Deep Learning for Speech Processing

Danna Gurari

University of Colorado Boulder

Spring 2022

Review

- Last week:
  - Efficient learning: curriculum learning
  - Efficient learning: active learning
  - Reinforcement learning

- Assignments (Canvas):
  - Final project outline due Friday
  - Final project video due in two weeks

- Questions?

Today's Topics

- Problem
- Applications
- Speech recognition evaluation
- Speech recognition models
- Video making tutorial

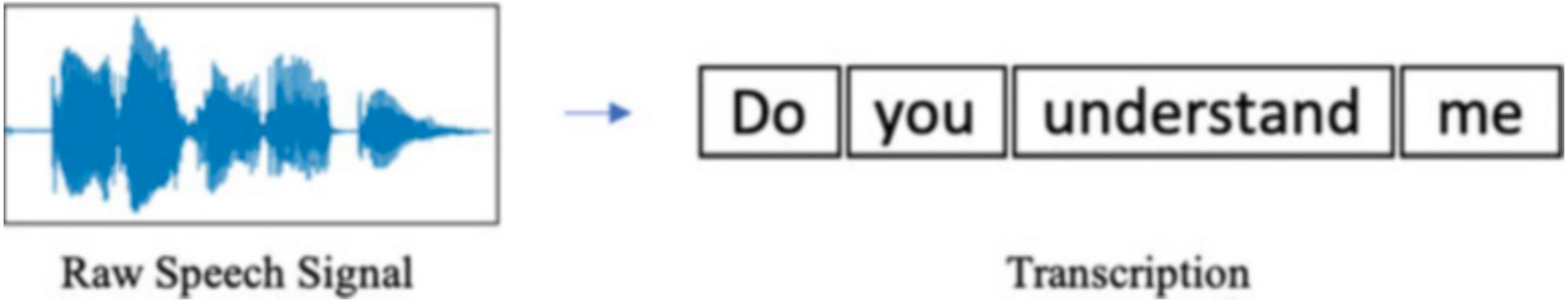
Today's Topics

- Problem
- Applications
- Speech recognition evaluation
- Speech recognition models
- Video making tutorial

Problem Definition

Input: spoken language

Output: machine readable text



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

What Is Speech?

Raw Speech Signal

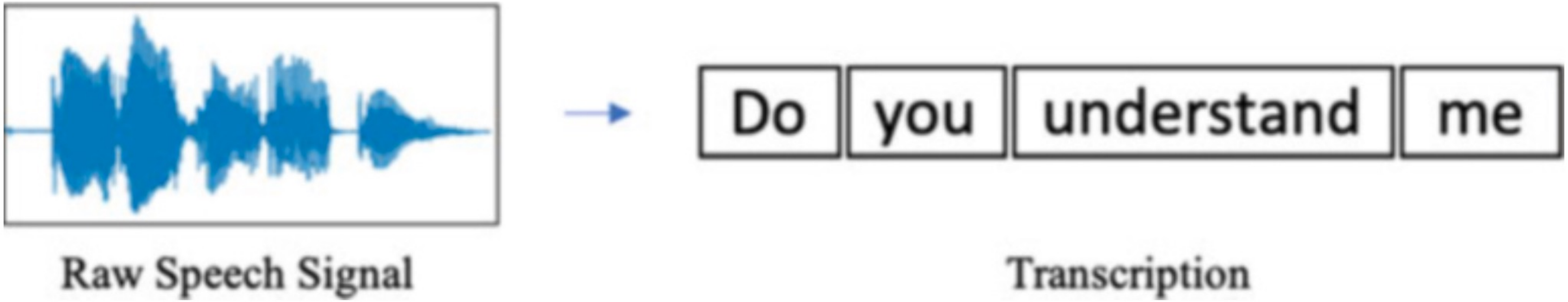
Compression waves created by pushing air from one's lungs and modulating it using one's tongue, teeth, and lips

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Why Is Speech Processing Challenging?

Input can be diverse including different accents, volumes, pace, and cadence

Temporal data needs to be segmented into distinct words



Technology can result in many artifacts including varying quality, echos, and background noise

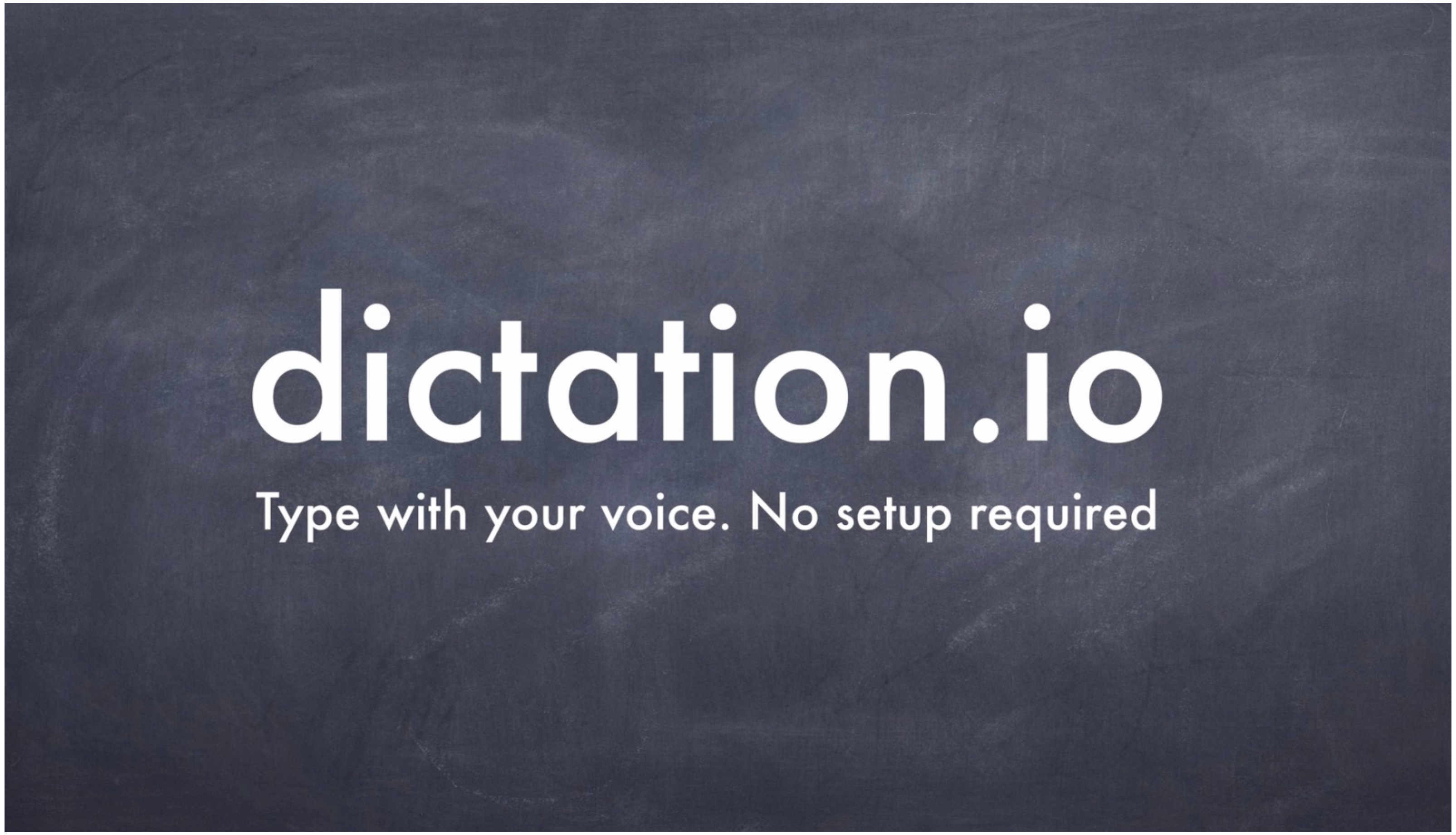
Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Today's Topics

- Problem
- Applications
- Speech representation
- Speech recognition models and evaluation
- Video making tutorial

Voice Typing on Mobile Devices

Voice Typing for Productivity Applications



Demo starting at 2:00: https://www.youtube.com/watch?v=5UK4vLzU9co&t=76s

Virtual Assistant

e.g., Amazon’s Echo with Alexa

e.g., Google Home

e.g., Baidu DuerOS

https://dueros.baidu.com/en/html/dueros/index.html

Virtual Assistant

Entertainment Video	Music, movie, television shows, variety show, short clip, audio book, and broadcasting, etc.	Information Inquiry	Cnsultation, weather, stocks, flight, sports (NBA), FAQ, cookbook, images, etc.
Lifestyle Services	Food, movie, take-out, hotel, shopping, taxi, cleaning service, travel, relaxation, and other O2O services.	Travel Conditions	Map, route, road condition, traffic restriction, endorsement, and surrounding environment query, etc.
Utility Tools	Ranslation, time, calculation, exchange rate, and unit conversion, etc.	Personal Assistant	Schedule management, alarm clock, reminder, memo and notepad, etc.
Learning	Encyclopedia, story, nursery rhyme, idiom, parenting, poetry and library, etc.	Chat and Relax	Chat, joke, poetry, idioms, and games, etc.

https://dueros.baidu.com/en/html/dueros/index.html

Audio Transcription (e.g., for Analysis & Situational/Permanent Hearing Impairments)

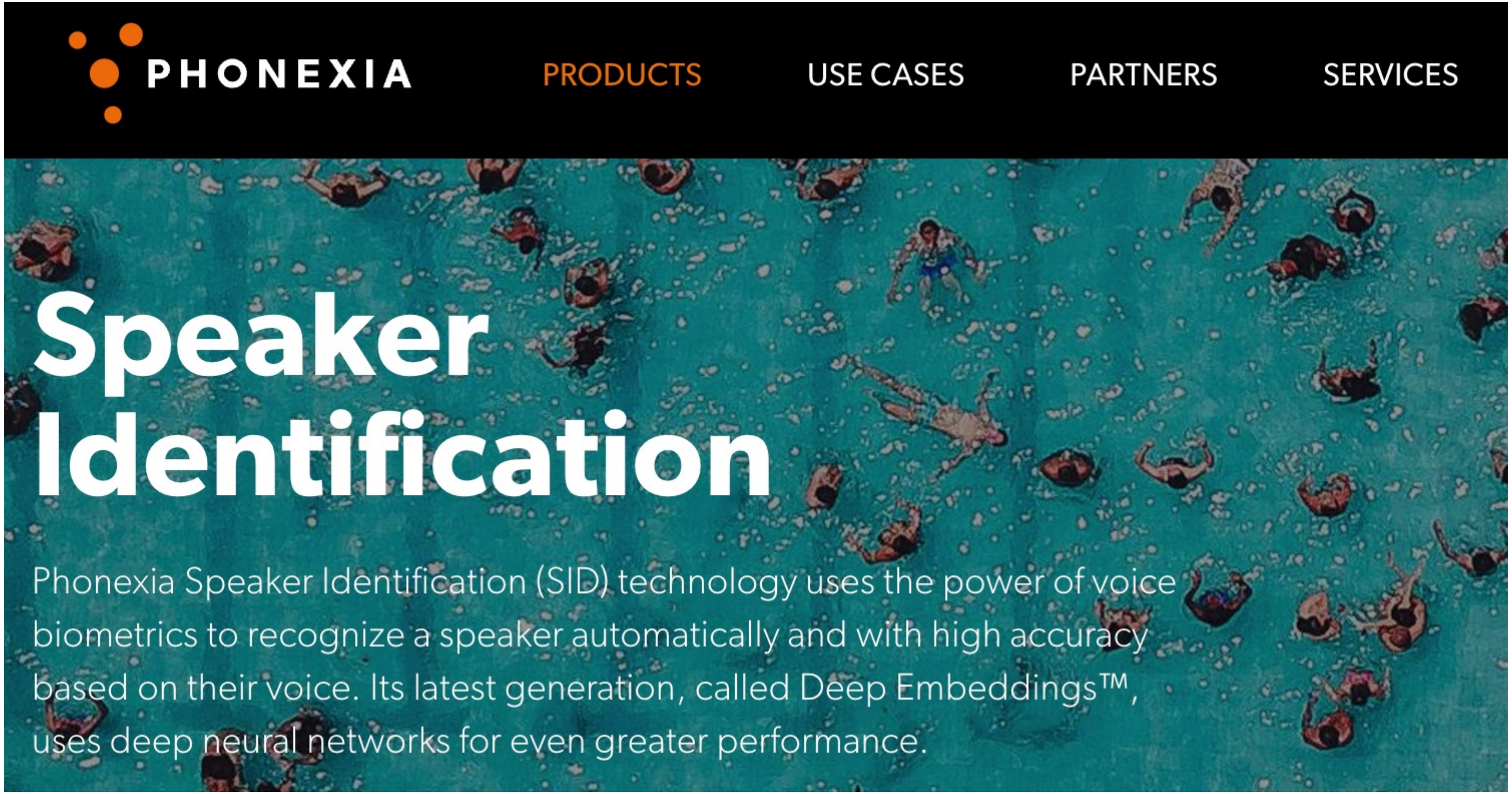
https://www.gmrtranscription.com/blog/podcast-transcription

Video/Movie Captioning (e.g., for Translation & Situational/Permanent Hearing Impairments)

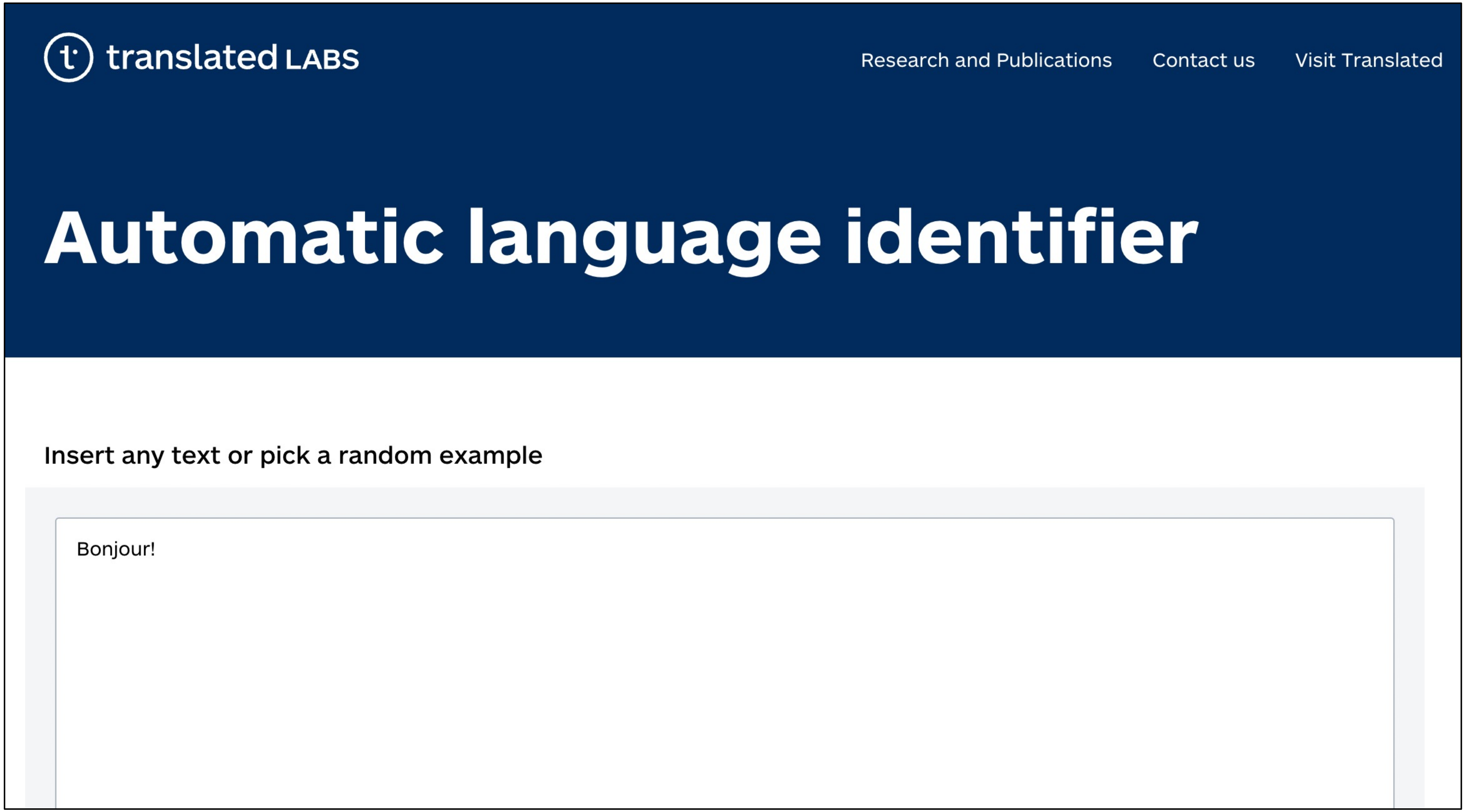
https://www.techsmith.com/blog/add-captions-subtitles-video/

Speech Emotion Recognition
(e.g., for Help Desks and Negotiators)

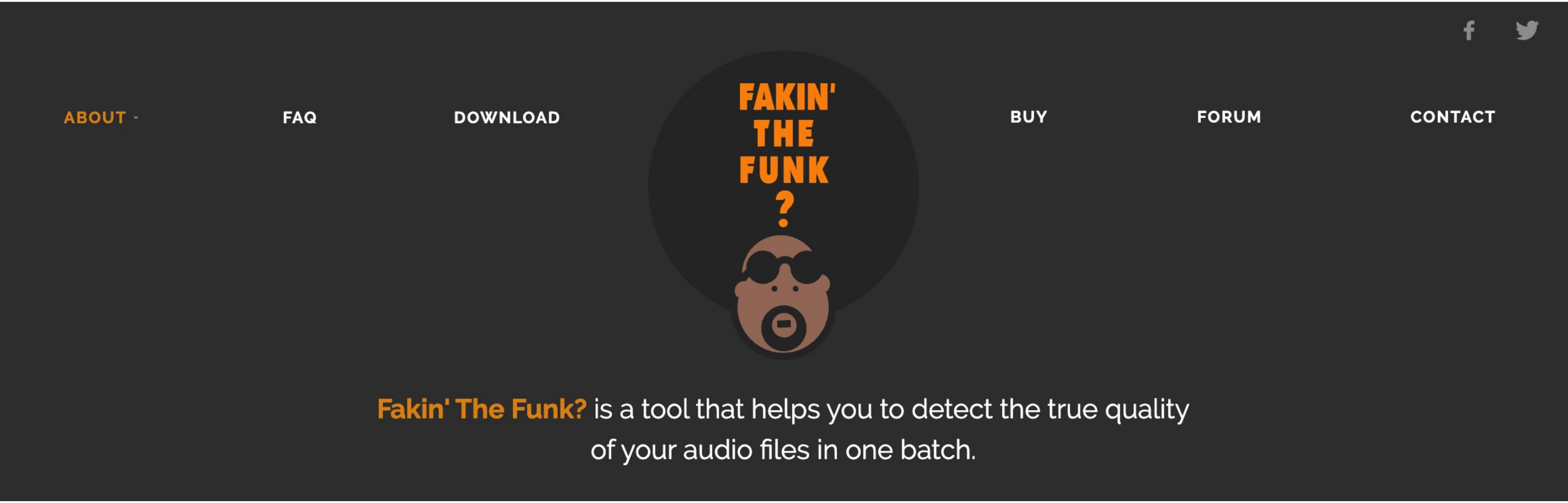
Speaker Identification (e.g., for Security)



Language Identification



Speech Enhancement



What are other potential applications for speech processing?

Today's Topics

- Problem
- Applications
- Speech recognition evaluation
- Speech recognition models
- Video making tutorial

Spectrum of Tasks

Commands

(brief with small vocabulary)

Discourse

(lengthy with large vocabulary)

Word Error Rate

- Indicates edit distance between the prediction and the target as follows:

$$\frac{\underbrace{\text{Substitutions}}_{\substack{\text{word replaced} \\ \text{(e.g., “goose”} \\ \text{transcribed as} \\ \text{“choose”)}}} + \underbrace{\text{Insertions}}_{\substack{\text{word added} \\ \text{(e.g., “caboose”} \\ \text{transcribed as} \\ \text{“the goose”)}}} + \underbrace{\text{Deletions}}_{\substack{\text{word omitted} \\ \text{(e.g., “it is a} \\ \text{fun” transcribed} \\ \text{as “it is fun”)}}}}{\text{Number of Words Spoken}}$$

- What indicates better performance: larger or smaller values?

Word Error Rate: Example

- Correct: The sun makes it look like uh a warm, day to go outside to adventure.
- Predicted: The son makes it to bike with a swarm to go outside to Denver today.
- Number of words spoken?
  - 15
- WER?

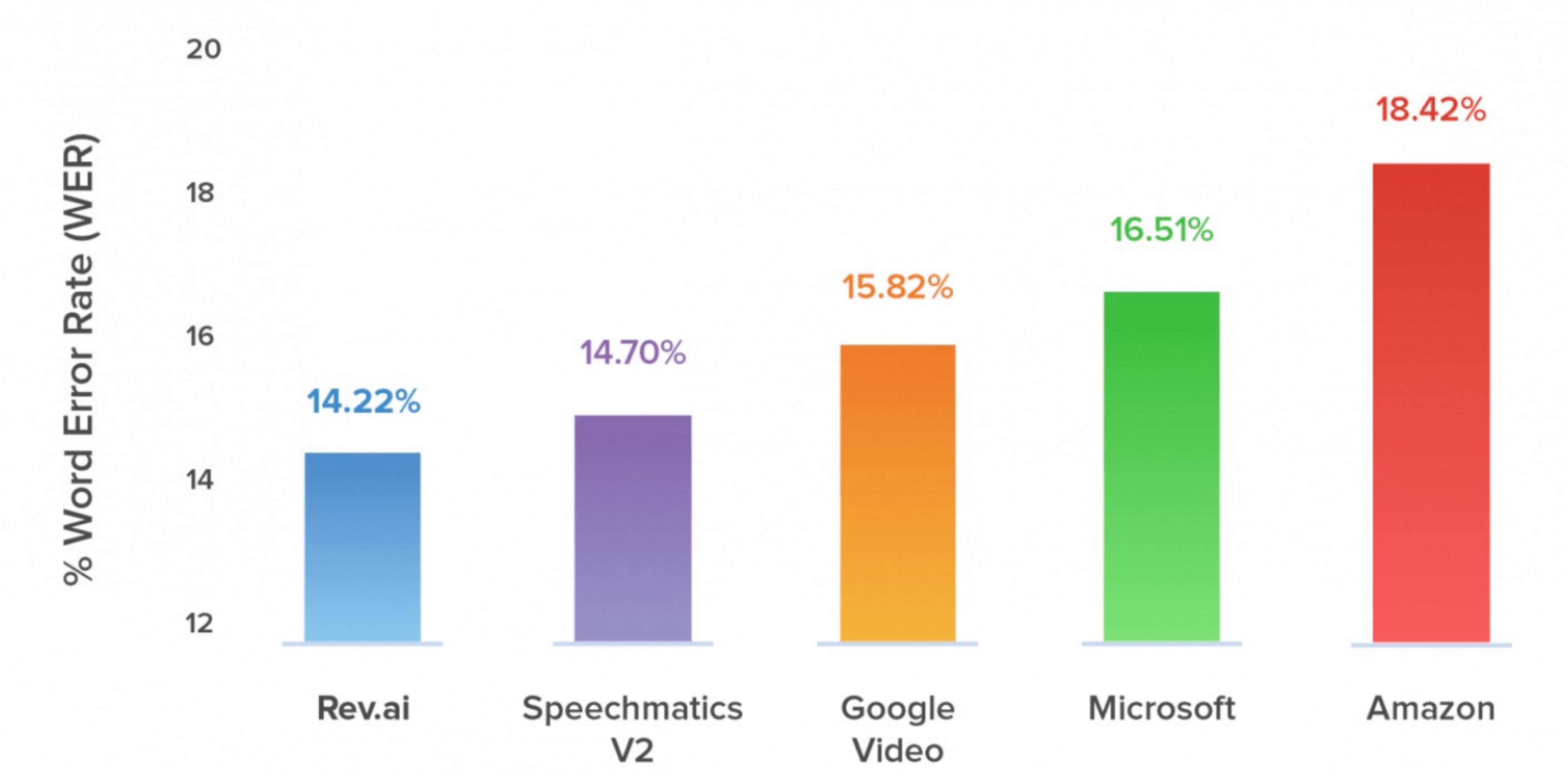
$$\frac{\text{Substitutions + Insertions + Deletions}}{\text{Number of Words Spoken}}$$

Word Error Rate: Example

- Correct: The sun makes it look like uh a warm, $\boxed{\text{day}}$ to go outside to adventure.
- Predicted: The $\boxed{\text{son}}$ makes it $\boxed{\text{to}}$ $\boxed{\text{bike}}$ $\boxed{\text{with}}$ a $\boxed{\text{swarm}}$ to go outside to $\boxed{\text{Denver}}$ $\boxed{\text{today.}}$
- Number of words spoken?
  - 15
- WER?

$$\frac{6 + 1 + 1}{15} = 0.53$$

Word Error Rate: Comparison Example



https://www.rev.com/blog/resources/what-is-wer-what-does-word-error-rate-mean

Word Error Rate: What Are Its Limitations as an Evaluation Metric?

- Does not indicate why errors occur
  - Background noise (e.g., music, other talking)
  - Specialized language (i.e., words reflecting domain expertise)
  - Speaker pronunciations/accent
- Does not reflect whether transcription correctly captures:
  - Capitalization
  - Punctuation
  - Numbers
  - Paragraphs
- May indicate poor quality when humans could understand the content
- Weights all word errors equally

Today’s Topics

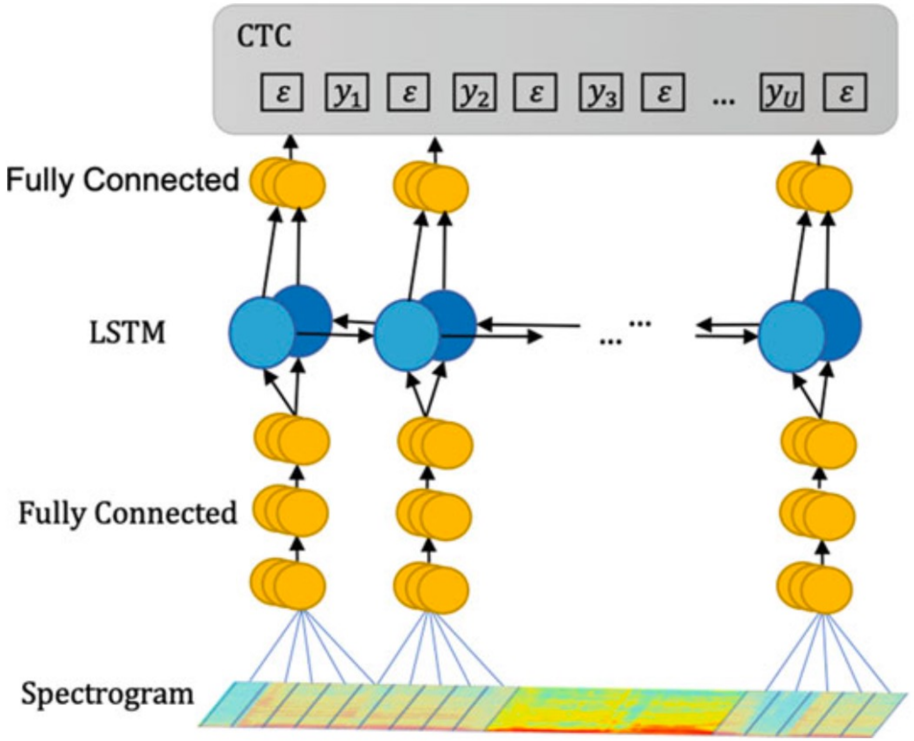
- Problem
- Applications
- Speech recognition evaluation
- Speech recognition models
- Video making tutorial

Popular Methods

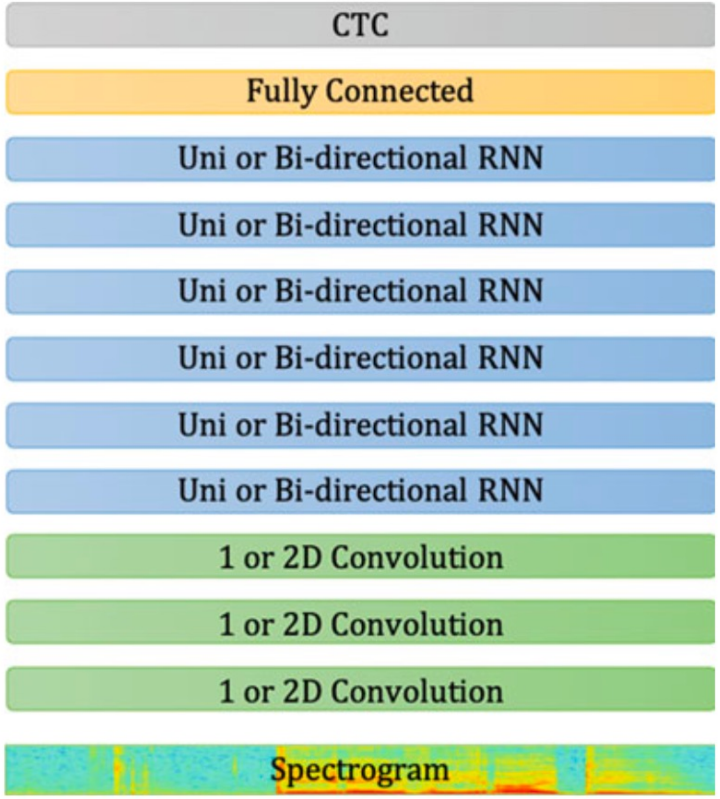
- Connectionist Temporal Classification (CTC)
- DeepSpeech
- DeepSpeech 2
- Listen, Attend, and Spell

Popular Methods

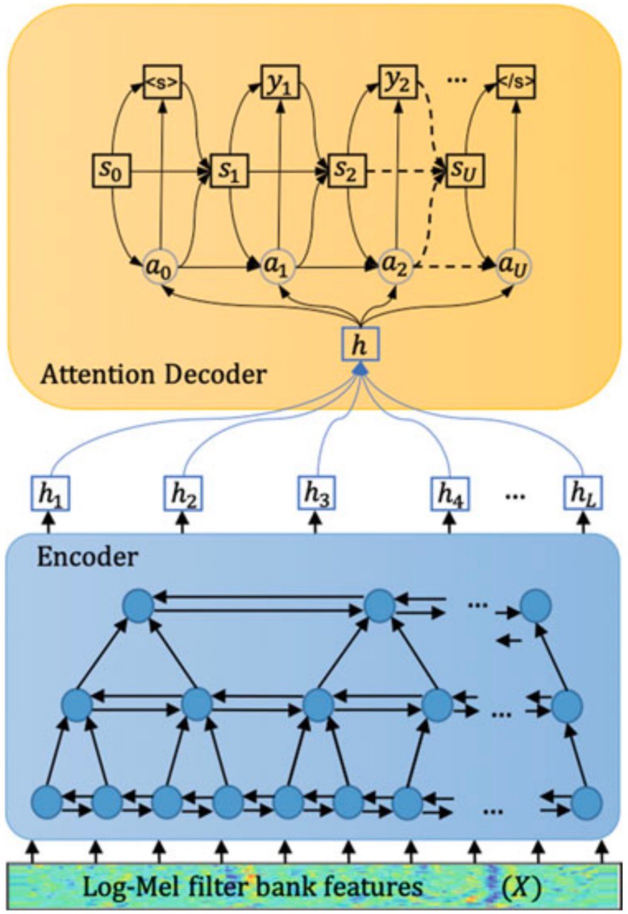
DeepSpeech



DeepSpeech2



Listen, Attend, and Spell

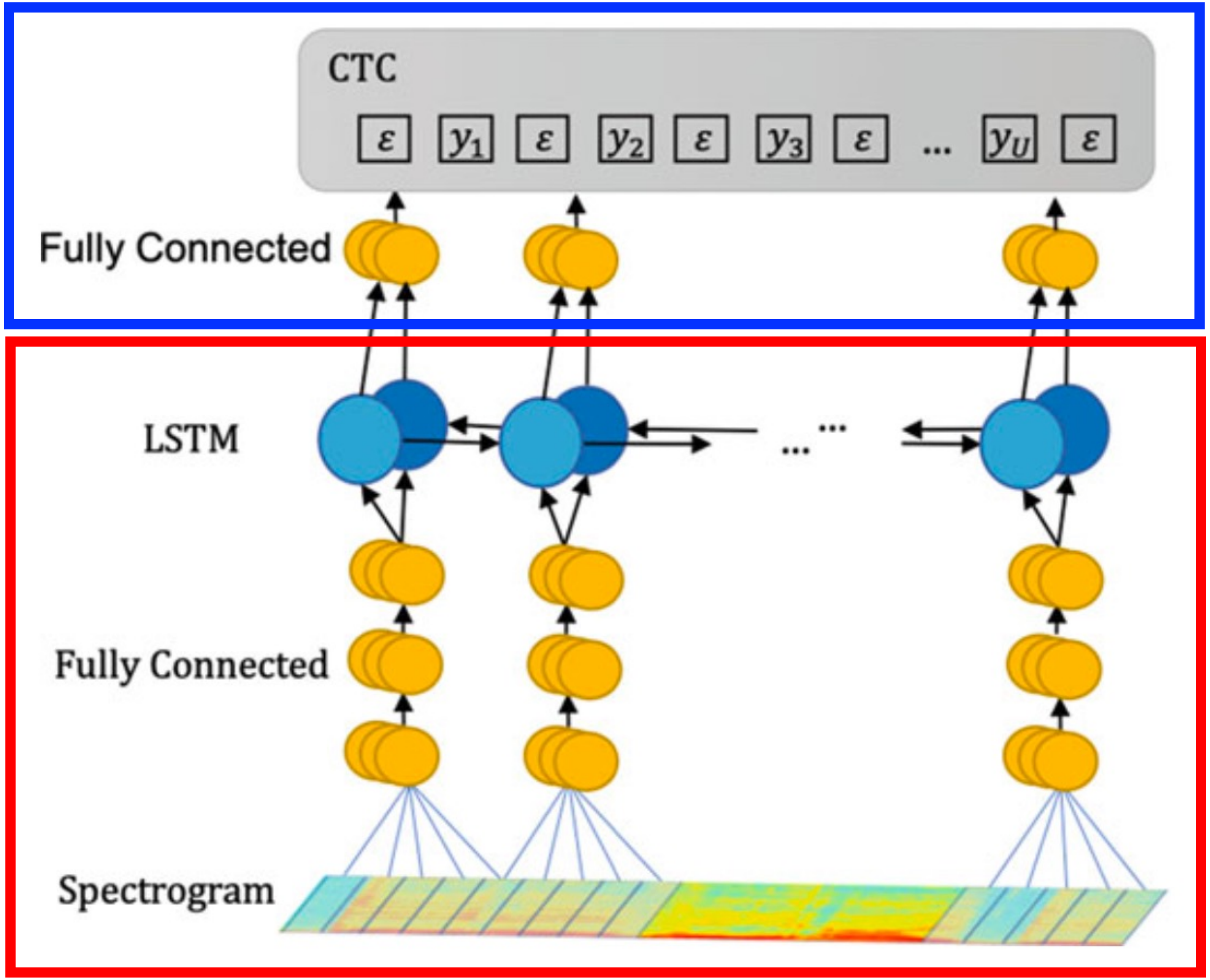


Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech

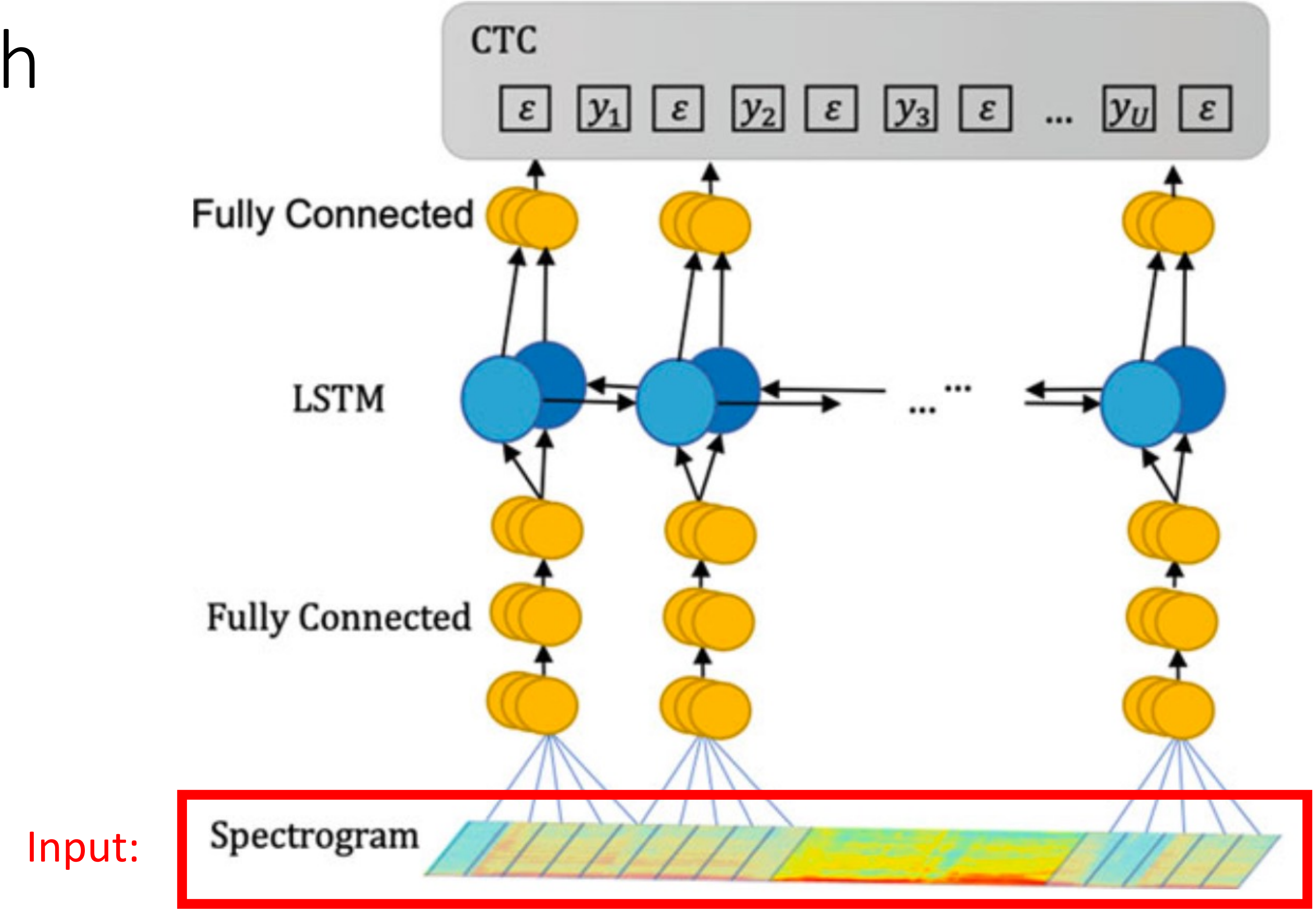
Decoder:

Encoder:



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

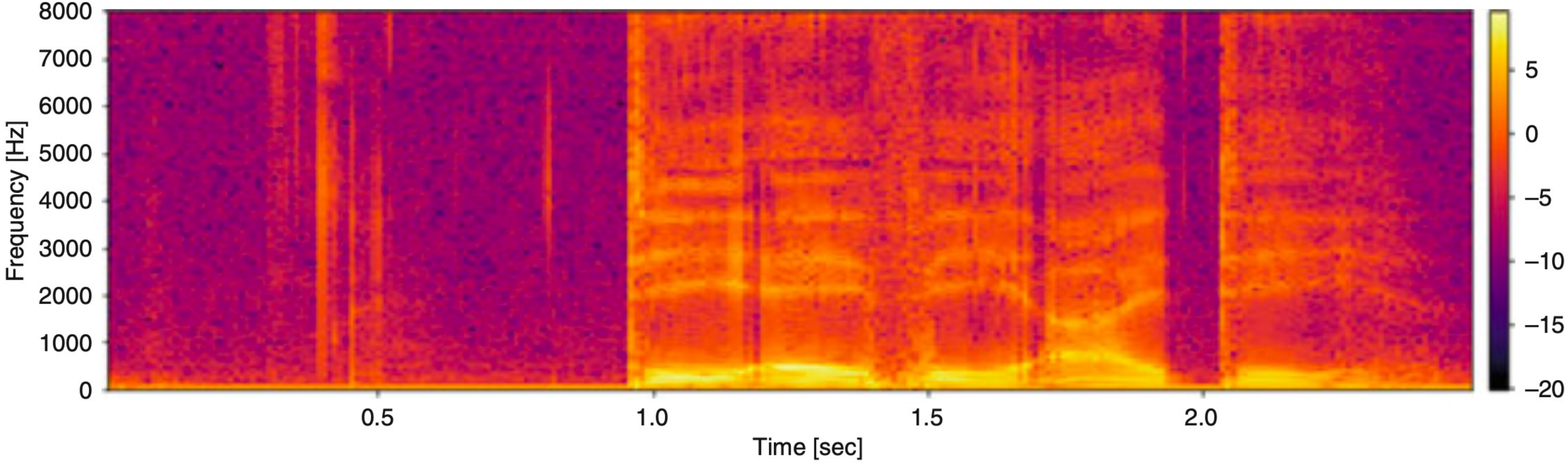
DeepSpeech



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Spectrogram: Visual Representation of Audio

Color: amplitude of frequency at a given time point

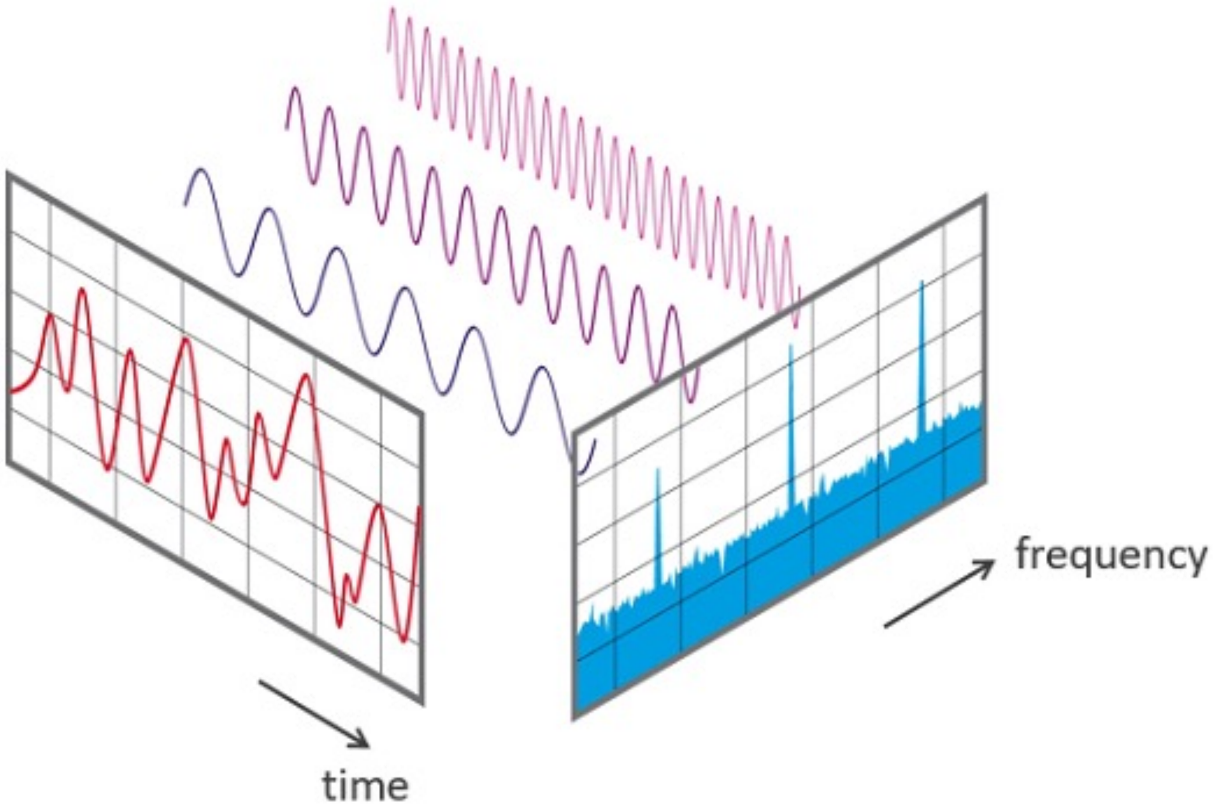


Created by sliding a short window across the audio signal and applying a Fourier transform to each window

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Background: Frequency Analysis of Audio Clip

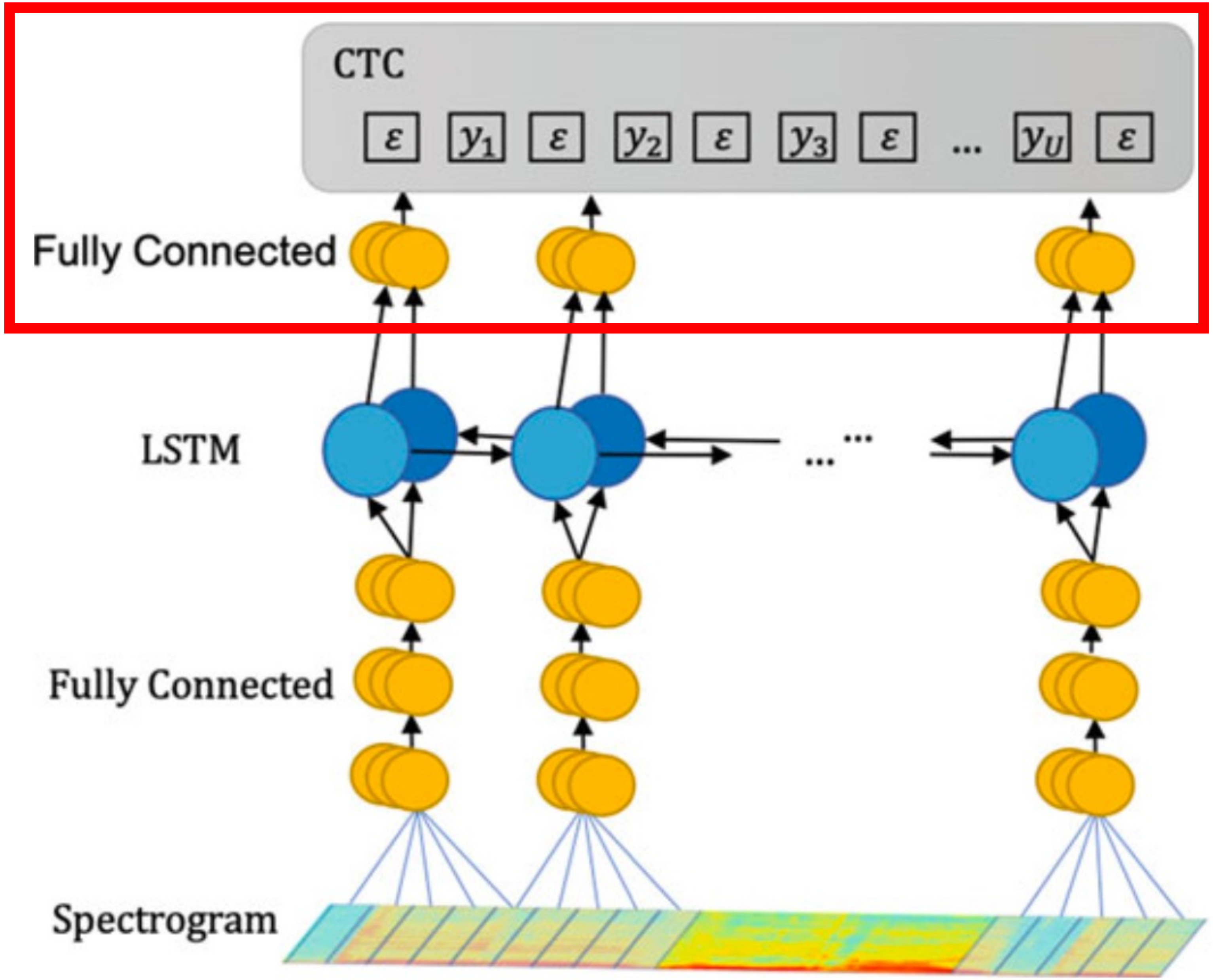
Fourier transform: represents a signal as a sum of sines and cosines (*frequency-domain*):



https://dev.to/trekhleb/playing-with-discrete-fourier-transform-algorithm-in-javascript-53n5

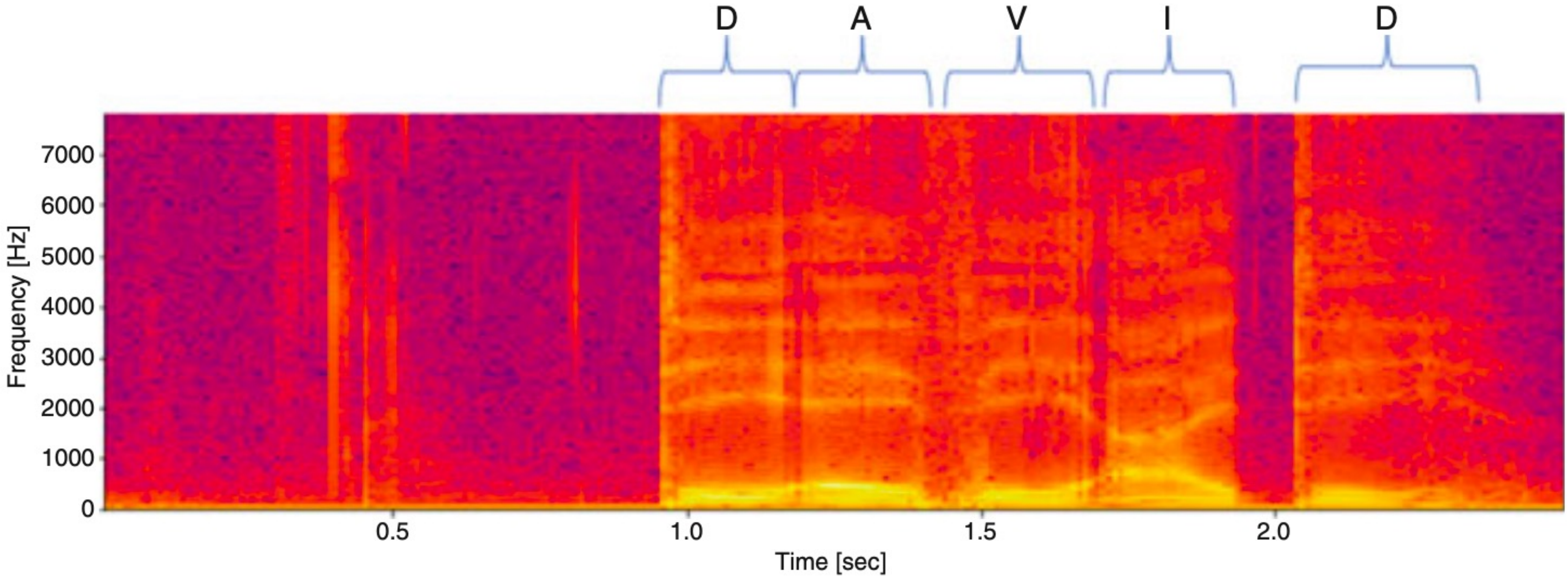
DeepSpeech

Output: character sequence predicted by a softmax layer



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

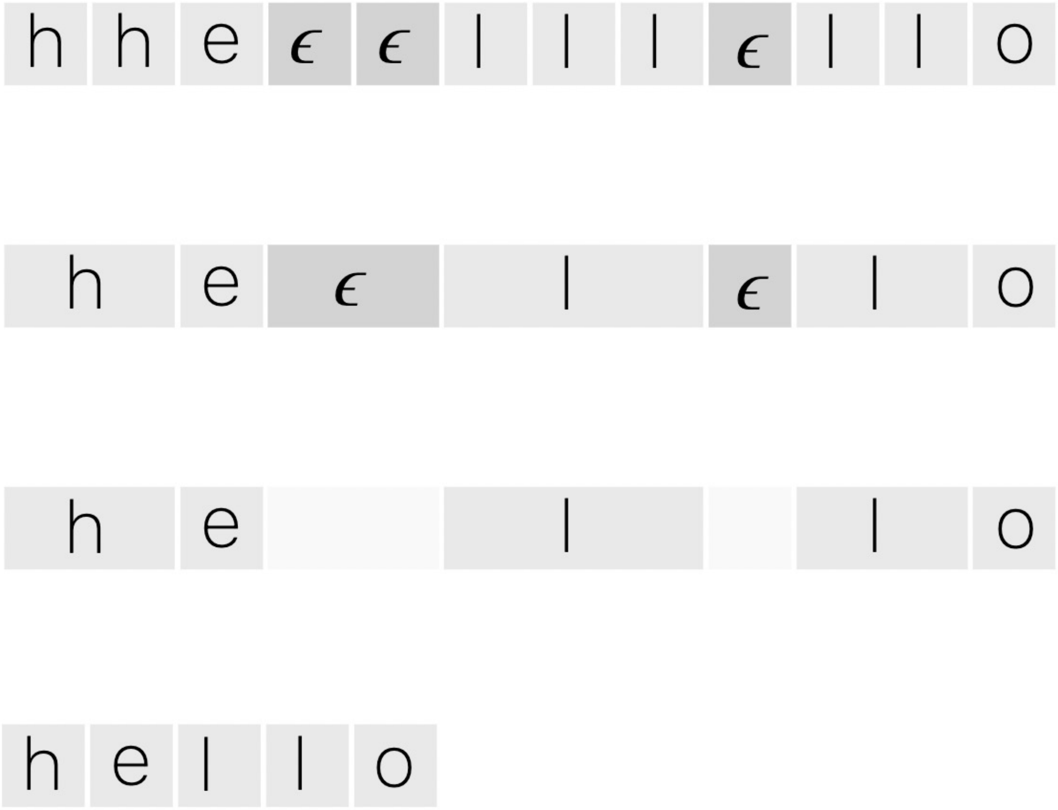
CTC: Input-Output Representation



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

CTC: Input-Output Representation

Key idea: blank token supports silent stretches and letter repeats (e.g., “hello” vs “helo”)



First, merge repeat characters.

Then, remove any ϵ tokens.

The remaining characters are the output.

https://distill.pub/2017/ctc/

CTC: Input-Output Representation

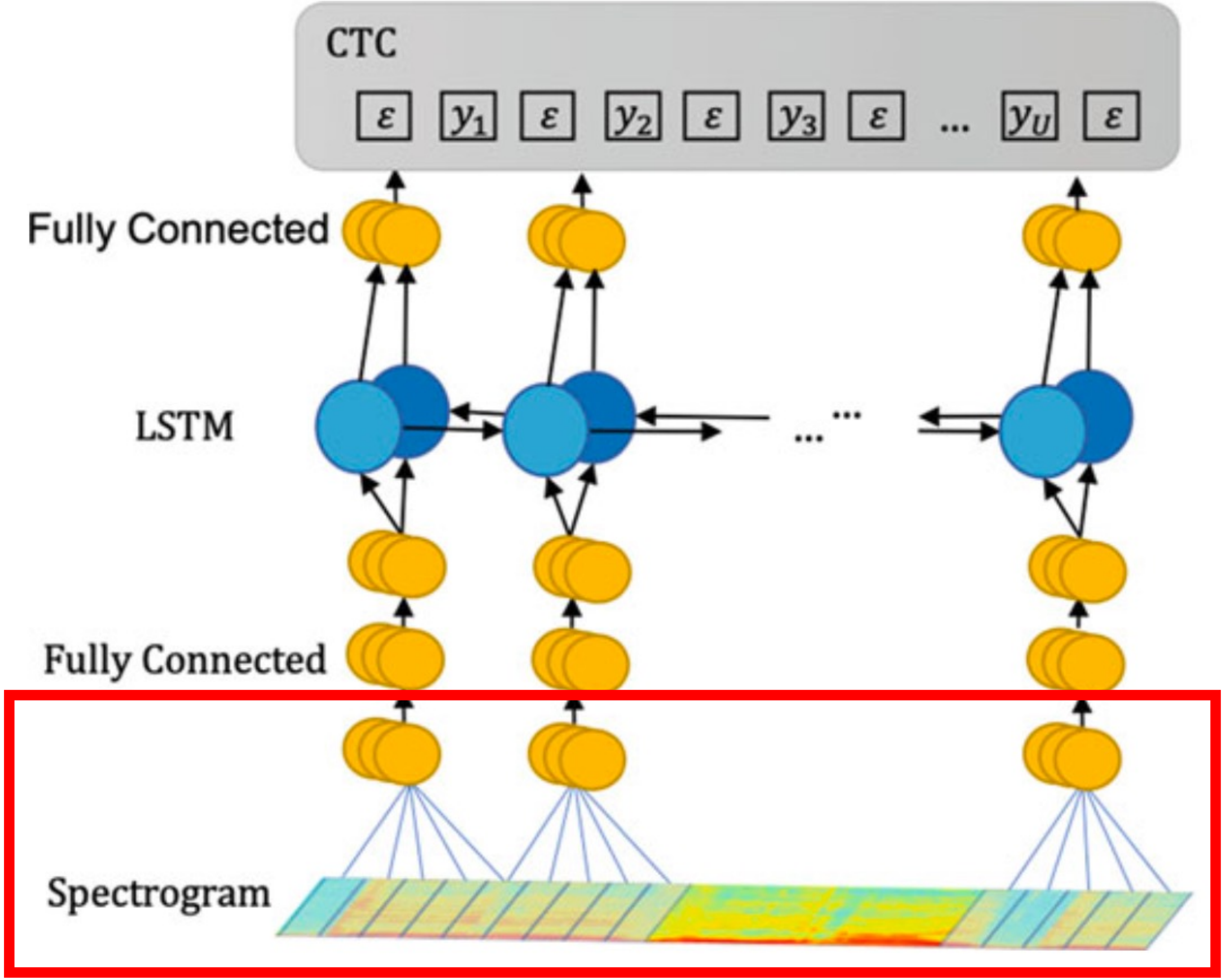
Key idea: blank token supports silent stretches and letter repeats (e.g., “hello” vs “helo”)

ϵ	c	c	ϵ	a	t
c	c	a	a	t	t
c	a	ϵ	ϵ	ϵ	t

Supports recognizing the same word when spoken differently!

https://distill.pub/2017/ctc/

DeepSpeech



First hidden layer looks at context around input:

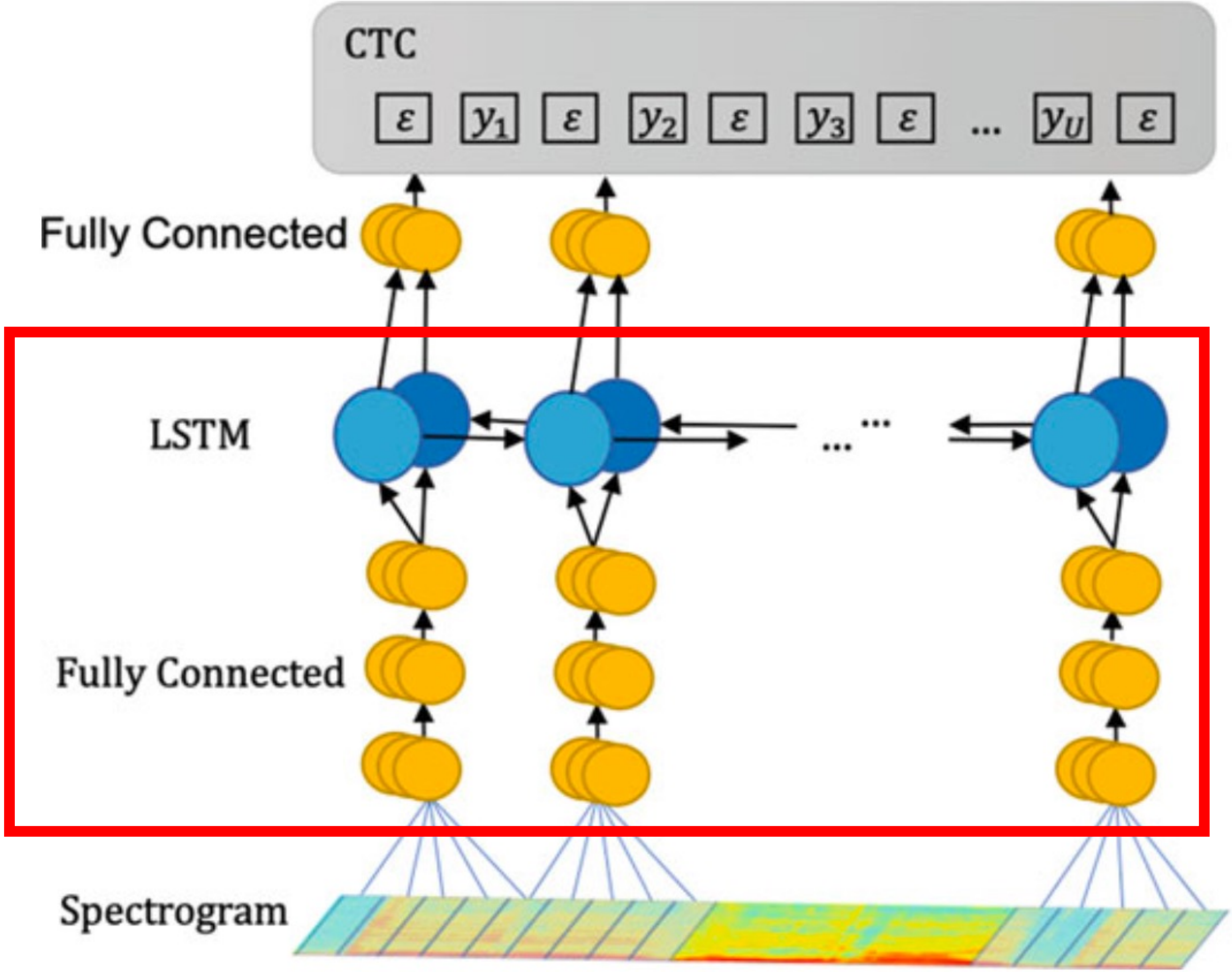
Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech

3 fully-connected layers followed by bidirectional LSTM:

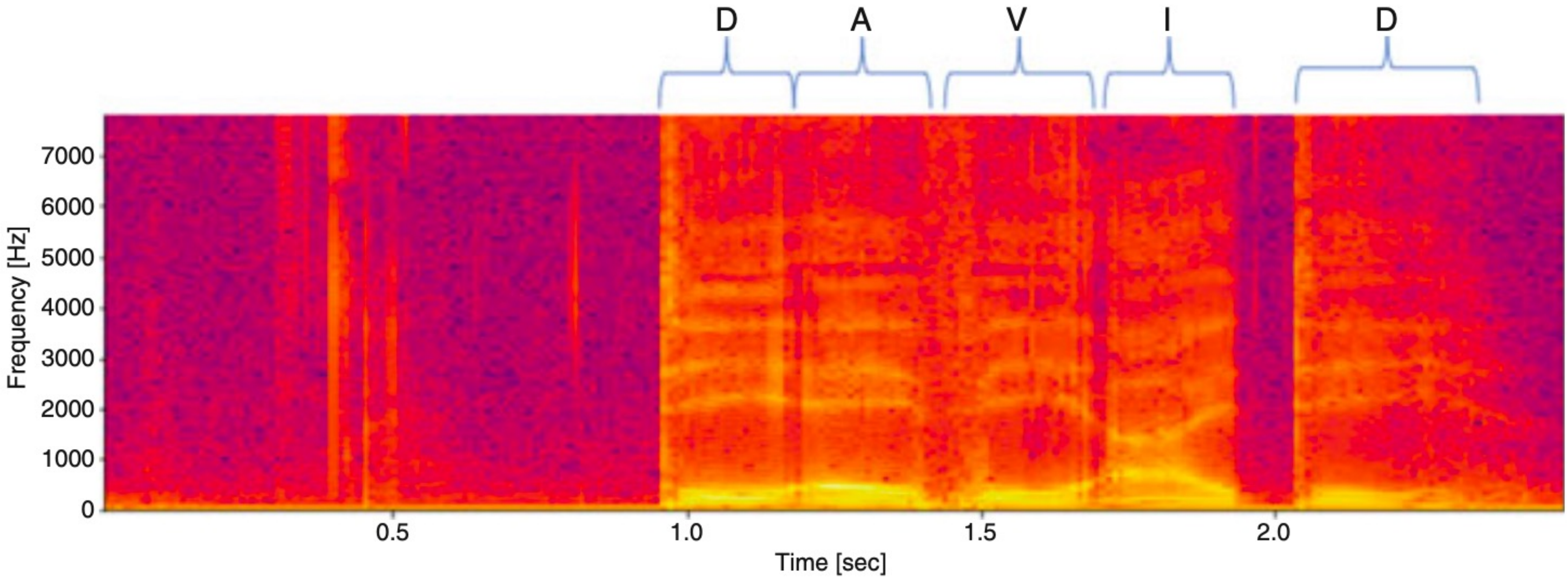
How is a bi-directional LSTM beneficial?

How is a bi-directional LSTM limiting?



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech: Optimization Function (CTC)



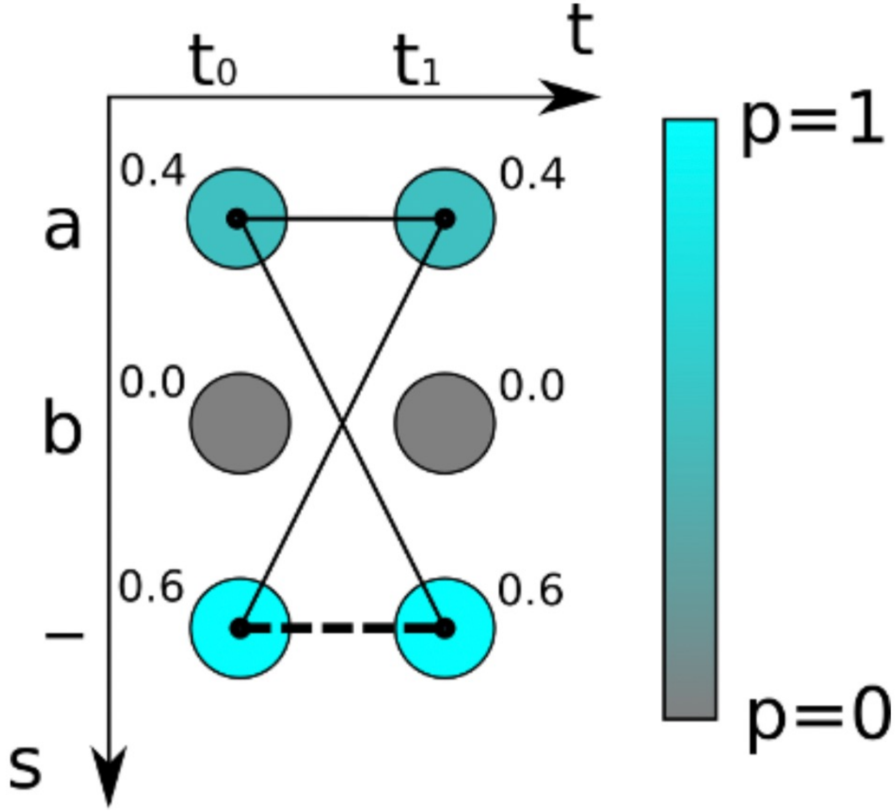
The CTC loss function enables learning output alignment without a per input label

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech: Optimization Function (CTC)

Most plausible from all possible alignments learned; e.g., 2 time steps with 2 potential characters and a blank token ("-")

- Probability of "a" is sum of all "a" representations
  - Probability of "aa"?
    - 0.4 x 0.4 = 0.16
  - Probability of "a-"?
    - 0.4 x 0.6 = 0.24
  - Probability of "-a"?
    - 0.6 x 0.4 = 0.24
  - Sum: 0.16 + 0.24 + 0.24 = 0.64

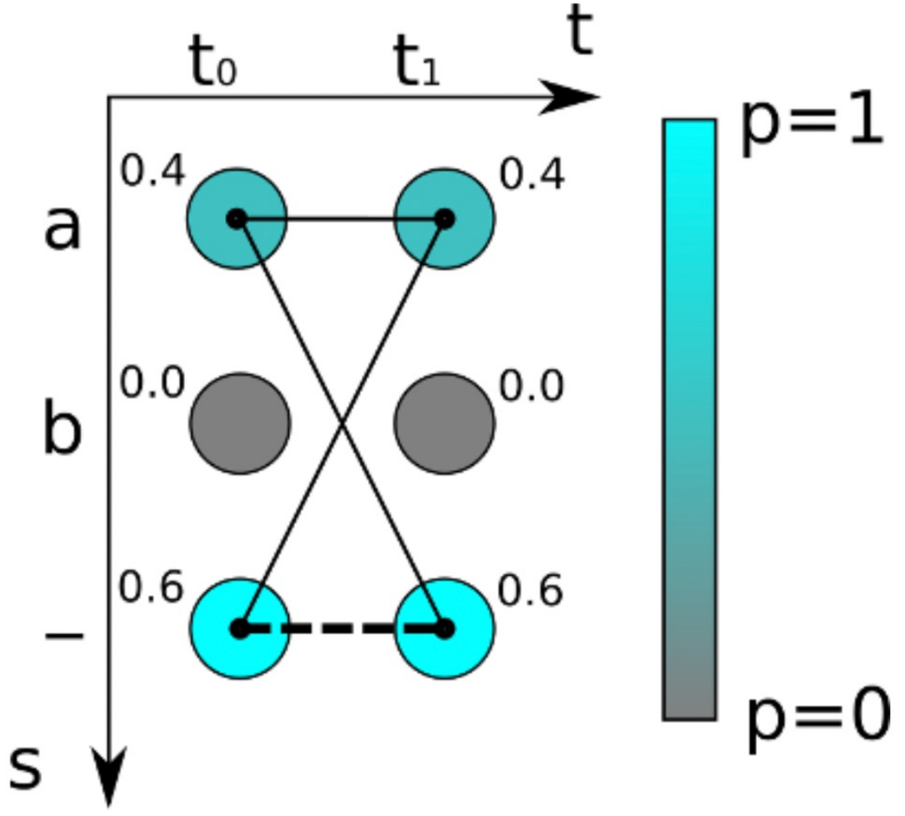


https://towardsdatascience.com/intuitively-understanding-connectionist-temporal-classification-3797e43a86c

DeepSpeech: Optimization Function (CTC)

Most plausible from all possible alignments learned; e.g., 2 time steps with 2 potential characters and a blank token ("-")

- Probability of "a": 0.64
- Probability of "" is sum of all "" representations
    - Probability of "--"?
        - 0.6 x 0.6 = 0.36

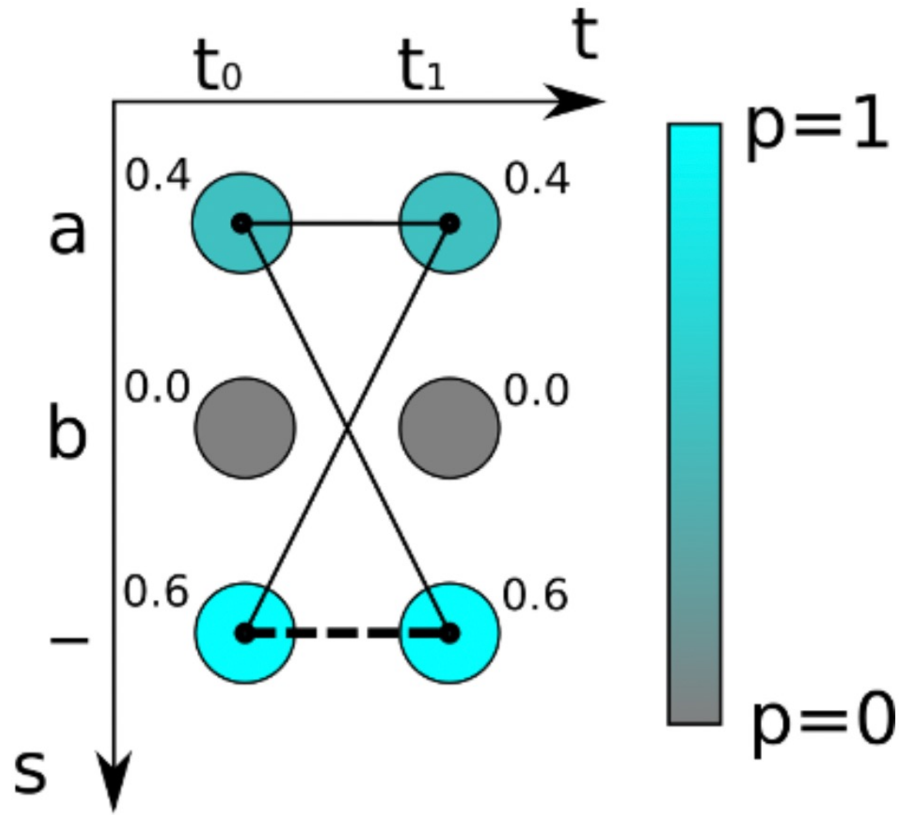


https://towardsdatascience.com/intuitively-understanding-connectionist-temporal-classification-3797e43a86c

DeepSpeech: Optimization Function (CTC)

Most plausible from all possible alignments learned; e.g., 2 time steps with 2 potential characters and a blank token ("-")

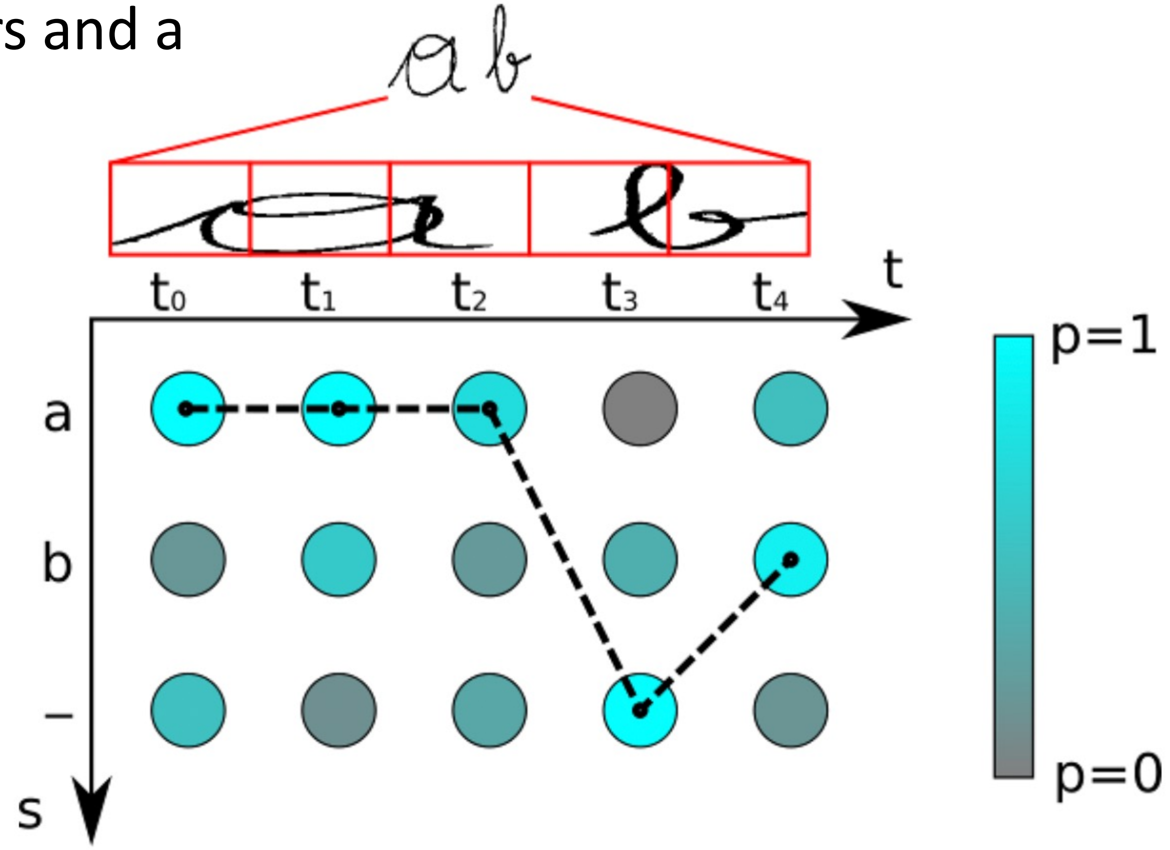
- Probability of "a": 0.64
- Probability of "": 0.36
- And so on for all possible alignments…



https://towardsdatascience.com/intuitively-understanding-connectionist-temporal-classification-3797e43a86c

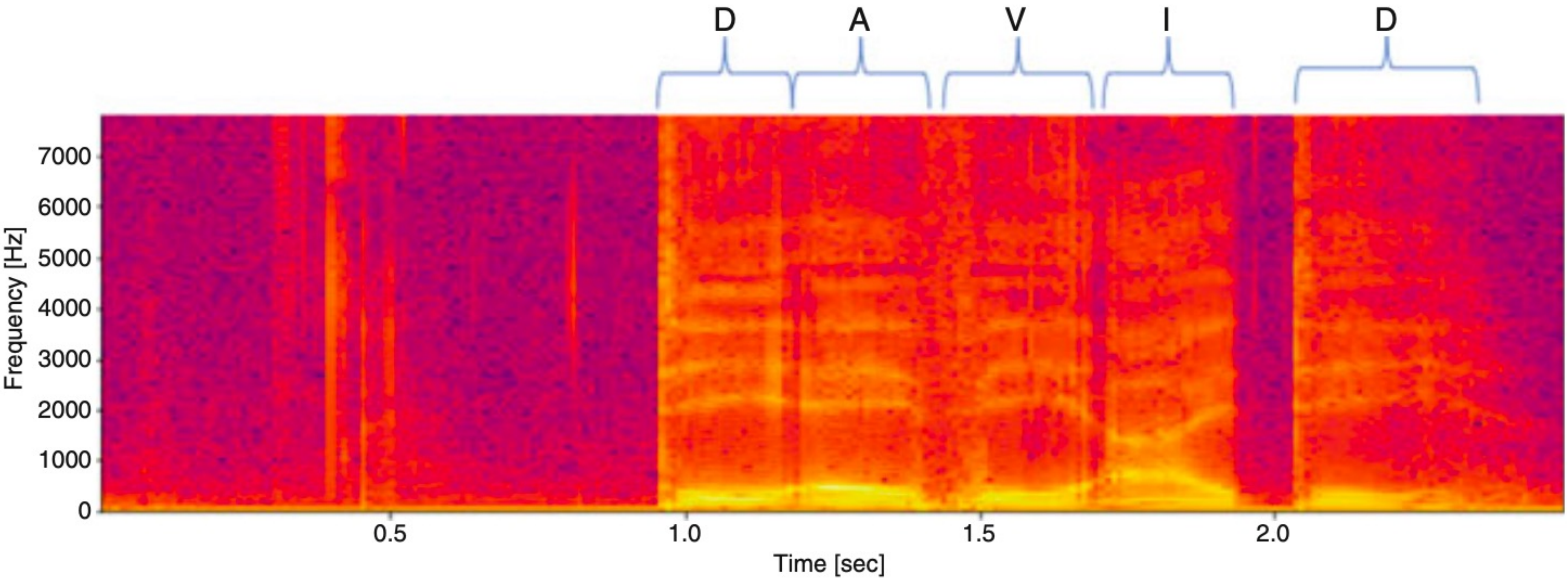
DeepSpeech: Optimization Function (CTC)

Most plausible from all possible alignments learned; e.g., 2 time steps with 2 potential characters and a blank token (“-”) with “best path decoding”



https://towardsdatascience.com/intuitively-understanding-connectionist-temporal-classification-3797e43a86c

DeepSpeech: Optimization Function (CTC)



CTC uses dynamic programming to accelerate computation and is differentiable

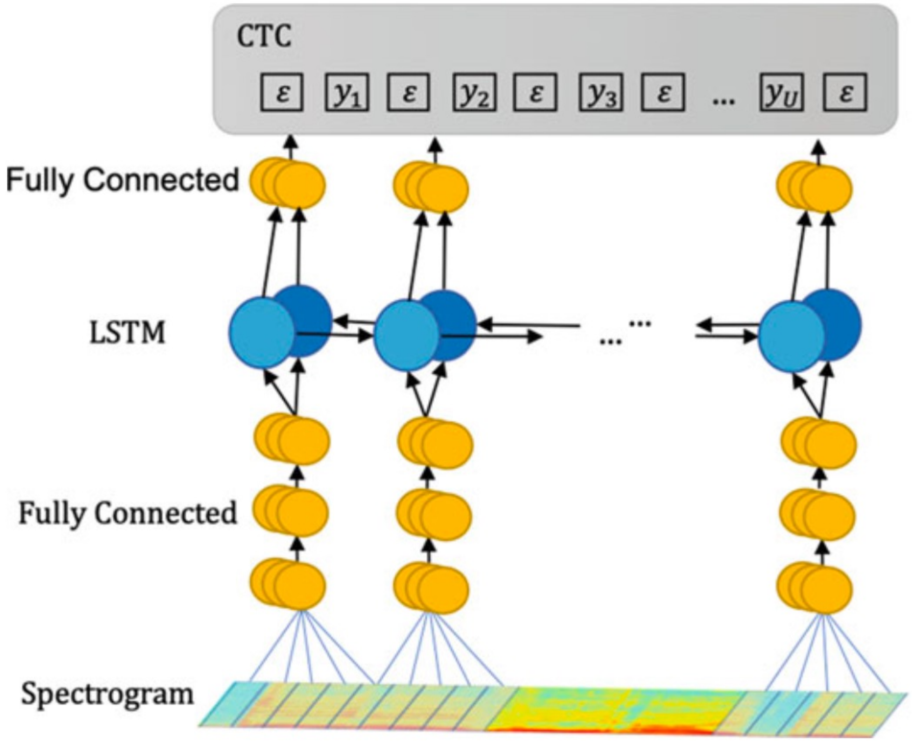
Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech: Training (Key Ideas)

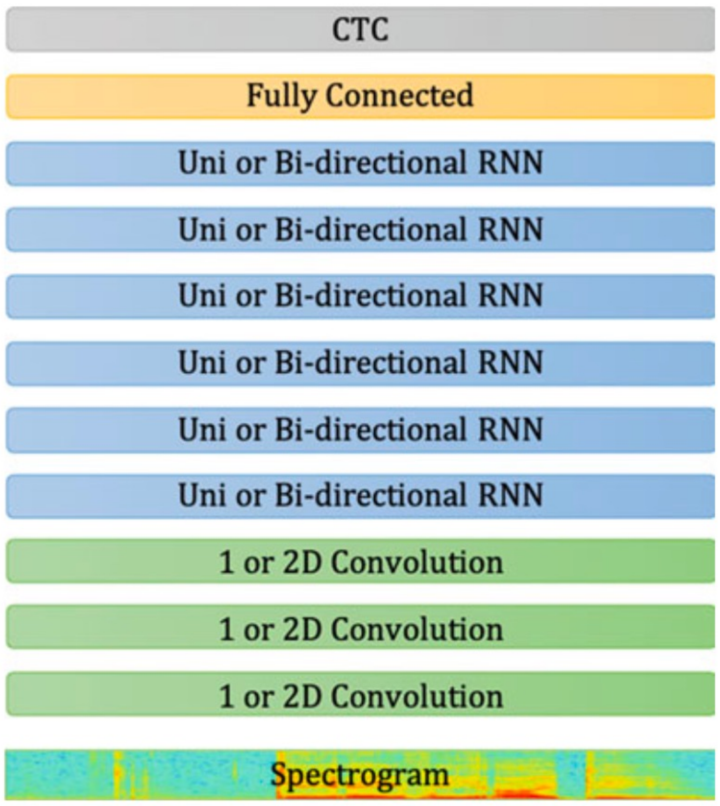
- 5000 hours from 9600 speakers
- Regularization
  - Dropout
  - Data augmentation: audio file translated 5 ms forward and backward
- Results boosted by incorporating a language model

Popular Methods

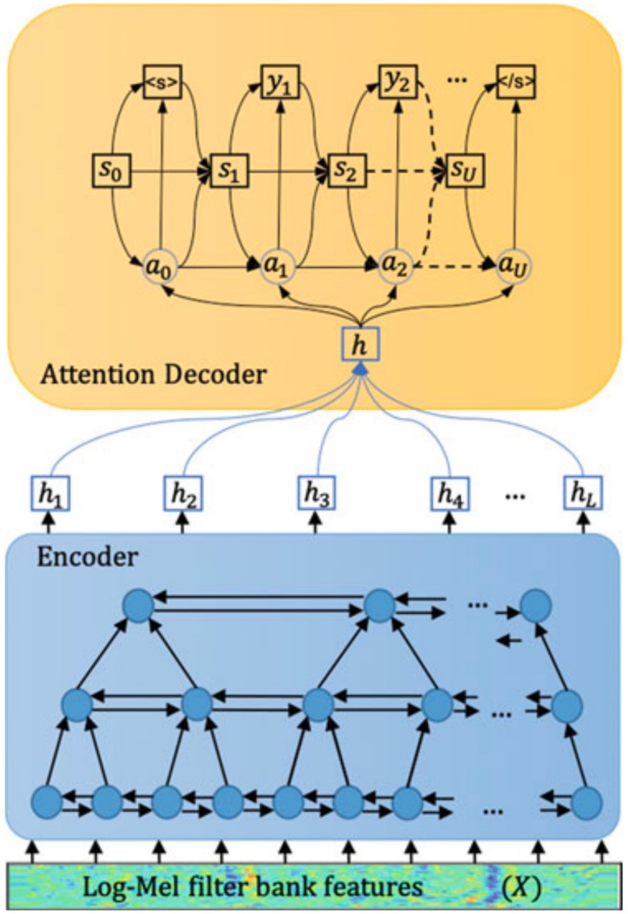
DeepSpeech



DeepSpeech2



Listen, Attend, and Spell



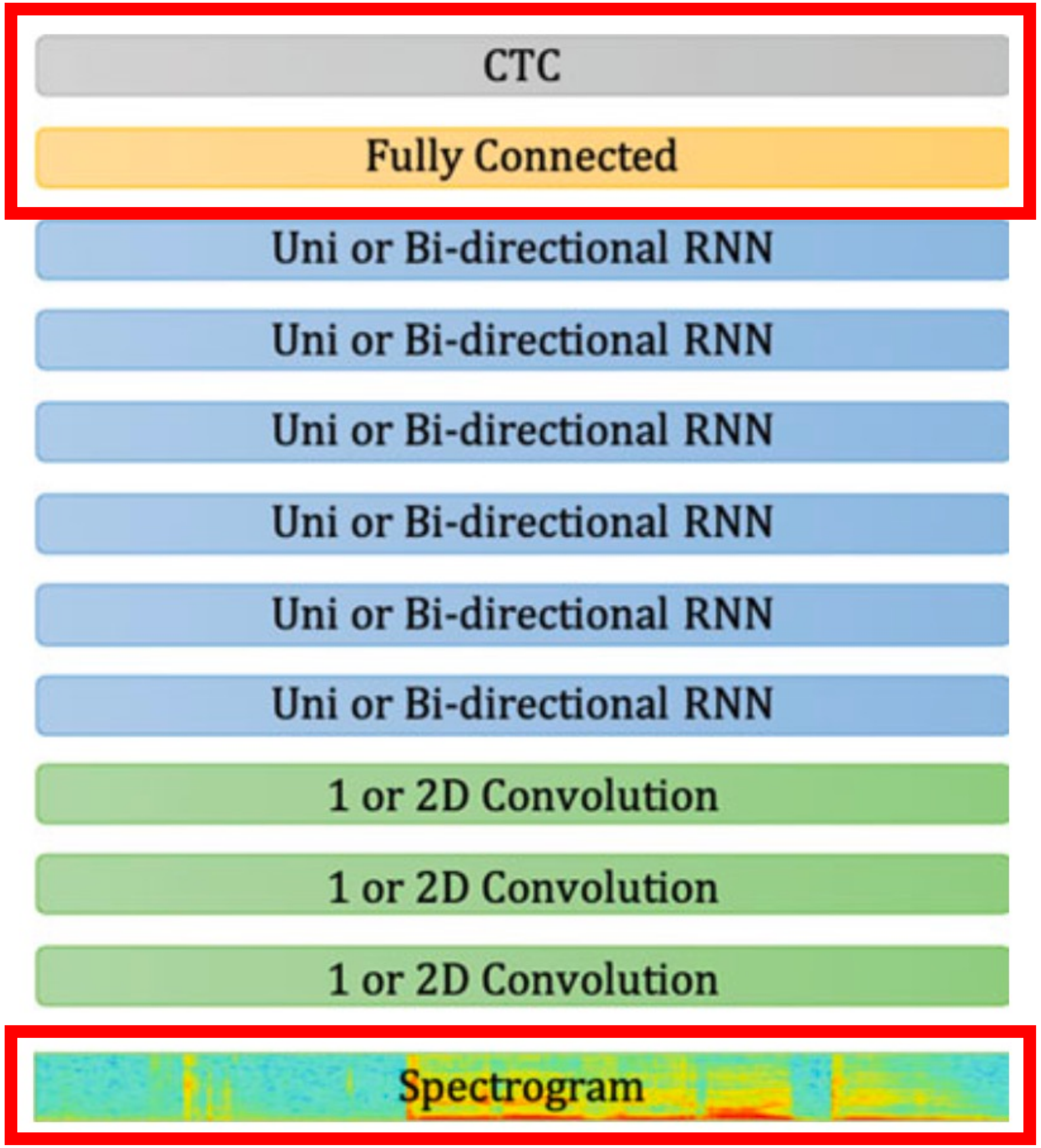
Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech2

Similar output:
(two architectures for
English and Mandarin)

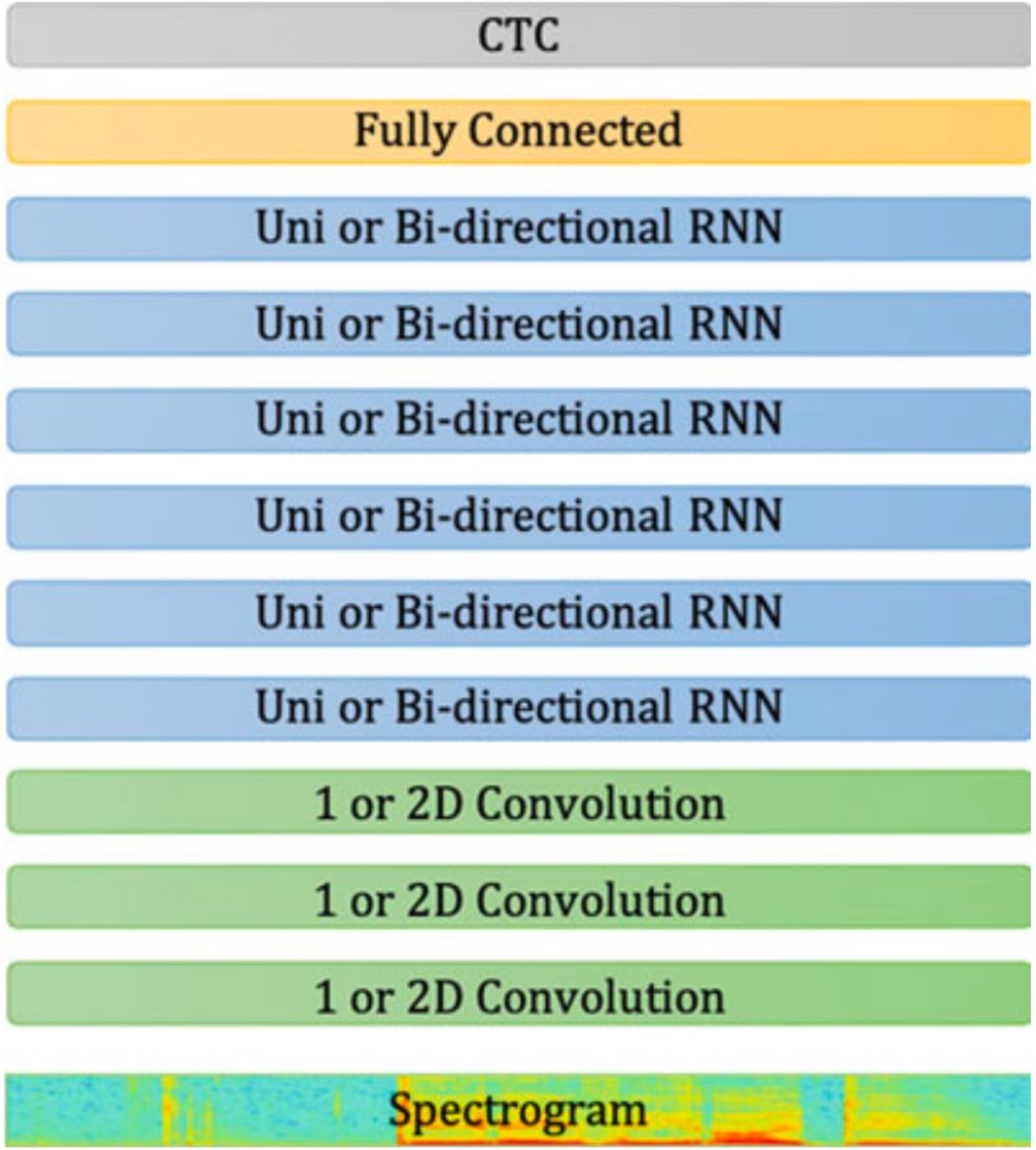
Extension of DeepSpeech that achieves a 7x speed-up and 43.4% relative WER improvement with a deeper architecture

Same input:



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech2



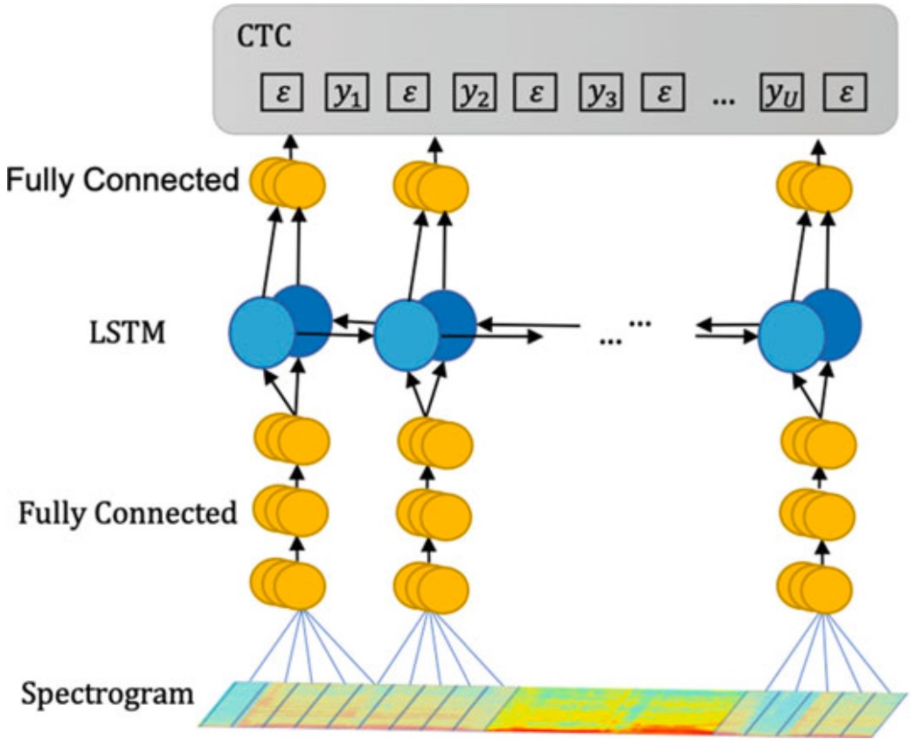
Training protocol difference from DeepSpeech:

- More training data (11,940 hours for English and 9,400 hours for Mandarin)
- Curriculum learning: trains based on length of utterances for first epoch with shorter ones first (improves WER by over 1 point)

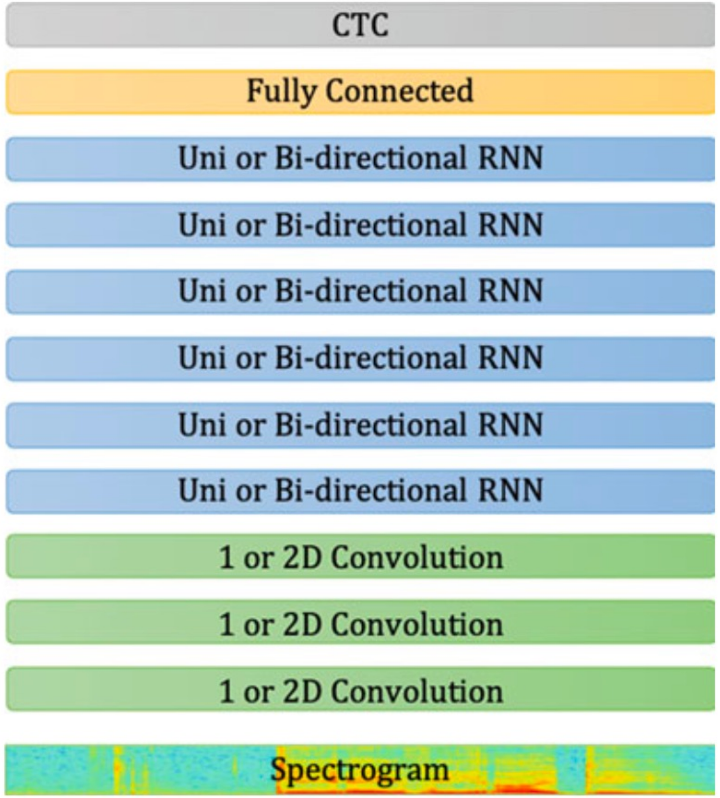
Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Popular Methods

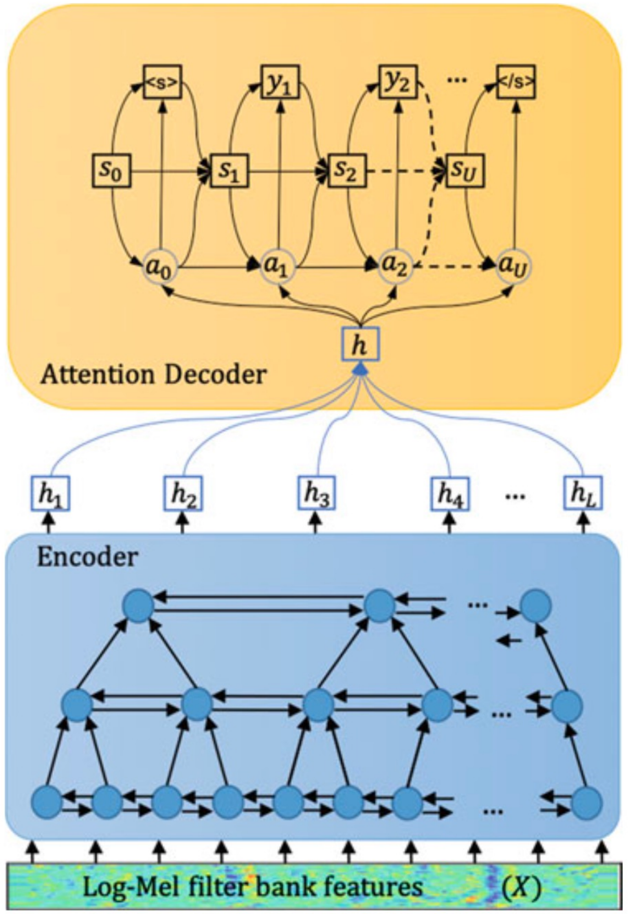
DeepSpeech



DeepSpeech2



Listen, Attend, and Spell

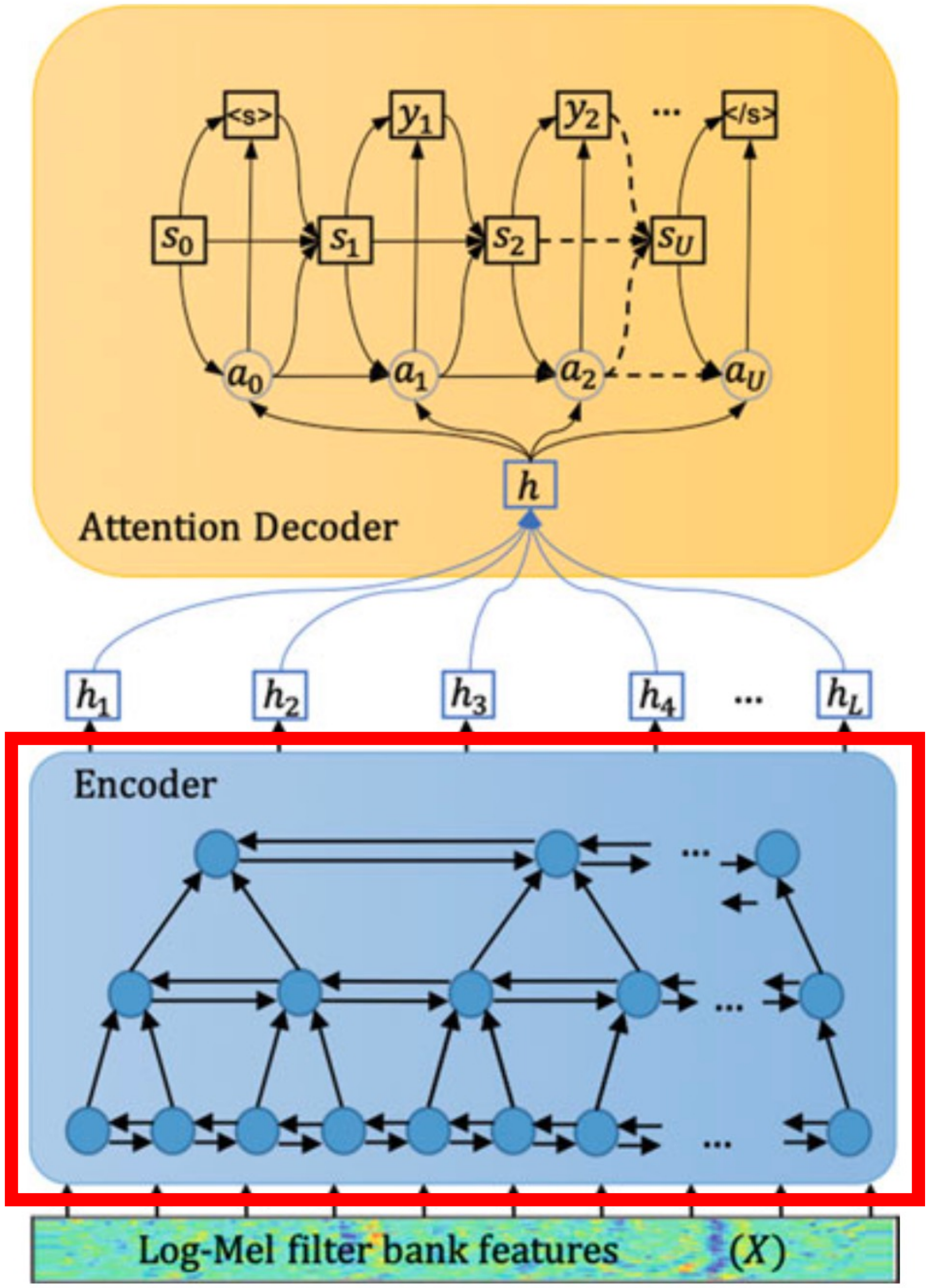


Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Listen, Attend, and Spell

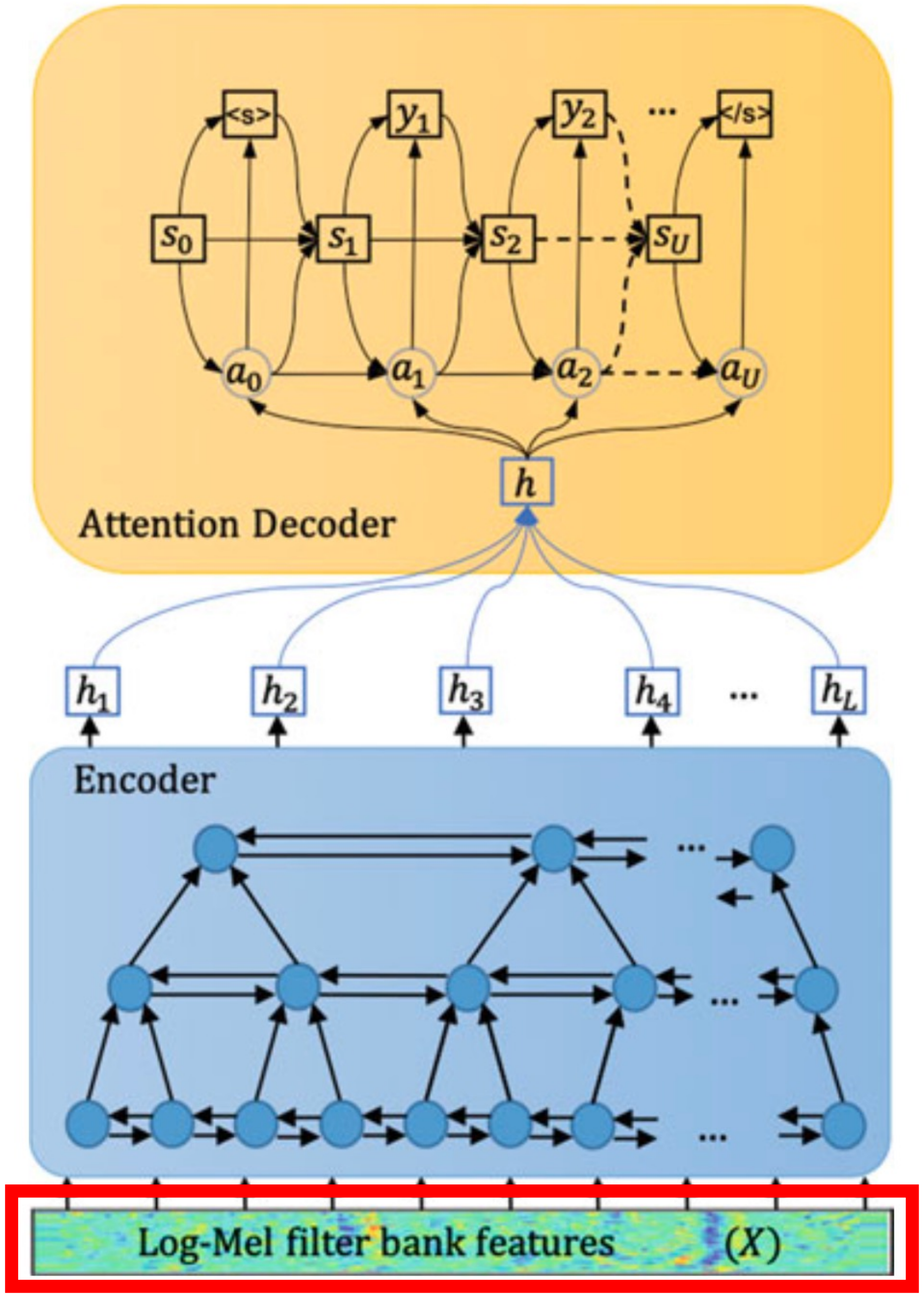
Mimics original paper on sequence to sequence learning with attention where the decoder learns what to attend to in the encoded representation

Pyramid structure reduces number of input time steps



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Listen, Attend, and Spell

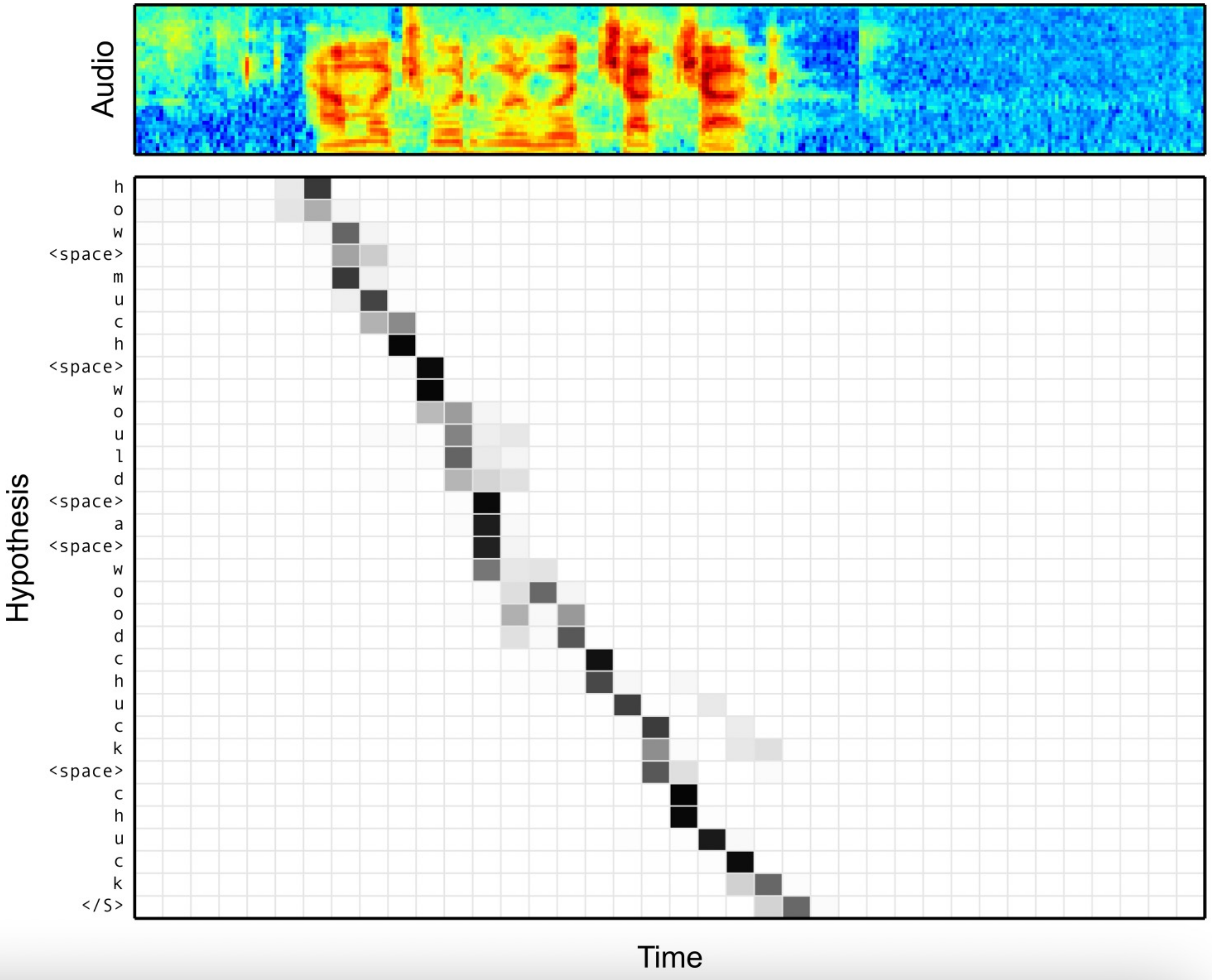


Input: more sophisticated hand-crafted audio representation then spectrogram

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Result

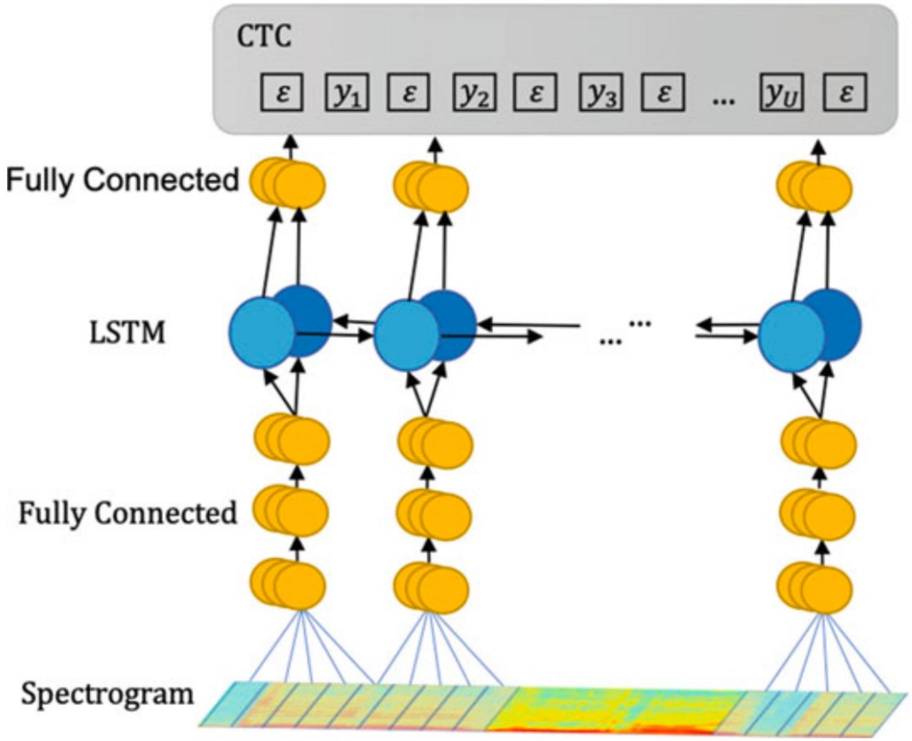
Attention enables visualizing alignment between audio signal and characters



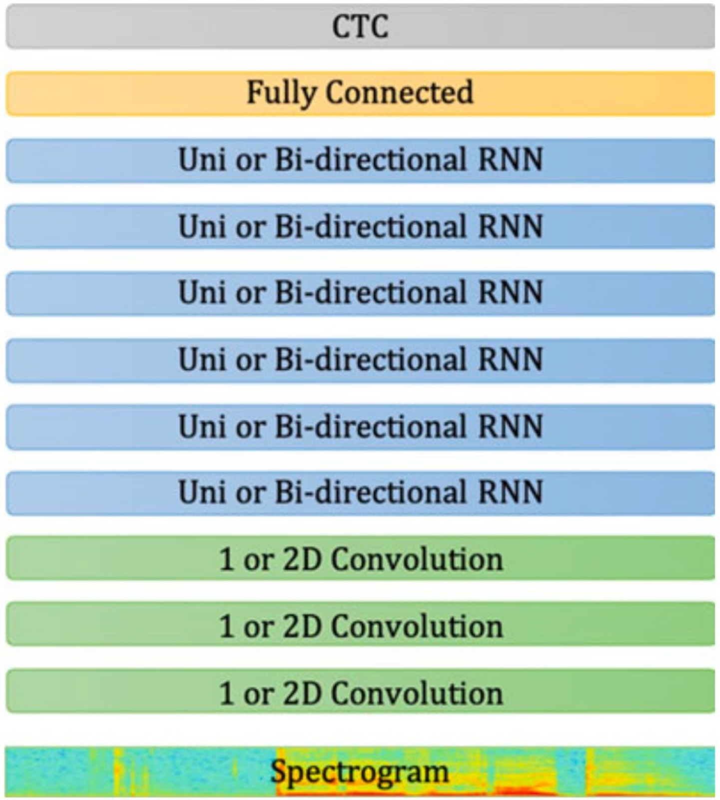
Chan et al. Listen, Attend and Spell. ICASSP 2016.

Popular Methods

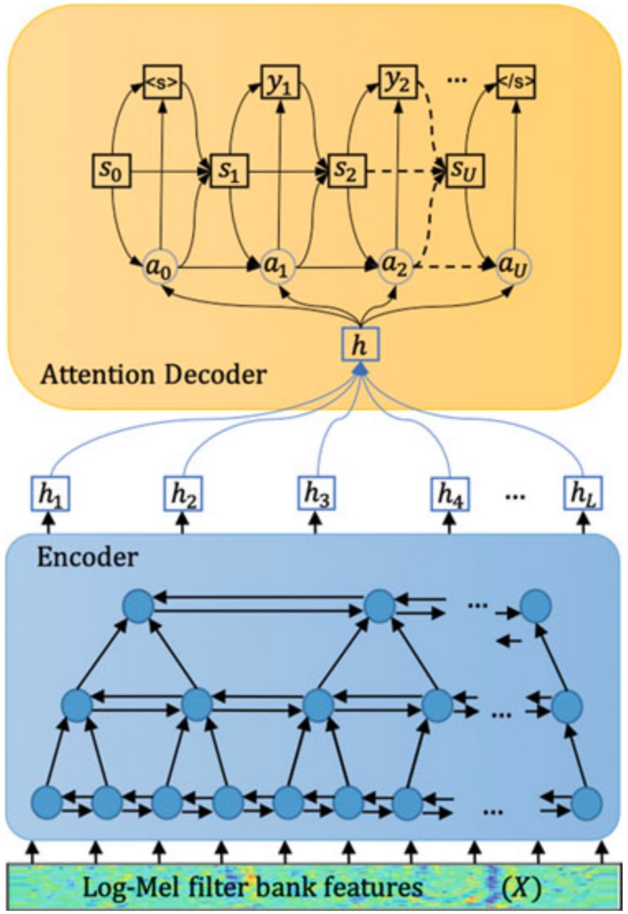
DeepSpeech



DeepSpeech2



Listen, Attend, and Spell



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Today's Topics

- Problem
- Applications
- Speech recognition evaluation
- Speech recognition models
- Video making tutorial

Today’s Topics

- Problem
- Applications
- Speech recognition evaluation
- Speech recognition models
- Video making tutorial

The End