Gridspace

IAP Program 2023
Lecture 7: Automatic Speech Recognition(ASR)

Jan 23, 2023

TODAY'S ROADMAP

- History of Automatic Speech Recognition(ASR)
- ASR Problem
- HMM based ASR
- E2E ASR
- Advanced E2E ASR

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

HISTORY OF ASR

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

HISTORY OF ASR

1952 Bell Lab's Audrey
(The first machine capable doing speech recognition)

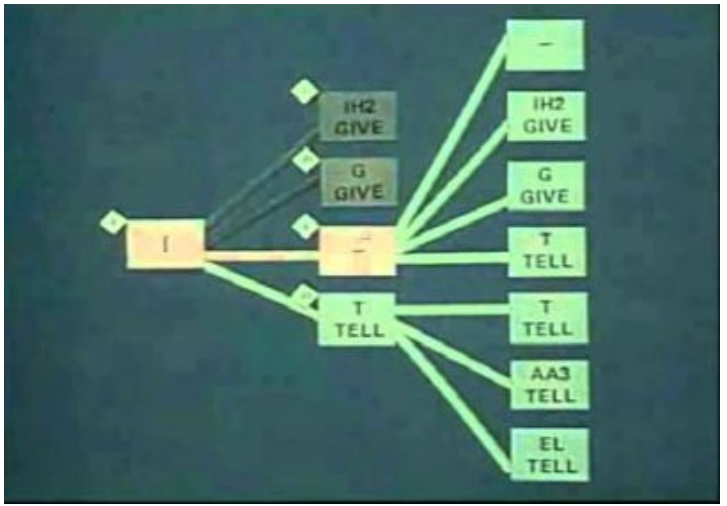
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

HISTORY OF ASR

1962 IBM’s Shoebox
(recognizing arithmetic words and digits)

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

HISTORY OF ASR



1970s CMU Harpy
(recognizing 1000+ words w/ beam search)

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

HISTORY OF ASR

A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition

LAWRENCE R. RABINER, FELLOW, IEEE

Although initially introduced and studied in the late 1960s and early 1970s, statistical methods of Markov source or hidden Markov modeling have become increasingly popular in the last several years. There are two strong reasons why this has occurred. First the models are very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of applications. Second the models, when applied properly, work very well in practice for several important applications. In this paper we attempt to carefully and methodically review the theoretical aspects of this type of statistical modeling and show how they have been applied to selected problems in machine recognition of speech.

In this case, with a good signal model, we can simulate the source and learn as much as possible via simulations. Finally, the most important reason why signal models are important is that they often work extremely well in practice, and enable us to realize important practical systems—e.g., prediction systems, recognition systems, identification systems, etc., in a very efficient manner.

These are several possible choices for what type of signal model is used for characterizing the properties of a given signal. Broadly one can dichotomize the types of signal models into the class of deterministic models, and the class

1980s HMM system dominates

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

HISTORY OF ASR

328 IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

1980s The first attempt use Neural Nets

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

HISTORY OF ASR

2000s

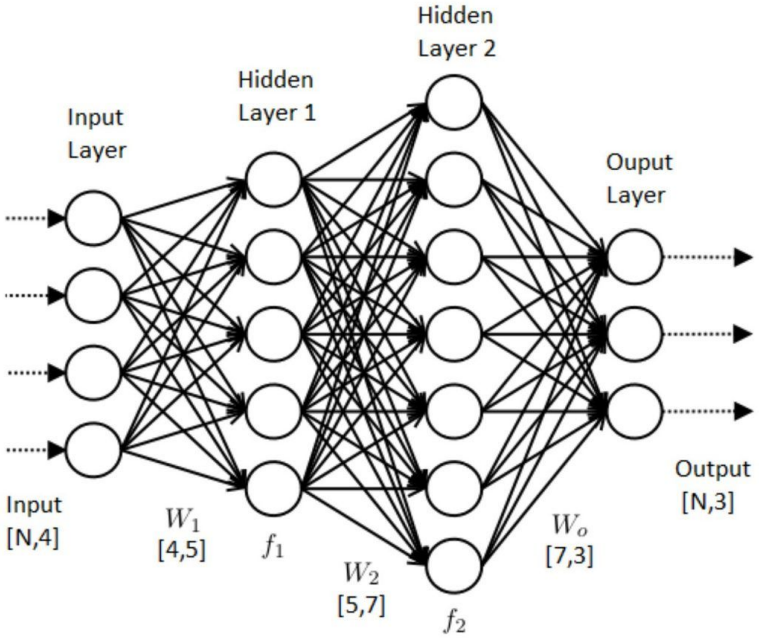
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

HISTORY OF ASR

2010s

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

HISTORY OF ASR



2010s

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

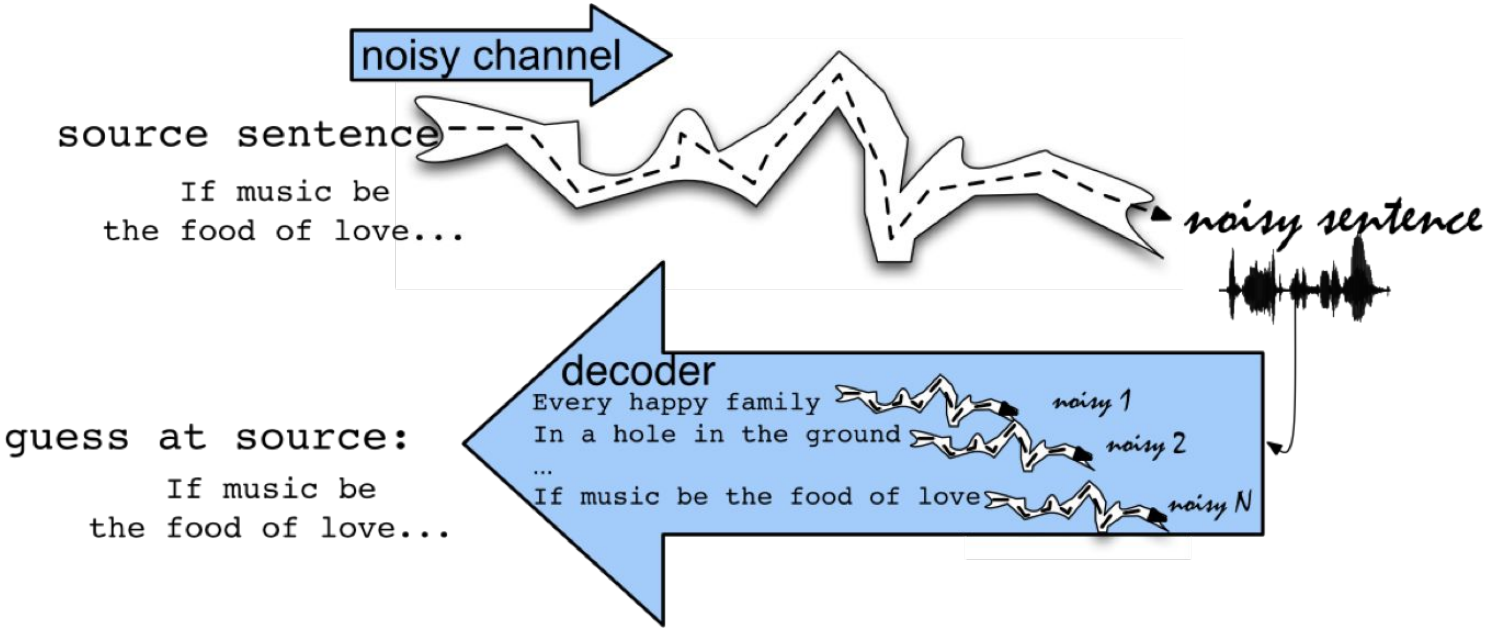
ASR PROBLEM

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Noisy Channel Model

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Noisy Channel Model



(Stanford CS224S)

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Statistical Model

$$W^* = \arg\max_W P(W|X)$$

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Statistical Model

$$W^* = \arg\max_W P(W|X)$$

$$W^* = \arg\max_W \frac{P(X|W)P(W)}{P(X)}$$

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Statistical Model

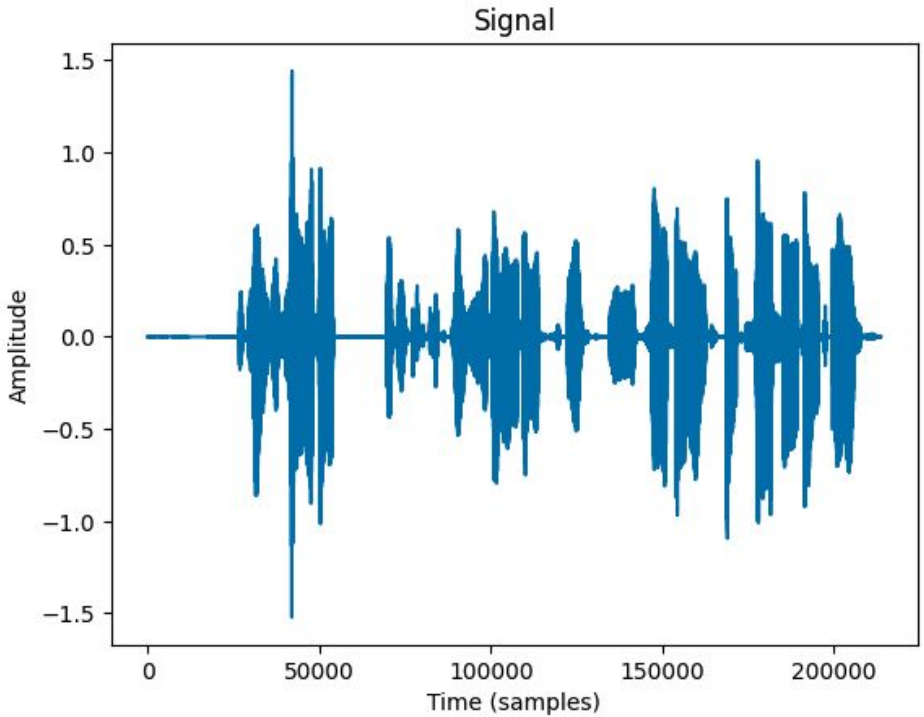
$$W^* = \arg\max_W P(W|X)$$

$$W^* = \arg\max_W \frac{P(X|W)P(W)}{P(X)}$$

$$W^* = \arg\max_W P(X|W)P(W)$$

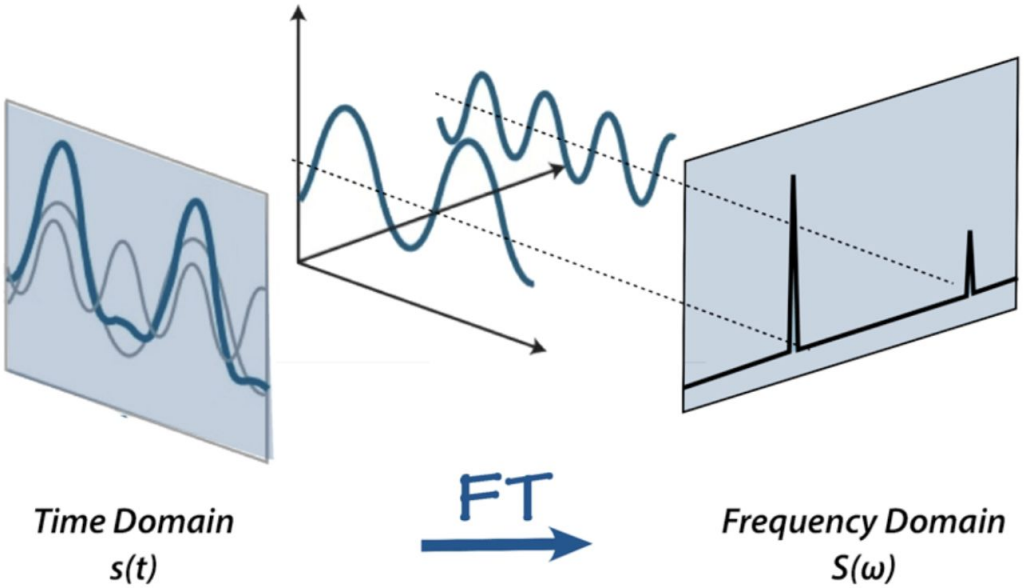
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

What is X?



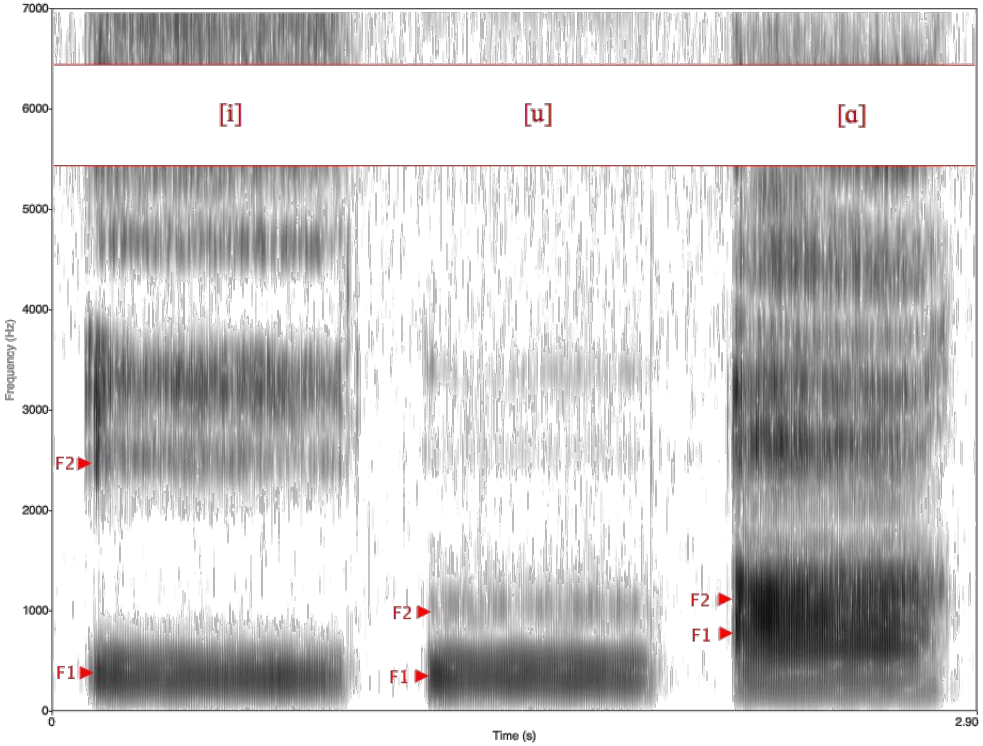
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

What is X?



GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

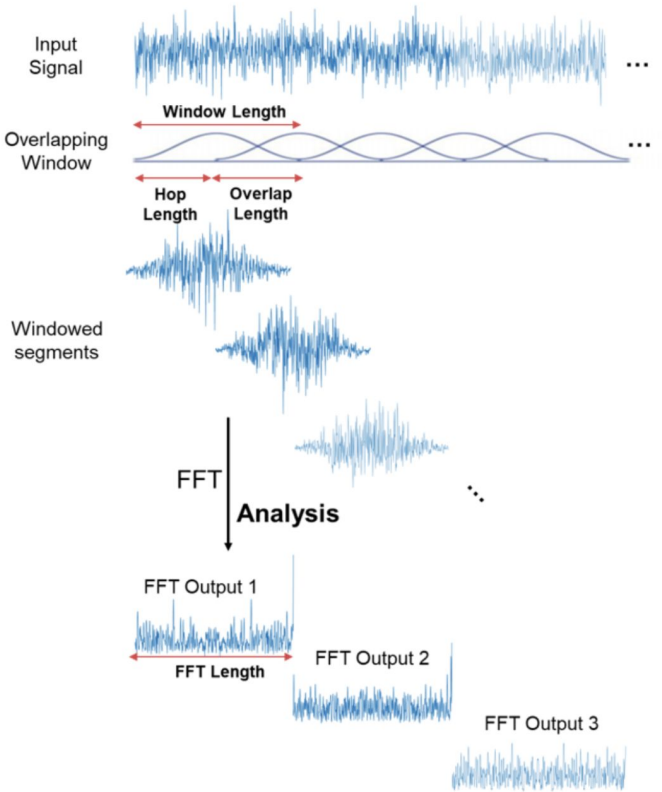
What is X?



Wikipedia: Formant

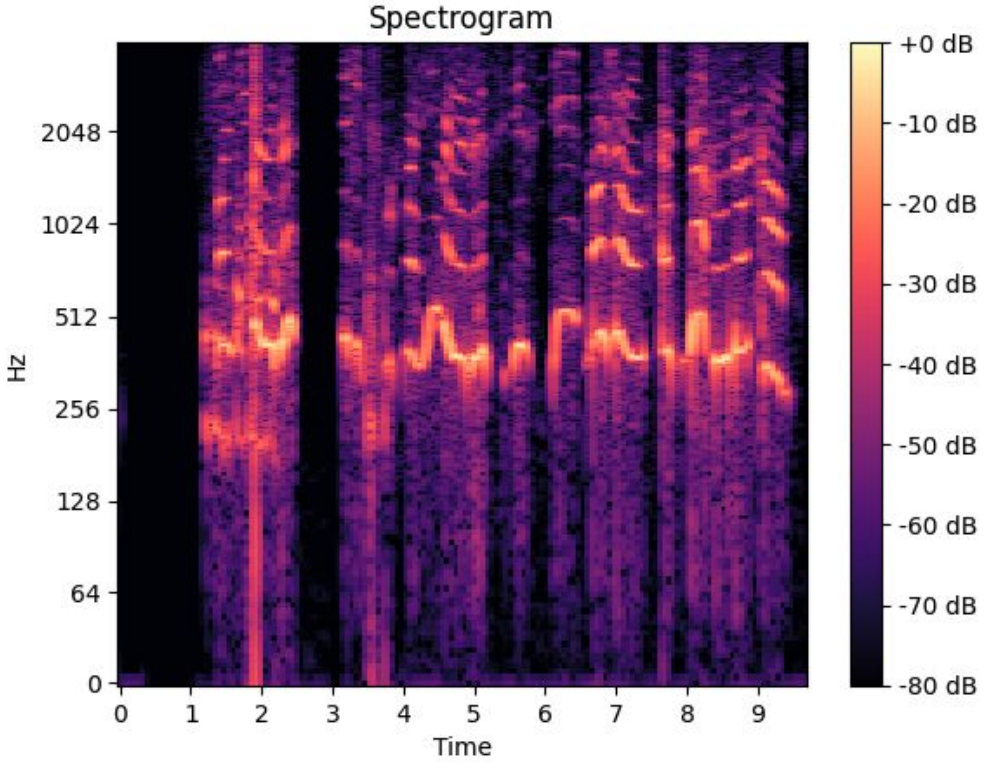
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

What is X?



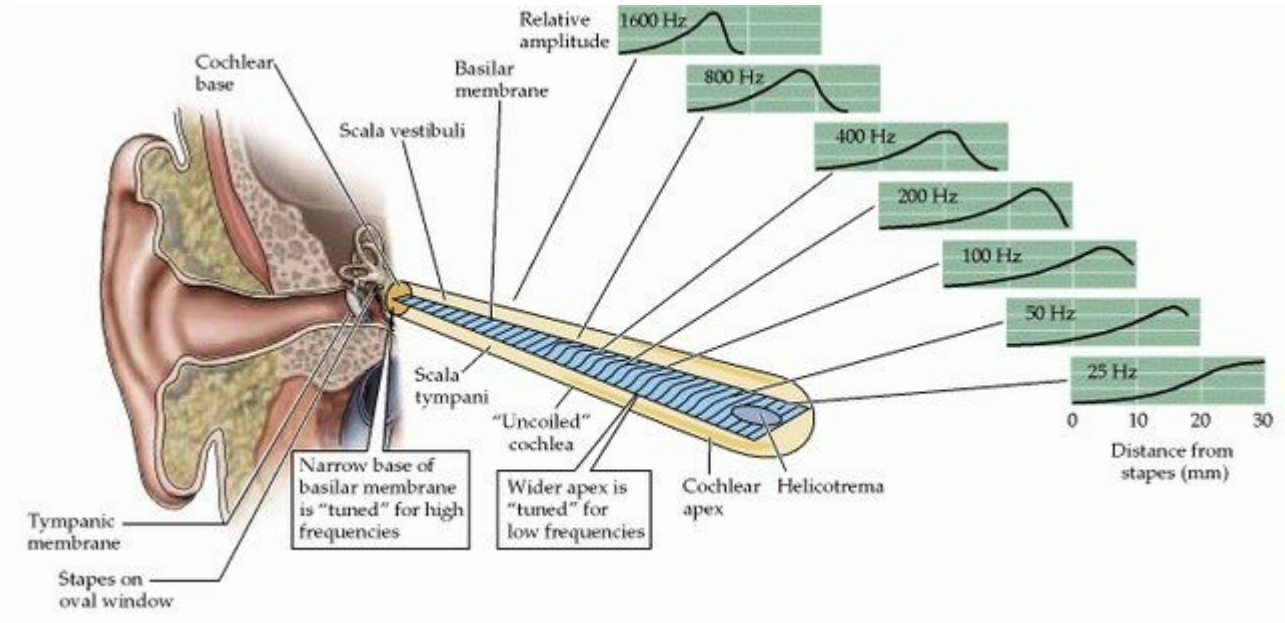
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

What is X?



GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

What is X?



GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

What is X?

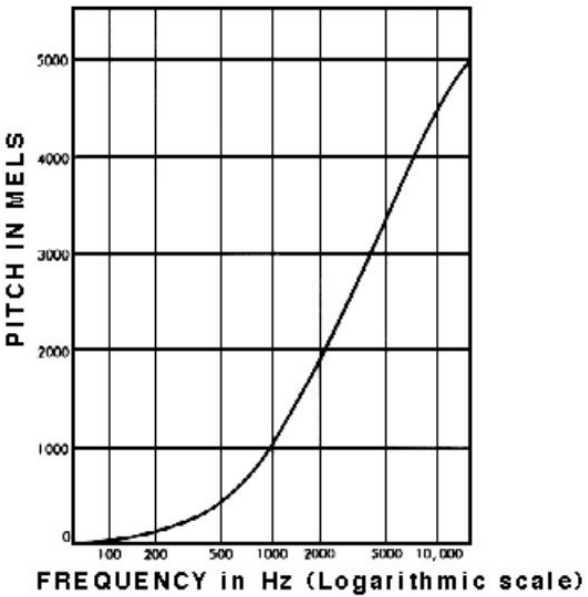
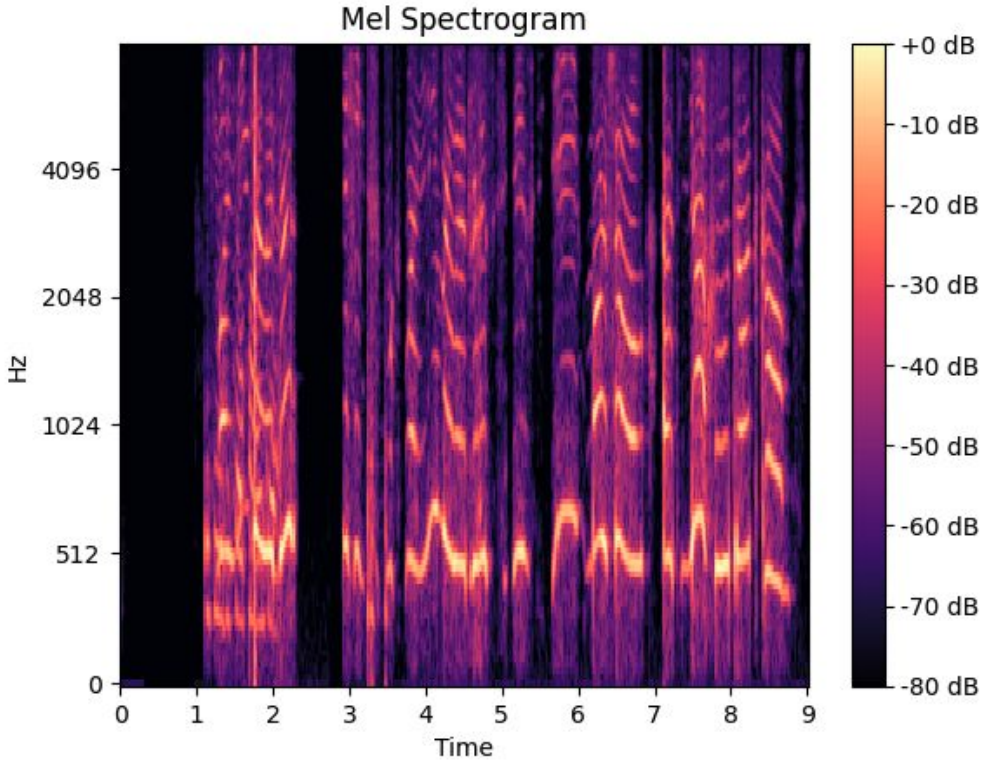


Image from Simon Fraser University

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

What is X?



GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

HMM based ASR

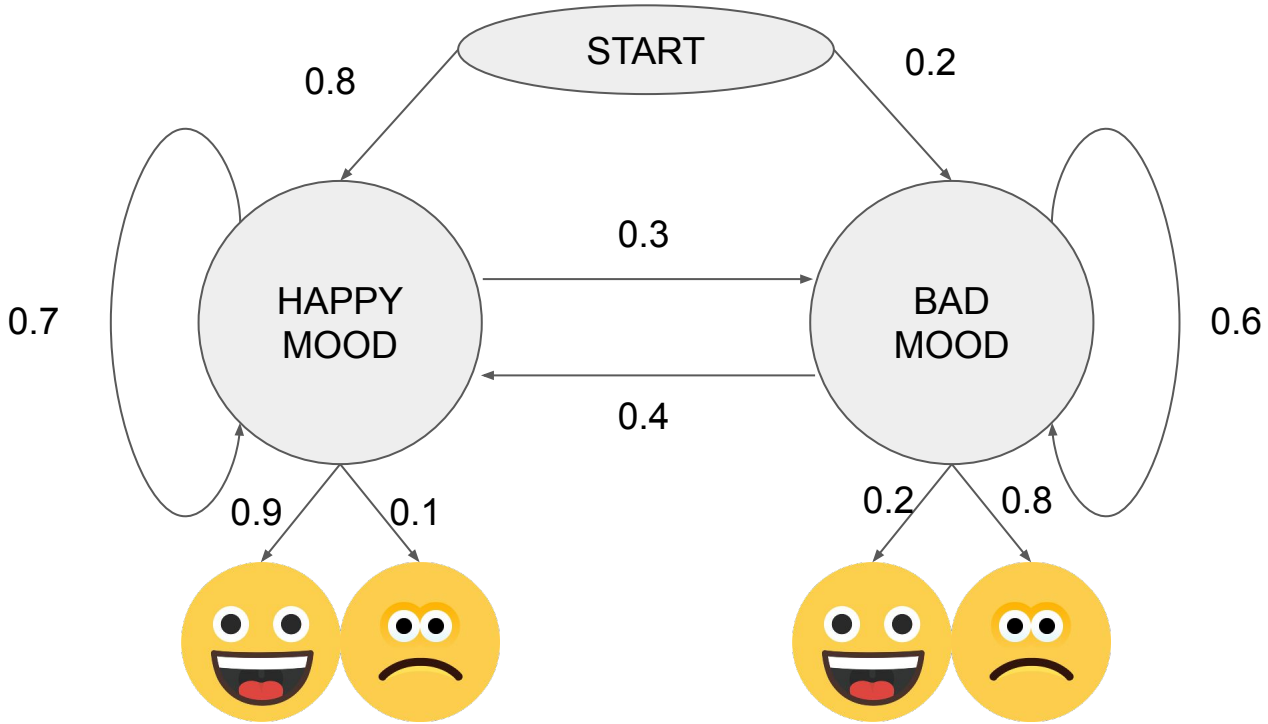
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

HMM(Hidden Markov Model)

1. Markov property
2. Stationary
3. Output Independence

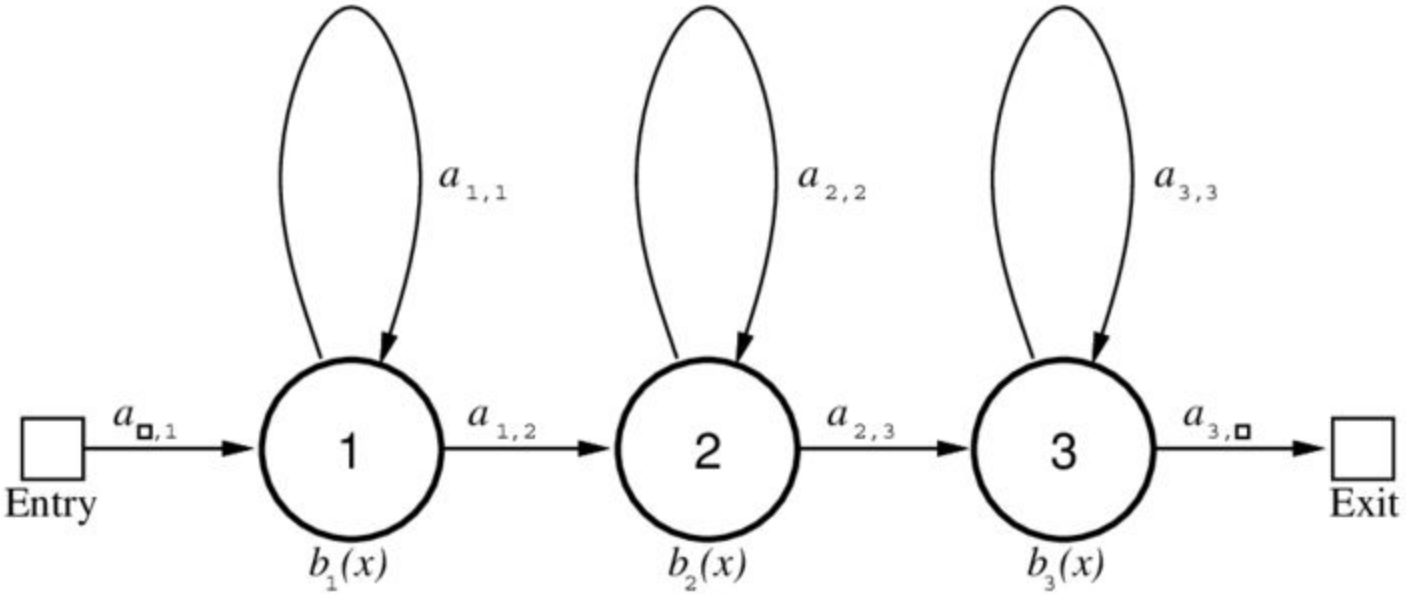
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

HMM(Hidden Markov Model)



GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

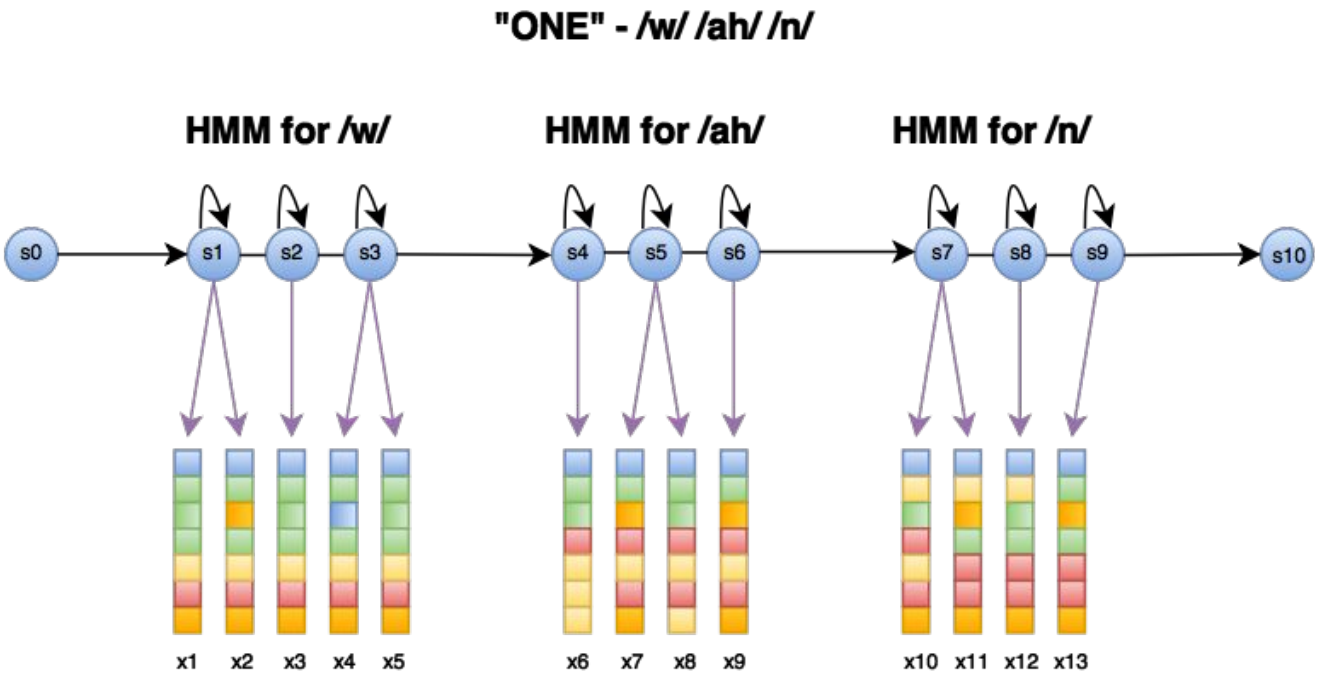
HMM in Speech



3-state left-to-right HMM

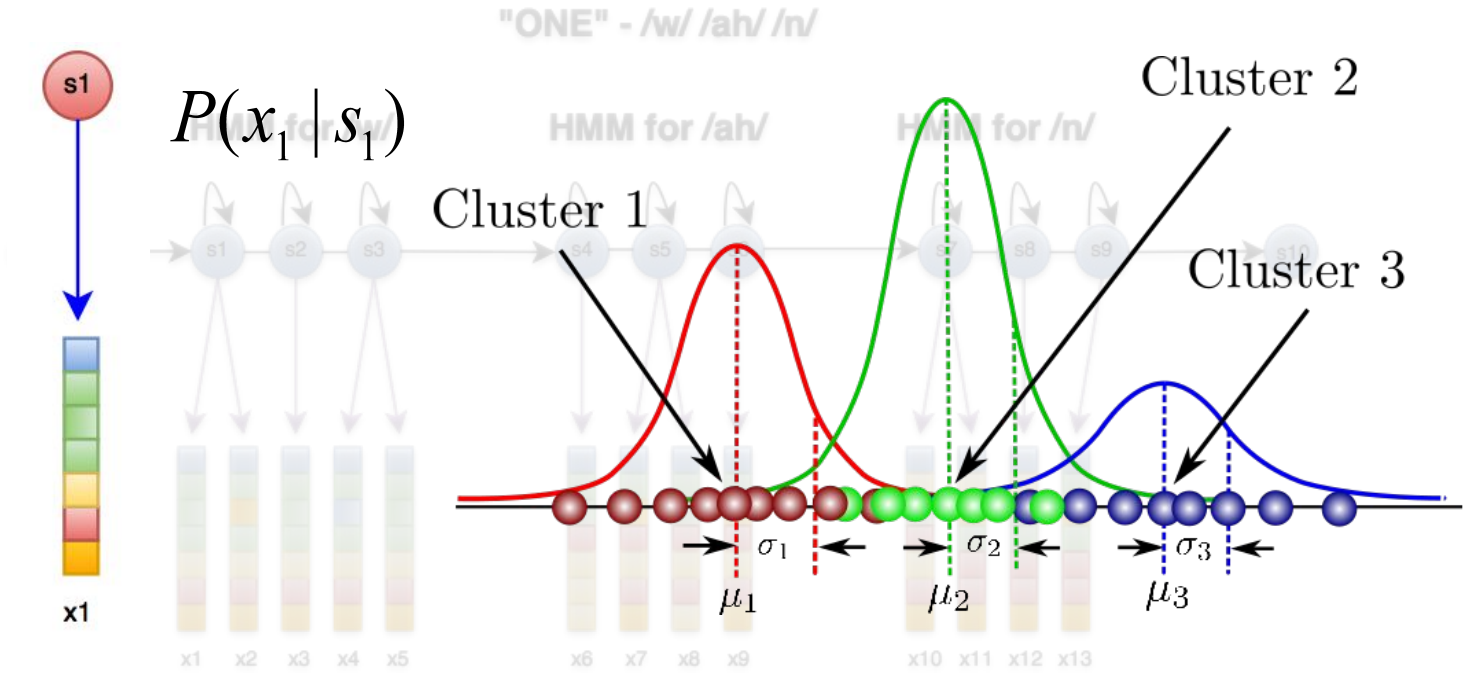
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Acoustic Model - HMM



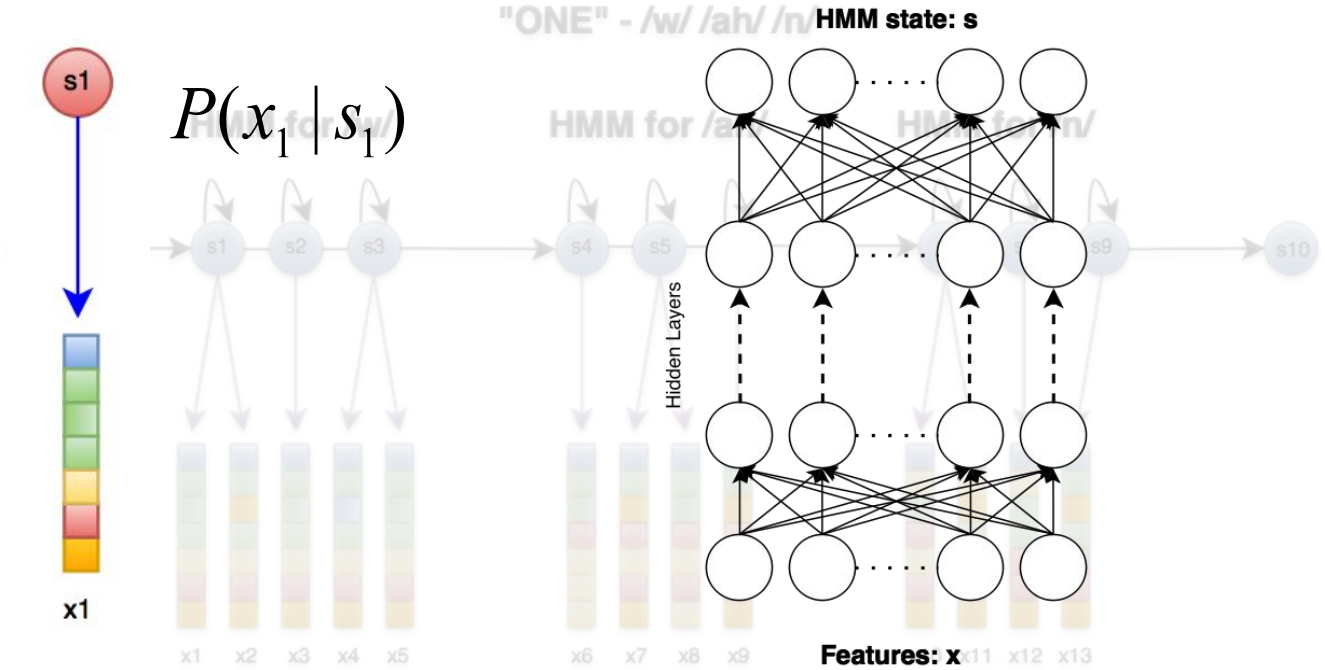
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Acoustic Model - HMM



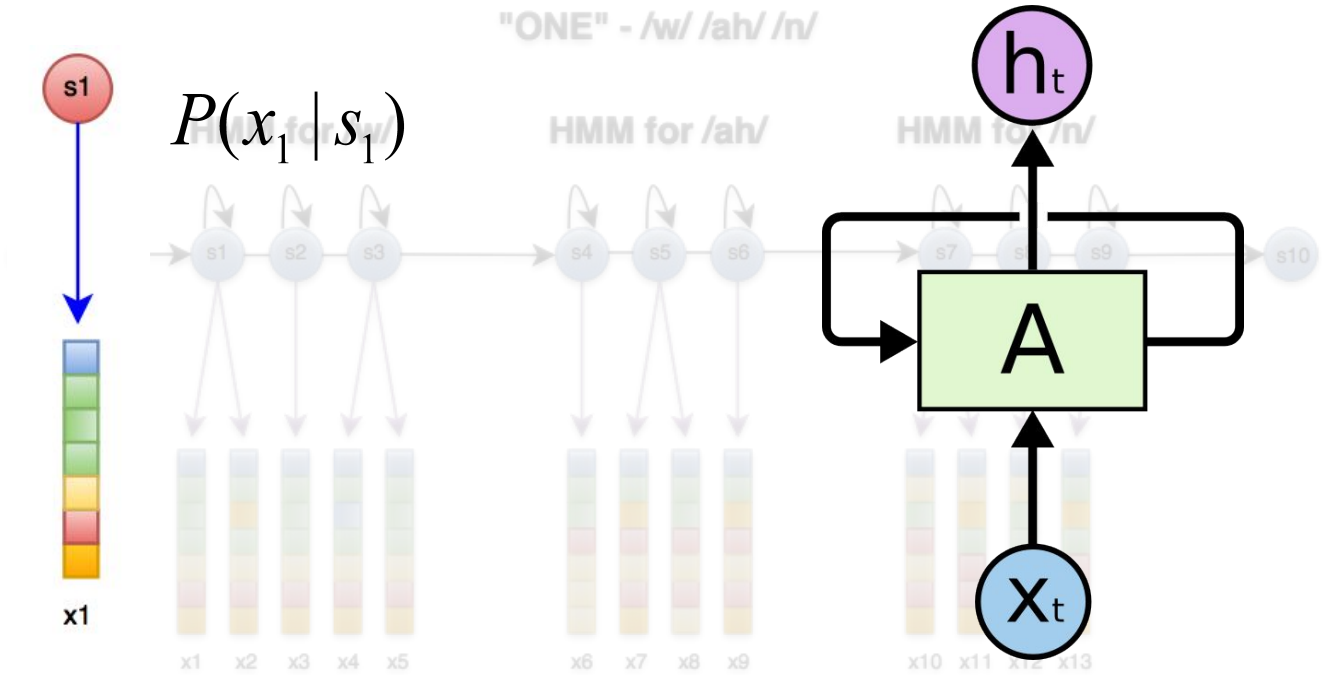
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Acoustic Model - HMM



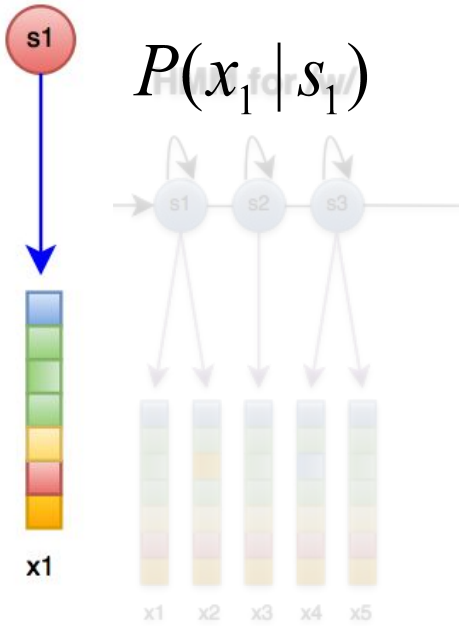
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

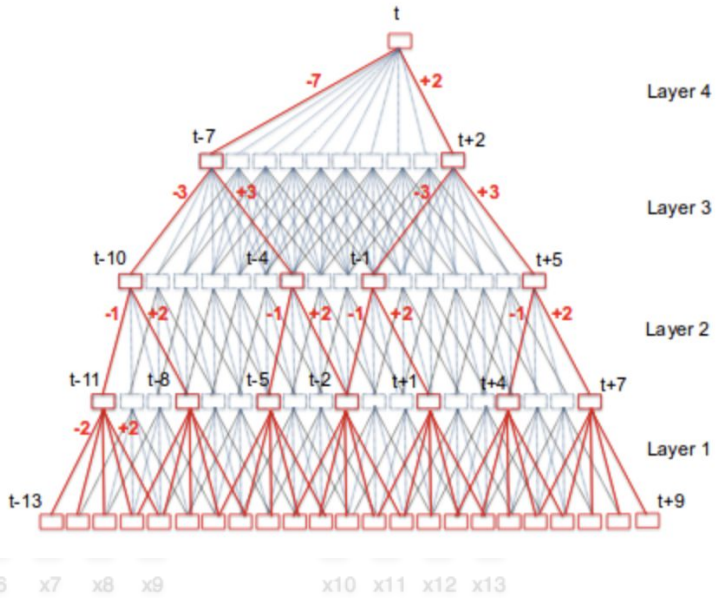
Acoustic Model - HMM



GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Acoustic Model - HMM

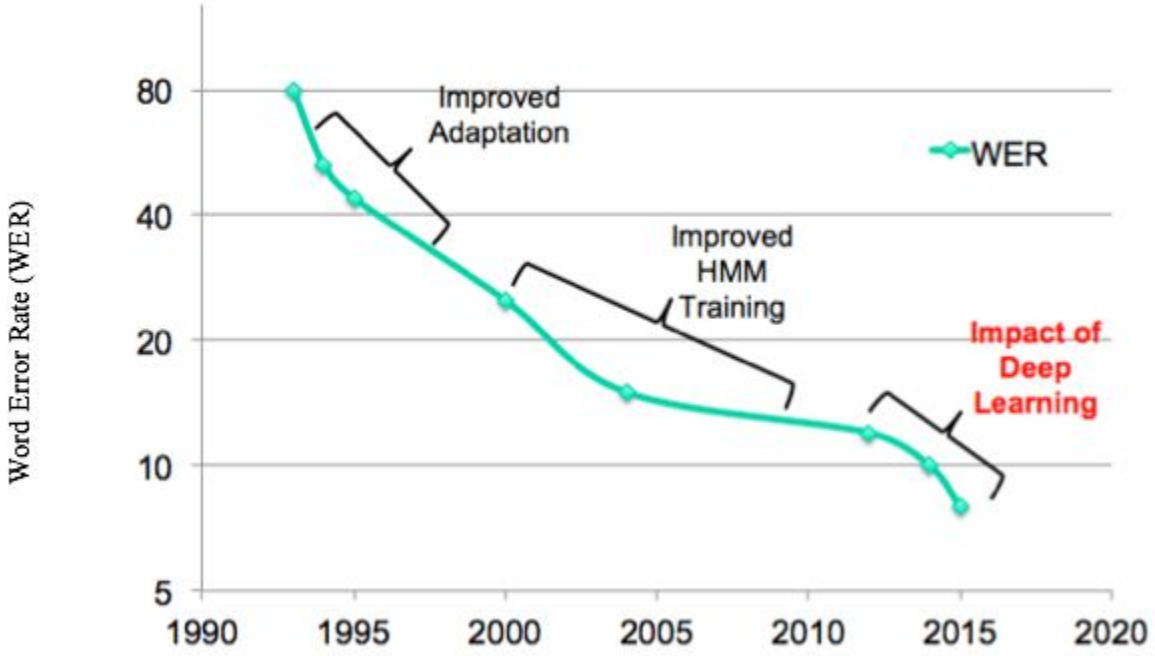




GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Acoustic Model - HMM

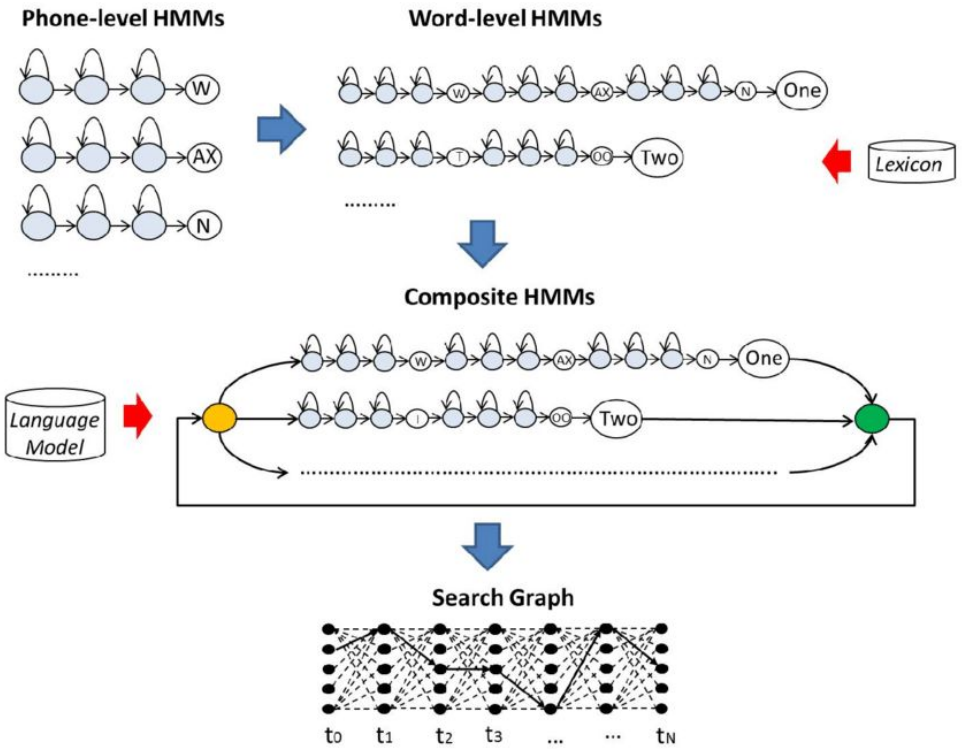
DNN-HMM system(Hybrid) brought huge impact until recently!



(from IBM)

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

HMM

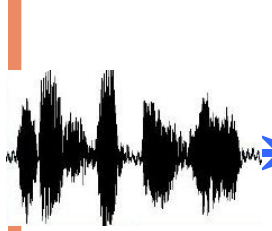


(image from: Mirco Ravanelli)

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Limitation of HMM

$$X = (x_1, x_2, \ldots, x_T)$$



Feature Extraction

Decoder

$$W^* = (w_1, w_2, \ldots, w_K)$$

Word Sequences

Acoustic Model (AM)

$P(X|Q)$

Lexicon

$P(Q|W)$

Language Model (LM)

$P(W)$

$$W^* = \arg\max_W P(X|W)P(W)$$

GRIDSPACE CONFIDENTIAL @2021 ALL RIGHTS RESERVED

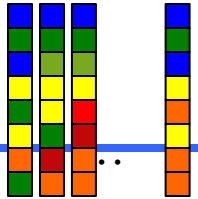
E2E ASR

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

End-to-End ASR

$$X = (x_1, x_2, \ldots, x_T)$$

Feature Extraction



Neural Net

$$W^* = (w_1, w_2, \ldots, w_K)$$

Word Sequences

- Direct Modeling of $P(W|X)$
- Sequence to Sequence Problem
  - Great Success in Machine Translation
  - Summarization
  - …

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

End-to-End ASR

1. Connectionist Temporal Classification (CTC)
2. Attention-Based Enc-Dec (AED)
3. RNN-Transducer (RNN-T)

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

CTC

- Alex Graves 2006
  - Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networ
- HMM-Free Speech Recognition

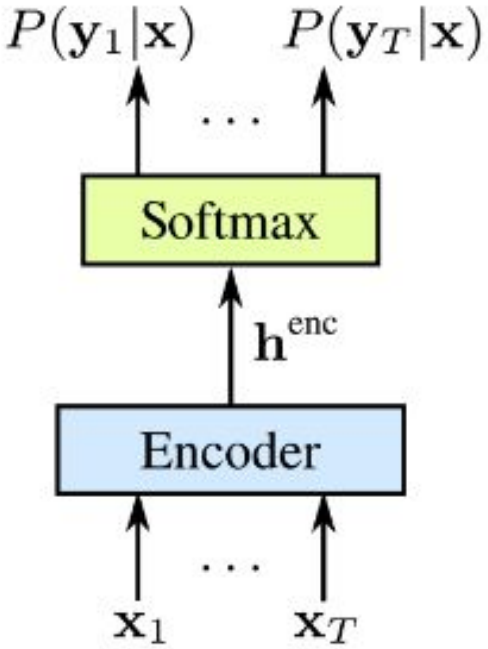


image from: Tara Sainath

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

CTC

h _ e _ _ l l o
h e _ l l o _ _
h e _ _ l l o _
...

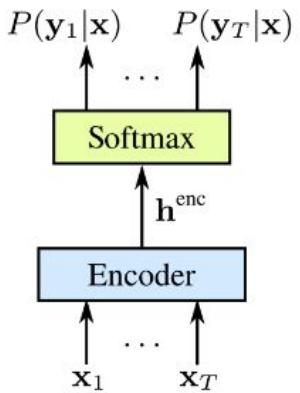


image from: Tara Sainath

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Limitation of CTC

1. Conditional Independence still exists
2. Strong LM is required to decode

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

AED

- Will Chan 2015
  - Listen Attend Spell
- No conditional independence assumed

- Listener: DNN part of Acoustic Model
- Attention: HMM part of Acoustic Model
- Spell: Language Model

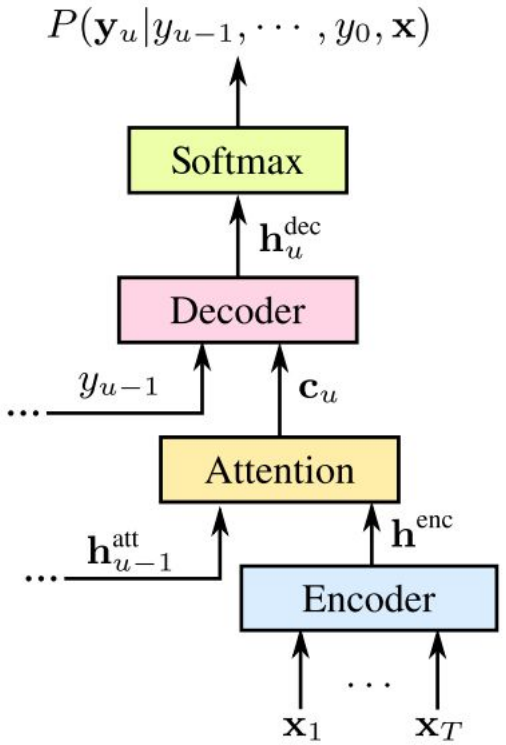
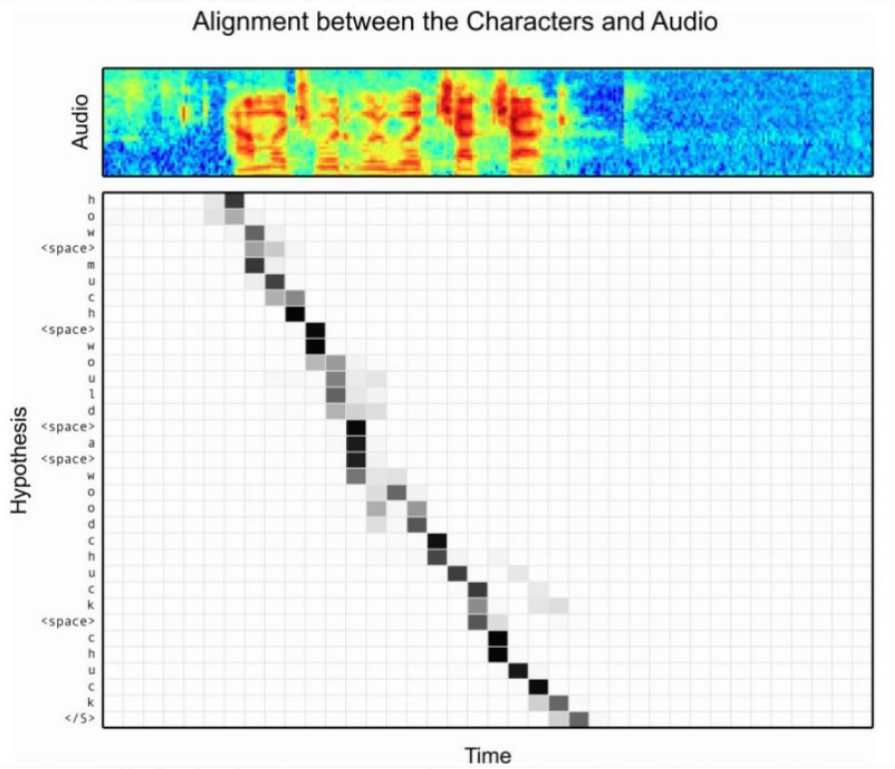


image from: Tara Sainath

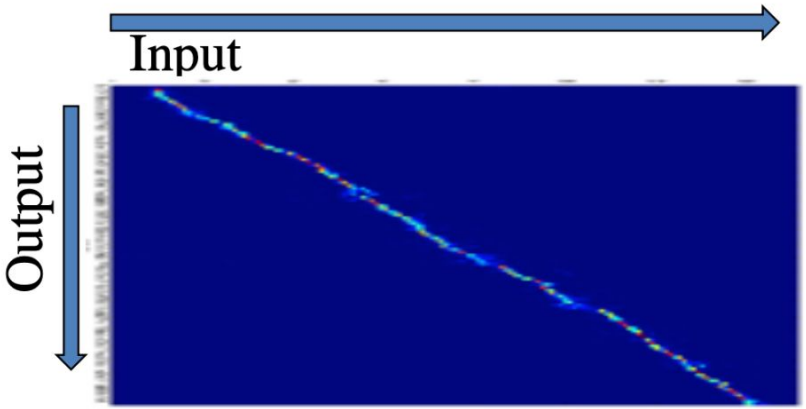
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

AED

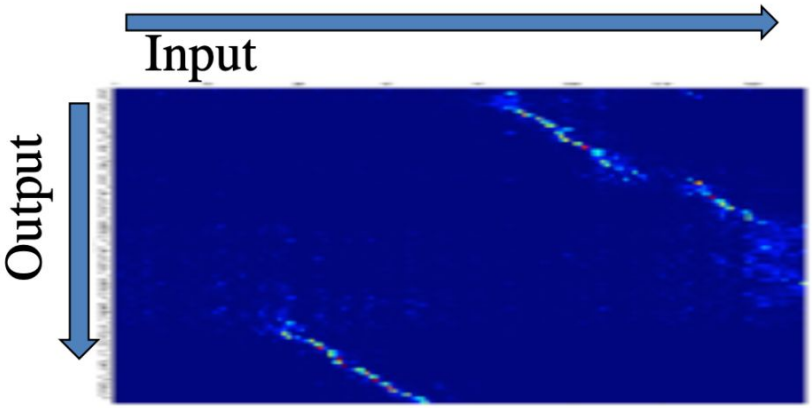


GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Joint Training for AED



Example of monotonic alignment



Example of distorted alignment

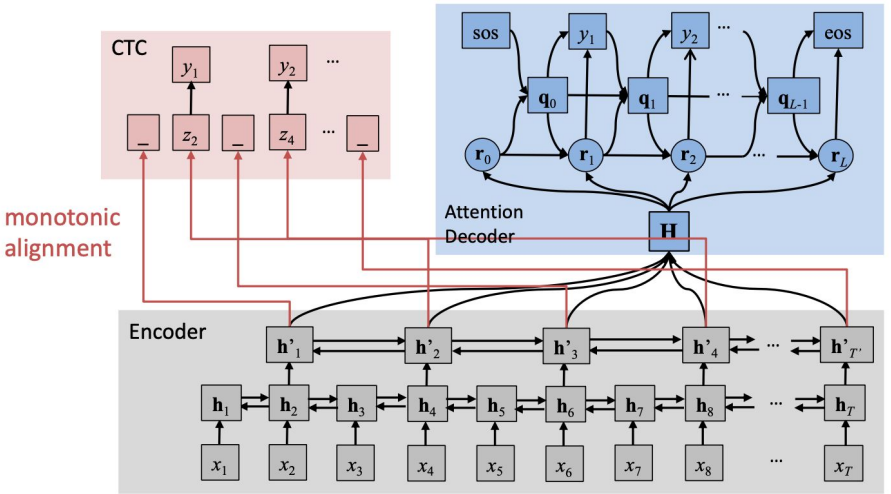
(Shinji Watanabe)

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Joint Training for AED

- Suyoun Kim 2017
- Joint training of CTC and LAS

Multitask learning: $\mathcal{L}_{\mathrm{MTL}} = \lambda \mathcal{L}_{\mathrm{CTC}} + (1-\lambda)\mathcal{L}_{\mathrm{Attention}}$ λ: CTC weight



CTC guides attention alignment to be monotonic

(Shinji Watanabe)

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Transducer(RNN-T)

- Alex Graves 2012, Kanishka Rao 2017
  - Sequence Transduction with Recurrent Neural Networks

- Encoder: Acoustic Model
- Predictor: Language Model
- Joiner: FC Net (Combining AM + LM)

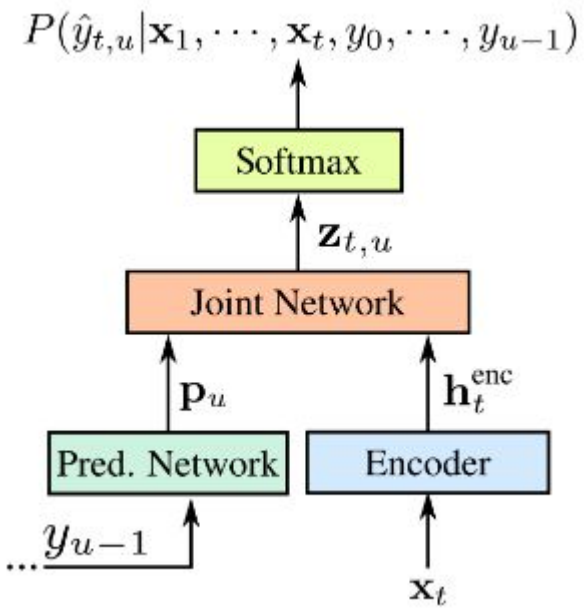
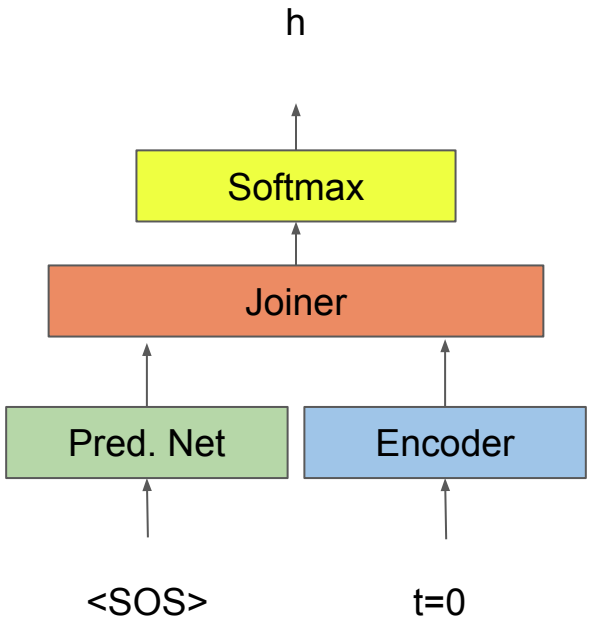


image from: Tara Sainath

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Transducer(RNN-T)

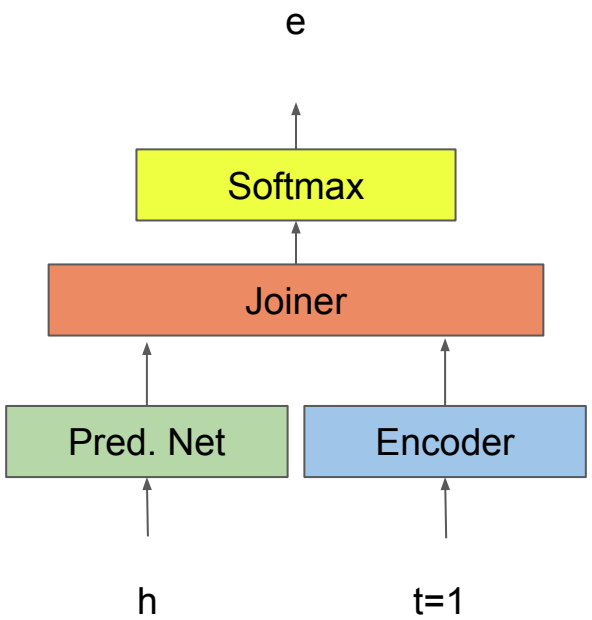


Result: h

image from: Tara Sainath

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Transducer(RNN-T)

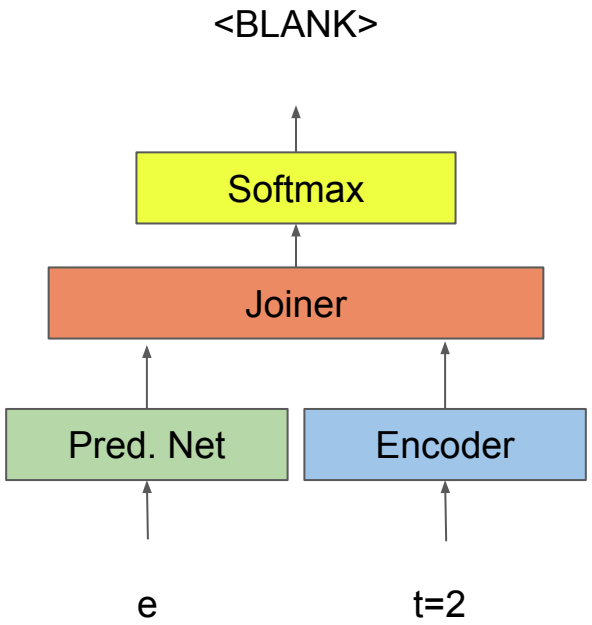


Result: h e

image from: Tara Sainath

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Transducer(RNN-T)

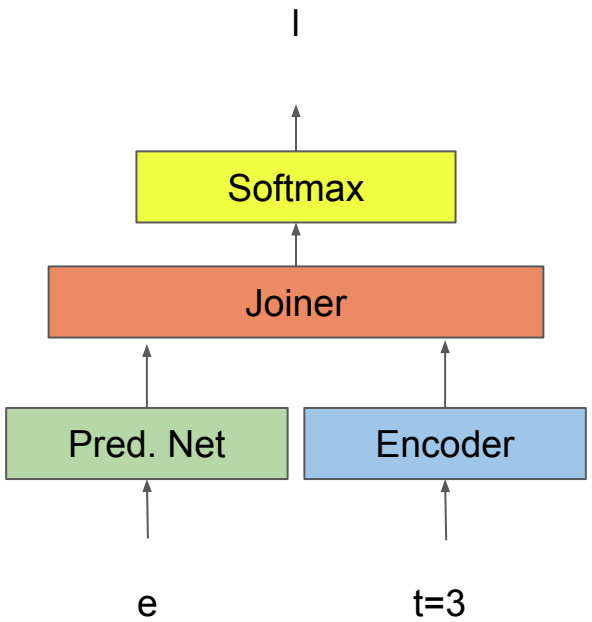


Result: h e <B>

image from: Tara Sainath

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Transducer(RNN-T)

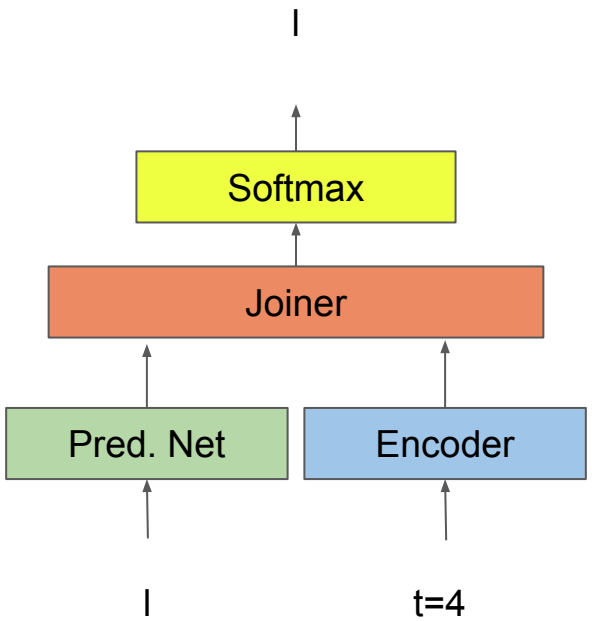


Result: h e <B> l

image from: Tara Sainath

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Transducer(RNN-T)

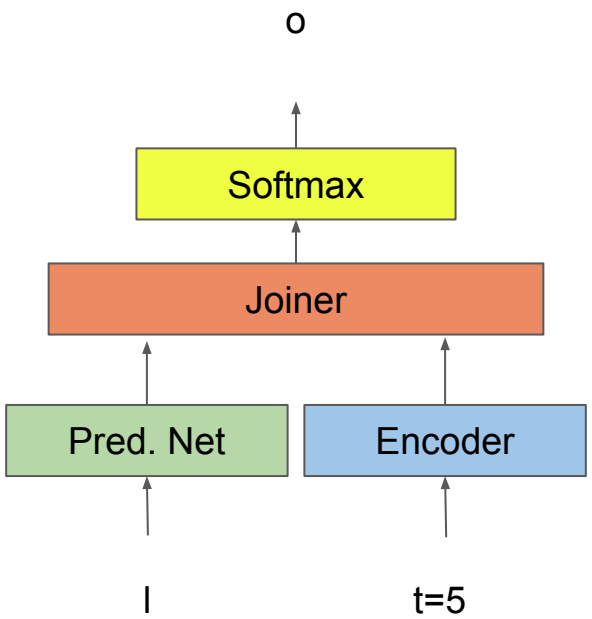


Result: h e <B> l l

image from: Tara Sainath

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Transducer(RNN-T)

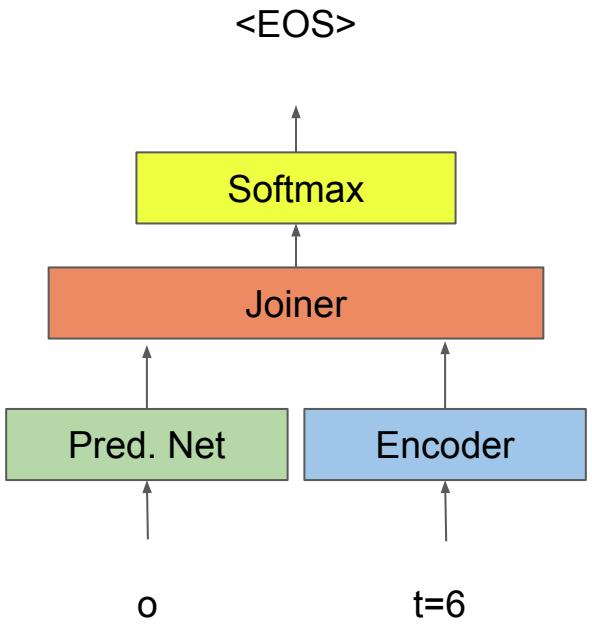


Result: h e <B> l l o

image from: Tara Sainath

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Transducer(RNN-T)



Result: h e <B> l l o

image from: Tara Sainath

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Advanced E2E ASR

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Tokenization

Subword unit
- subword unit by data-driven way
- ex: SentencePiece

VS Compared char, phoneme token
- lower perplexity
- less often decoding

VS Word token
- rare words are not enough for training
- much less than output dimension

GRIDSPACE CONFIDENTIAL @2021 ALL RIGHTS RESERVED

Augmentation

Daniel Park(2020)

- Spec Augment

- 1. Time Warping
- 2. Freq Masking
- 3. Time Masking

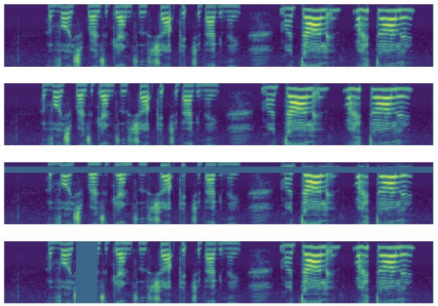


Figure 1: *Augmentations applied to the base input, given at the top. From top to bottom, the figures depict the log mel spectrogram of the base input with no augmentation, time warp, frequency masking and time masking applied.*

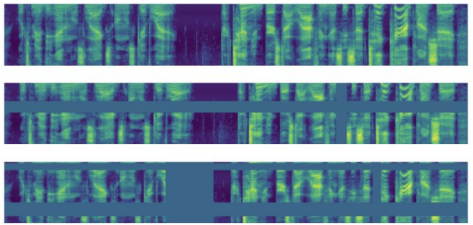


Figure 2: *Augmentation policies applied to the base input. From top to bottom, the figures depict the log mel spectrogram of the base input with policies None, LB and LD applied.*

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Transformer ASR

- Since “Attention is all you need”
- There are lots of variants

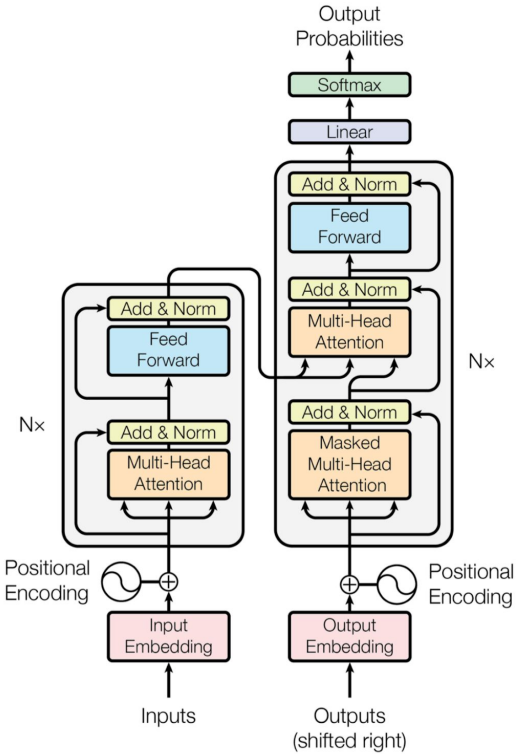


Figure 1: The Transformer - model architecture.

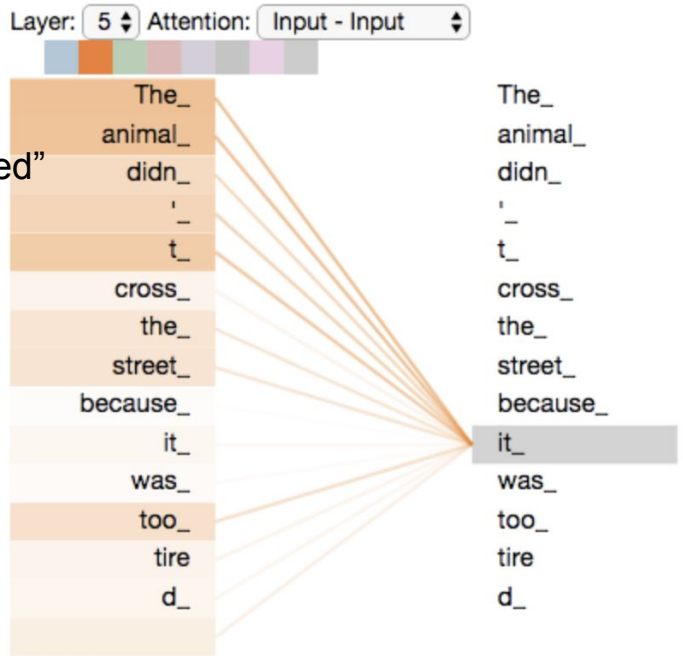
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Transformer ASR

- Self-Attention

Let's say we are trying to model the sentence below:

"The animal didn't cross the street because it was too tired"



(https://jalammar.github.io/illustrated-transformer)

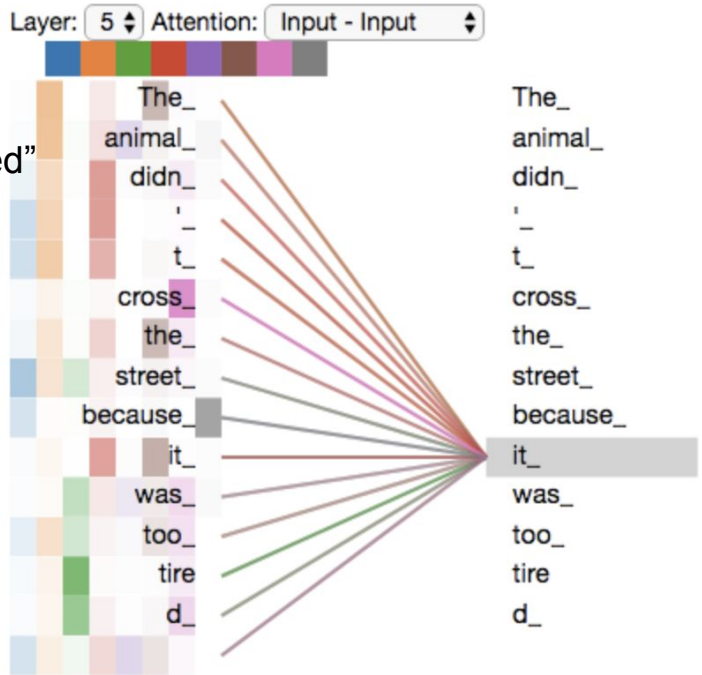
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Transformer ASR

- Multi-Heads Attention

Let's say we are trying to model the sentence below:

"The animal didn't cross the street because it was too tired"



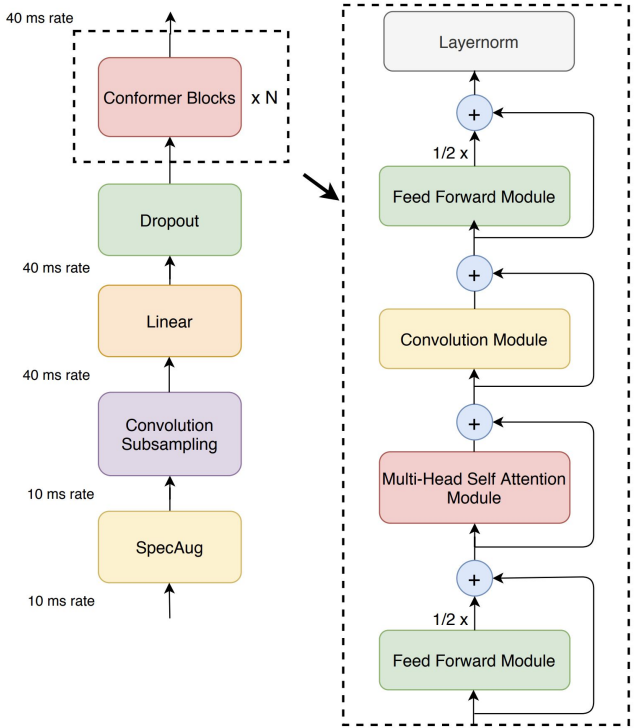
(https://jalammar.github.io/illustrated-transformer)

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Conformer ASR

Conformer

- Sequence-to-sequence transformer with multi-headed self attention. Directly optimizes target word sequence
- Combines attention (global context) with convolution (local invariance)



GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Attention for Online ASR

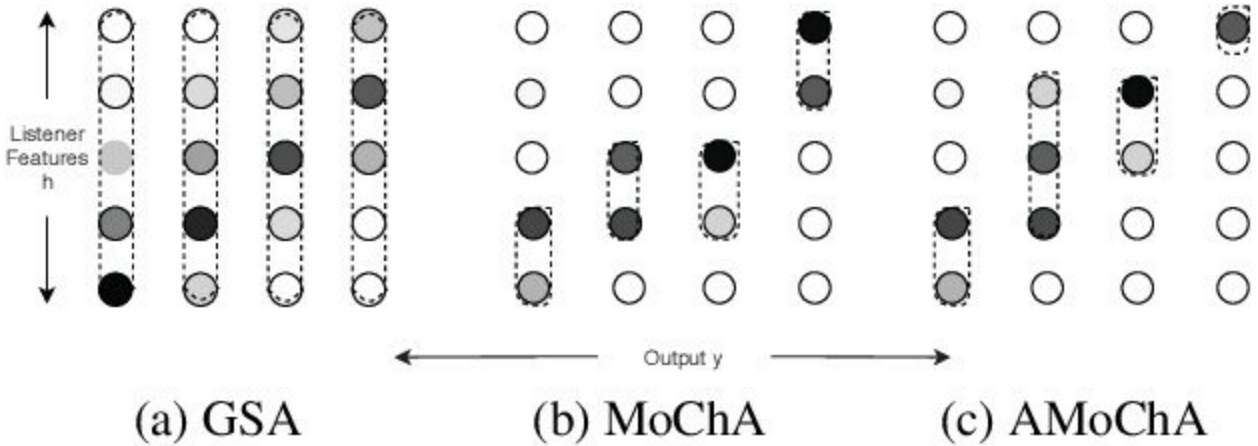
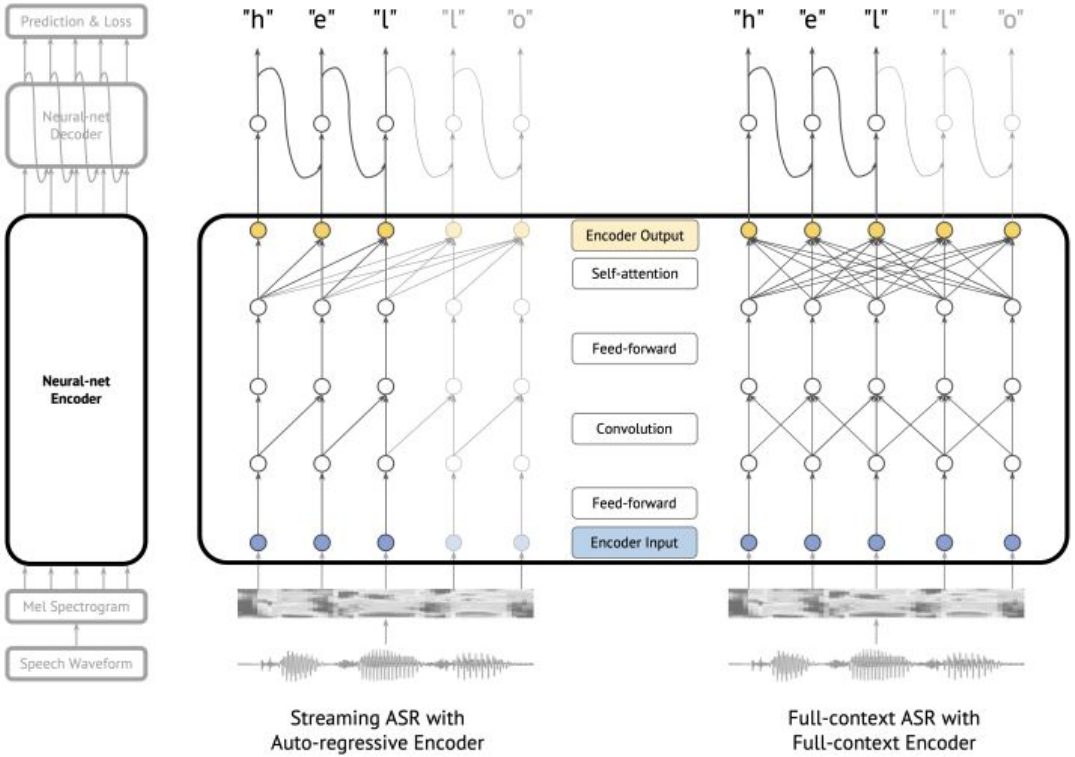


image from: “An Online Attention-based Model for Speech Recognition”

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Dual Mode ASR



GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Joint E2E ASR

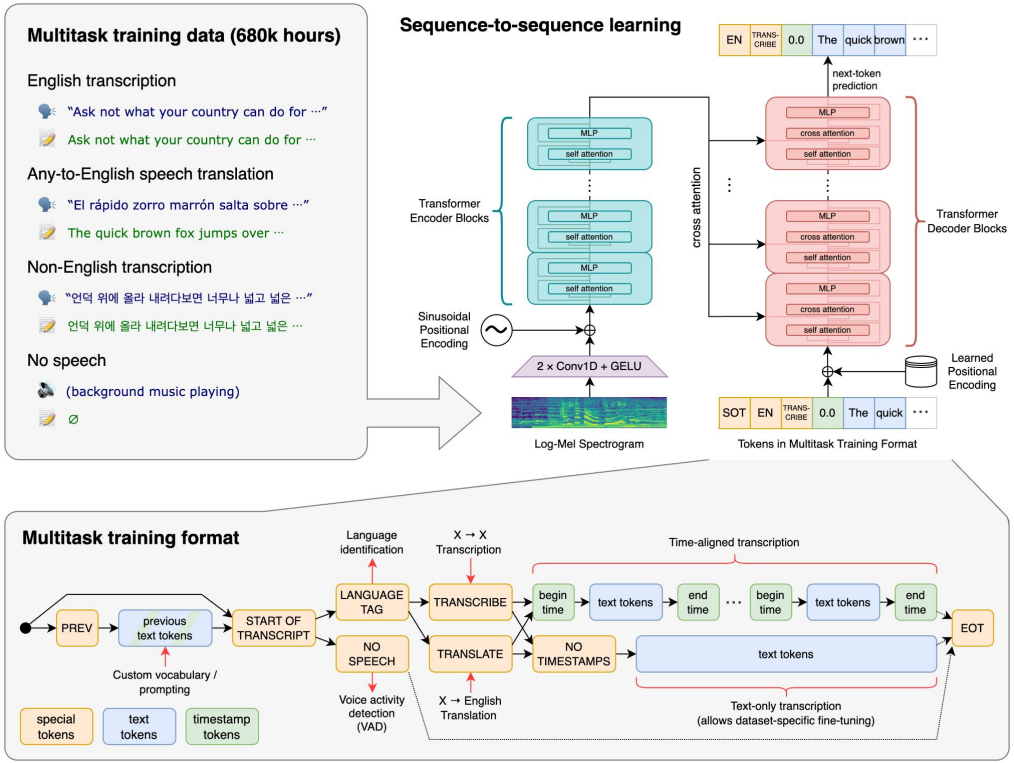


Image from: OpenAI Whisper

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Go Bigger

Year	Model Name	Type	SIZE	DATA
2015	DeepSpeech2	CTC	24M	1k+
2016	LAS	AED	25M	2k+
2019	RNN-T(Google)	Transducer	90M	18K+
2020	Transformer-Based(Meta)	AED	150M	13.7K
2022	Whisper	AED	1.5B	680K

GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Exercises for Next Time

- With HMM model in 27p, what would be the most probable state when your spouse sent 😀☹️😀 ?
- When you have 1 hours of labeled speech data, how do you want to train speech recognition?
- To make noise robust speech recognition, what can you do with modern e2e speech recognition?

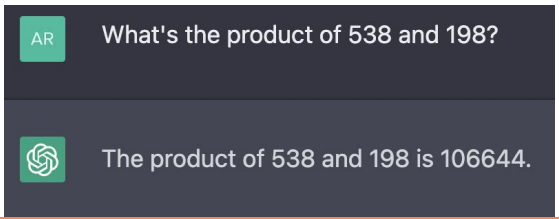
GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED

Exercises from Last Time

- Recurrent networks exploit translation symmetry in text (you can shift text and it's still text). What symmetry do convolutional neural networks exploit?
  Translation symmetry in space rather than time.

- GPT-2 had a context window of 1024 tokens. ChatGPT / GPT-3.5 has a context window of 8,192 tokens. How might this change the number of parameters in the model, keeping embeddings and number of attention heads / layers the same?
  $O(n^2)$ where n is context size, so by a factor of ~64X.

- Why does ChatGPT struggle to perform arithmetic when it can write functional code?
  While LLM's understand numbers and their properties in a sentence, they are not easily-equipped to perform algorithms like arithmetic on large numbers which don't resemble linear text generation.
  See this hallucination here:



GRIDSPACE CONFIDENTIAL @2023 ALL RIGHTS RESERVED