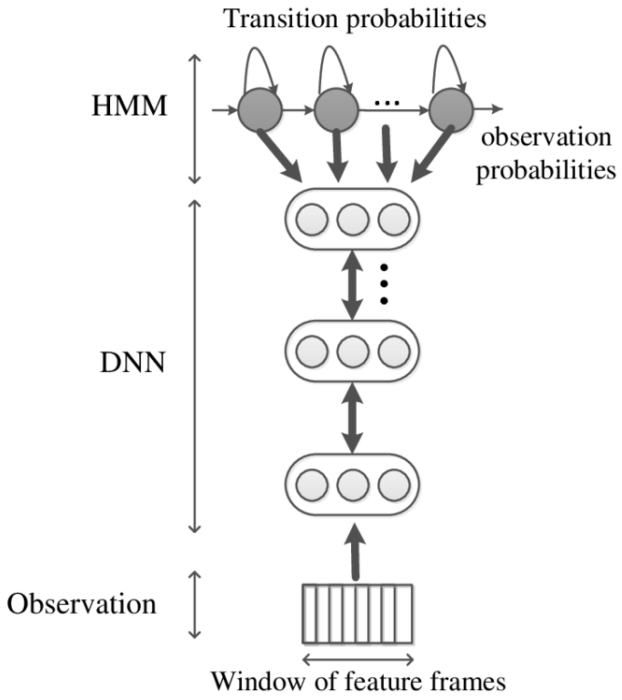


Figure 1: Esquema de arquitectura HMM-DNN

Sin embargo, estos modelos siguen teniendo gran importancia en el estado del arte de ASR, dado que su sencillez de entrenamiento frente a los end-to-end y sus buenos resultados (figura 6) los presentan como una buena opción.

# 3 Modelos end to end

En oposición a los modelos híbridos basados en HMM surgen los modelos end-to-end Los modelos end-to-end reciben una señal de audio y devuelven la secuencia de palabras reconocida de este audio. Encapsulan todo el proceso de reconocimiento de habla en un solo modelo. La mayoría de modelos end-to-end cuentan con dos partes: encoder, que recibe la señal acústica y crea un vector de características; decoder, que recibe el vector de características y descifra la combinación de palabras asociadas a la señal acústica. Esta división no se encuentran en todos los modelos end-to-end pero es una de las arquitecturas más usadas, también tener en cuenta que al tratarse de un modelo end-to-end de carácter neuronal donde se entrena todo a la vez es difícil hacer una clara diferenciación de qué parte hace cada función.

En los últimos años han surgido distintas arquitecturas end-to-end y vamos a tratar tres de las más conocidas en las siguientes secciones.

## 3.1 Connectionist temporal classification

Los modelos CTC (Connectionist temporal classification) fueron propuestos por [4] y crean un cambio de paradigma al ser el primer modelo entrenable end-to-end basado en redes neuronales para reconocimiento del habla. La mayor contribución de estos modelos es la creación de una función de pérdida CTC que permite entrenar modelos sin datos alineados (frente a los modelos híbridos previamente explicados). El modelo propuesto genera una matriz de probabilidades de caracteres para cada instante temporal que representa la probabilidad de que un carácter este presente en el instante t.

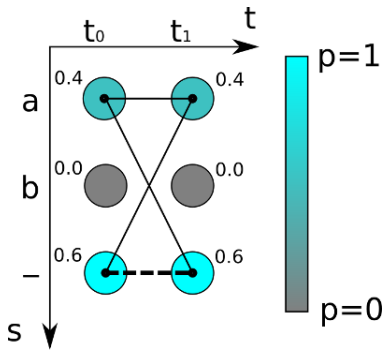


Figure 2: Matriz de salida del modelo neuronal CTC

La nueva función de pérdida calcula su valor sumando sobre la probabilidad de todos los alineamientos posibles (también llamados *paths*) y aplica el logaritmo negativo sobre el resultado de la suma. Cada uno de estos alineamientos es uno de los posibles caminos en la matriz de probabilidades que acabamos de ver. Podemos ver un ejemplo en la siguiente figura:

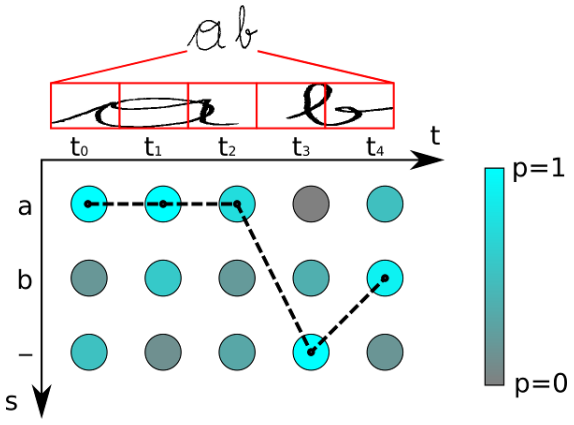


Figure 3: Alineamiento para la secuencia "ab".

Los modelos CTC usan distintas arquitecturas neuronales (LSTM, RNN, BLSTM, GRU) pero todos usan la función de pérdida previamente descrita. Todos estos modelos tienen dos puntos negativos:

- No son capaces de aprender dependencias entre tokens de la cadena de salida. La función de perdida de CTC trata cada elemento del la cadena de salida como independiente por lo que no es capaz de aprender relaciones entre los tokens que la componen. Es decir, no es capaz de aprender un modelo de lenguaje, es puramente un modelo acústico.
- Solo pueden crear secuencias de salida de misma longitud o menor que la señal acústica de entrada.

## 3.2 RNN-Transducer

Para solucionar los problemas que presentaban los modelos con función de perdida CTC Graves creo un nuevo modelo, RNN-Transducer[3]. Estos modelos siguen solucionando el problema de datos alineados, pero añaden una nueva forma de generar posibles alineamientos/*paths* y el cálculos de probabilidades de los mismos.

Los RNN-Transducer se basan en tres componentes: Encoder, Prediction Network y Joint Network.

- Encoder: Esta parte juega el rol de modelo acústico y dada una secuencia acústica $X$ crea un vector de características $F$.
- Prediction Network: Forma parte del rol de decoder y modela las dependencias entre los tokens de la secuencia de salida. Este componente es el gran añadido frente a modelos CTC
- Joint Network: Se encarga de crear el alineamiento entre la secuencia de entrada y la de salida.

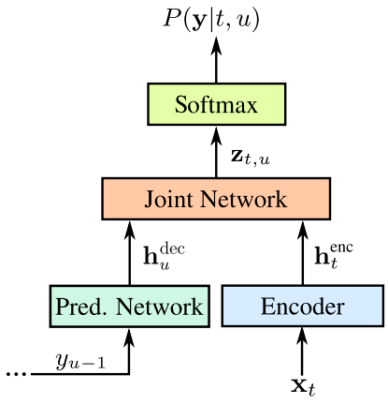


Figure 4: RNN-Transducer architecture

Con esta nueva arquitectura se consiguen tres grandes avances:

- Esta nueva arquitectura es capaz de generar secuencias de salida de mayor longitud que las de entrada.
- Tiene un componente específico para modelar dependencias entre los elementos de la secuencia de salida y se basa en redes RNN que ayudan a encontrar dependencias entre elementos.
- Al contar con componentes que simulan el modelo acústico y modelo de lenguaje pero que son encapsulados en una sola red esta arquitectura es la primera que consigue un entrenamiento completo de ambas funcionalidades en un solo sistema.

### 3.3 Modelos basados en *attention*

En este caso no se presenta una arquitectura con el propósito de mejorar flaquezas de modelos anteriores sino que se trata un cambio estructural de la manera de tratar redes neuronales *seqtoseq*. Estas redes reciben secuencias como entrada al modelo y devuelven otra secuencia. Esta arquitectura no es exclusiva para reconocimiento del habla y está presente en otros problemas cono traducción automática o condensado de texto.

Los modelos basados en *attention* surgieron primero en el campo traducción automática con [11] y generan un salto en rendimiento tan alto que han cambiado el paradigma de modelos *seqtoseq*. No tardaron en ser aplicados en modelos de reconocimiento del habla y tenemos ejemplos de modelos CTC [7] y RNN-Transducer[10] con esta capa de *attention* implementada.

La capa de *attention* suele implementarse entre el encoder y el decoder del modelo. Esta capa actúa como ayuda para el decoder, indicándole a qué parte

del vector de características creado por el encoder debería darle más importancia (*attention*) para generar la cadena de salida

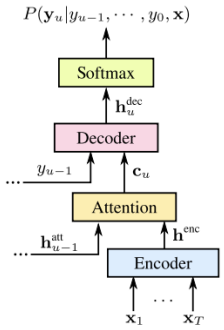


Figure 5: Capa de *attention* entre el encoder y el decoder [9].

Esta capa de atención ha generado grandes avances en modelo neuronales en los últimos años y dado pie a modelos basados en estas capas como Transformers donde se usan varias de estas de forma consecutiva.

# 4 Estado del arte - LibriSpeech

En esta sección nos gustaría mostrar el estado del arte con los mejores modelos listados públicamente en el ranking [1] de test-other para LibriSpeech. Estos modelos han sido evaluados con WER (Word Error Rate) y hemos excluido aquellos que usaban datos extra para entrenamiento, ya sean otros conjuntos de datos o *data augmentation*.

| Ranking | Modelo | WER |
|---|---|---|
| 1 | Conformer | 3.9% |
| 2 | wav2vec 2.0 | 4.1% |
| 3 | ContextNet 2.0 | 4.1% |
| 4 | CTC + Transformer LM 2.0 | 4.2% |
| 5 | Transformer Transducer 2.0 | 4.2% |
| 9 | Hybrid + Transformer LM rescoring | 4.85% |

Figure 6: Ranking WER en test-other LibriSpeech

En primer lugar, comentar cómo de cerca están todos los modelos del top 5 y que el modelo hibrido con mejores resultados esta muy arriba en la tabla (noveno puesto). Pasamos ahora a explicar brevemente los modelos del ranking:

- El Conformer [5] se trata de un modelo que combina los Transformers (modelos basados en *attention*) junto con redes convolucionales.

- Wav2vec 2.0 [2] está también basado en Transformers pero trata el problema con un enfoque distinto, haciendo uso de datos no supervisados para extender el modelo original.
- ContextNet [6] es el modelo sin *attention* que mejores resultados obtiene, en este caso se trata de una arquitectura CNN-RNN-Transducer que combina convolucionales con RNN-Transducer.
- CTC + Transformer LM [13] combina el modelo CTC junto con un modelo de lenguaje basado en Transformers para suplir las desventajas de CTC.
- Transformer-Transducer [14] combina la arquitectura de Transformers con *attention* con RNN-Transducers.
- Hybrid + Transformer ML rescoring [12], donde se combina un modelo híbrido con un modelo de lenguaje basado en Transformers para hacer rescoring de las secuencias de salida.

Podemos ver que 5 de los 6 modelos analizados del ranking cuentan con algún tipo de *attention* en ellos. Ahora mismo este tipo de arquitectura es dominante no solo en Reconocimiento del Habla sino en otros campos también. Vemos que se usa en combinación con CTC y RNN-Transducers para exprimir al máximo el rendimiento de estos modelos o que se crean modelos nuevos como el Conformer donde se unen Transformers con convolucionales. Los modelos híbridos no se quedan atrás y también se juntan con estas arquitecturas de *attention* consiguiendo el noveno puesto.

# 5 Conclusiones

Este trabajo nos ha servido para indagar de una manera más profunda en los modelos end-to-end y ver como han tomado por completo con el top 5 de benchmarks famosos como LibriSpeech. En trabajos pasados de la asignatura implementamos arquitecturas end-to-end como DeepSpeech, pero estas no alcanzaban resultados cercanos a los vistos en el ranking, y nos gustaría poder implementar alguna de estas arquitecturas en el futuro ahora que contamos con más experiencia en toolkits de redes neuronales.

Viendo los rankings y los últimos artículos de investigación en Reconocimiento del Habla parece que el estado del arte esta dominado por modelos end-to-end de carácter neuronal. La investigación en Machine Learning solo hace que crecer y esta en constante cambio por lo que no nos sorprendería que lo descrito en este trabajo vuelva a quedar obsoleto en un par de años. Un ejemplo de esto son las capas de *attention*, las cuales revolucionaron múltiples campos de investigación y parecen dominar actualmente.

# References


[1] Speech recognition on librispeech test-other. https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-other. Accessed: 2021-02-19.

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[3] Alex Graves. Sequence transduction with recurrent neural networks. *CoRR*, abs/1211.3711, 2012.

[4] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA, 2006. Association for Computing Machinery.

[5] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition, 2020.

[6] Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context, 2020.

[7] Takaaki Hori, Shinji Watanabe, and John Hershey. Joint CTC/attention decoding for end-to-end speech recognition. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 518–529, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[9] Rohit Prabhavalkar, Kanishka Rao, Tara Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly. A comparison of sequence-to-sequence models for speech recognition. 2017.

[10] Zhengkun Tian, Jiangyan Yi, Jianhua Tao, Ye Bai, and Zhengqi Wen. Self-attention transducers for end-to-end speech recognition. *Interspeech 2019*, Sep 2019.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[12] Yongqiang Wang, Abdelrahman Mohamed, Due Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, and et al. Transformer-based acoustic modeling for hybrid speech recognition. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.

[13] Frank Zhang, Yongqiang Wang, Xiaohui Zhang, Chunxi Liu, Yatharth Saraf, and Geoffrey Zweig. Faster, simpler and more accurate hybrid asr systems using wordpieces, 2020.

[14] Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss, 2020.