ASR State-of-the-Art

Vincenzo Norman Vitale - UrbanECO Research Center at UniNa

Urban/Eco

AISV
Associazione Italiana Scienze della Voce

Outline

- Intro to ASR
- Basic Concepts
- Most Advanced ASR Models
  - Encoder:
    - Conformer family
    - Self-Supervised
    - Semi-Supervised: Whisper
    - Mamba-models
  - Decoder:
    - CTC
    - Transducer
    - Transformer
- Limitation and Challenges
- HandsOn & Discussion

Defining Automatic Speech Recognition

Jurafsky and Martin define the Automatic Speech Recognition (ASR) task as:

- Mapping any waveform like this:

- To the appropriate string of words: It’s time for lunch!

What is ASR model (or system)?

Input

ASR Model

Word Sequence

Output

Data Collection & Preparation

- Computational Modeling
- Architecture Design

Evaluation & Benchmarking

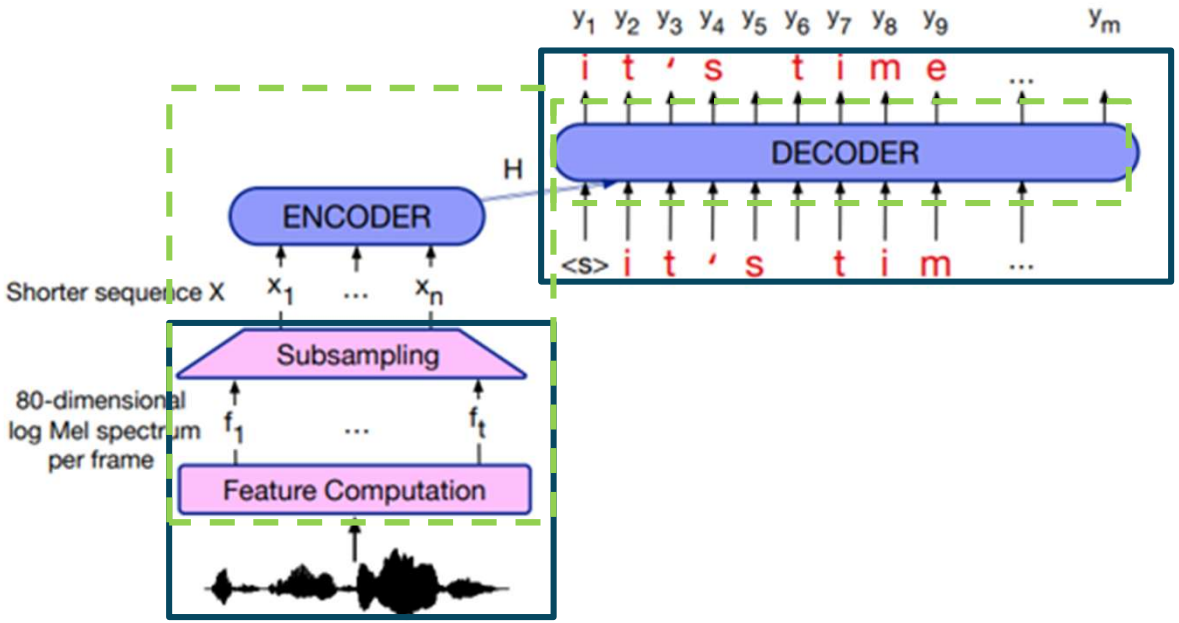
Colliding worlds: Linguistics & Computer Science

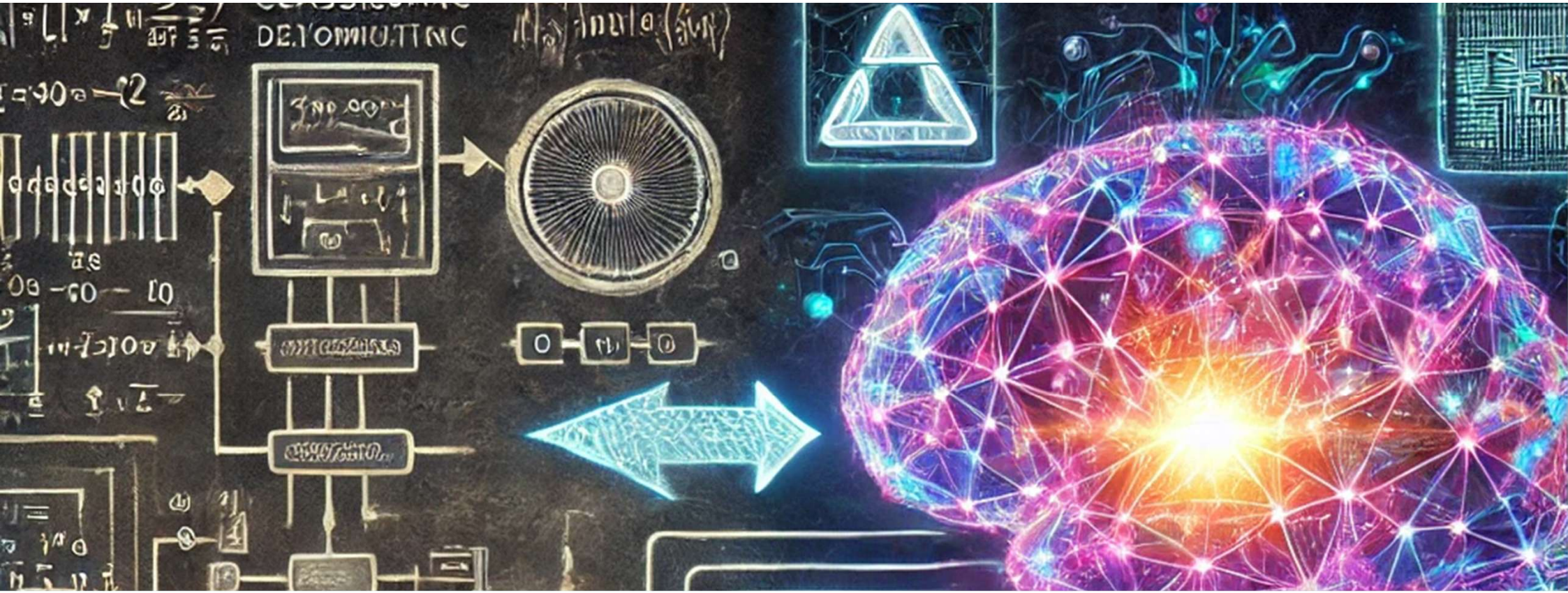
- Data Collection & Preparation
- Computational Modeling
- Architecture Design
- Evaluation & Benchmarking

Lingustics

CS

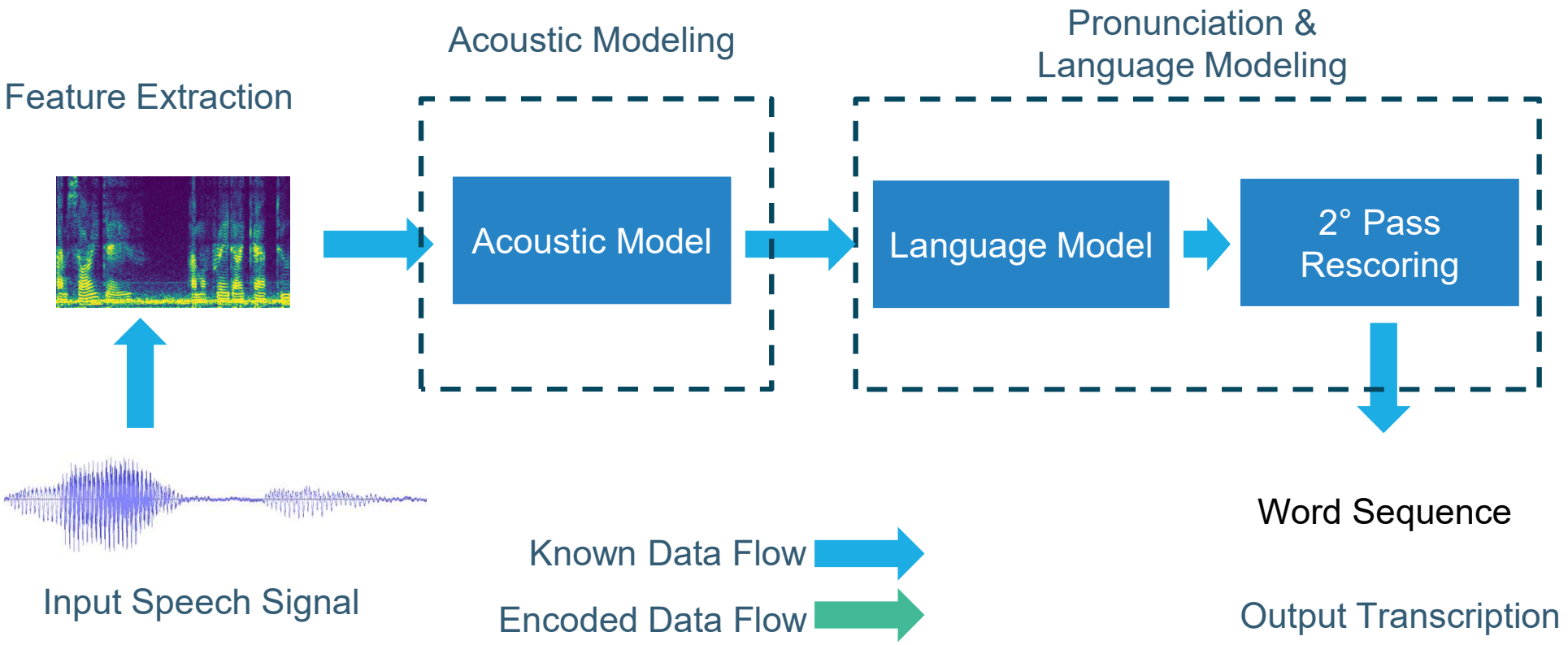
Living together is possible!



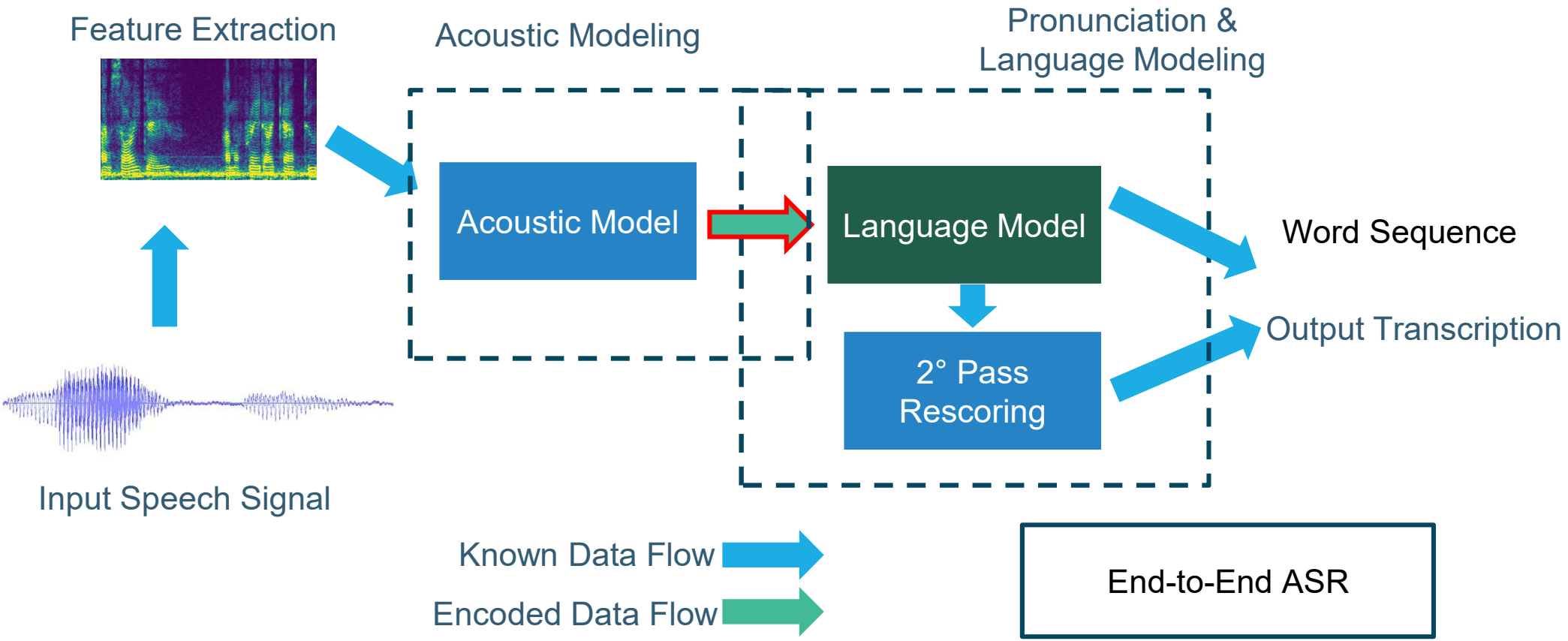


Classic & Modern ASR

Classic ASR Pipeline



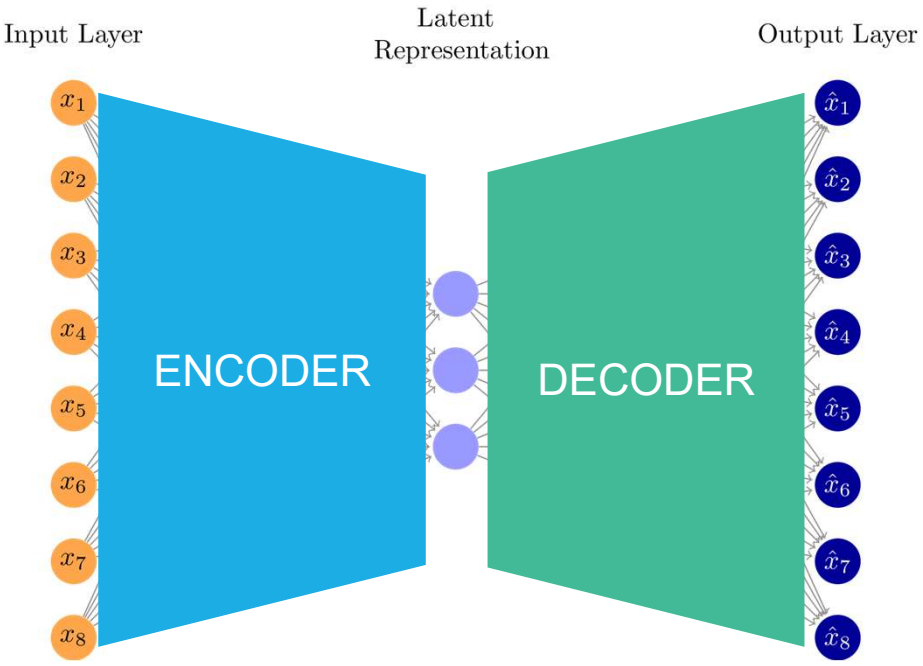
Modern ASR Pipeline

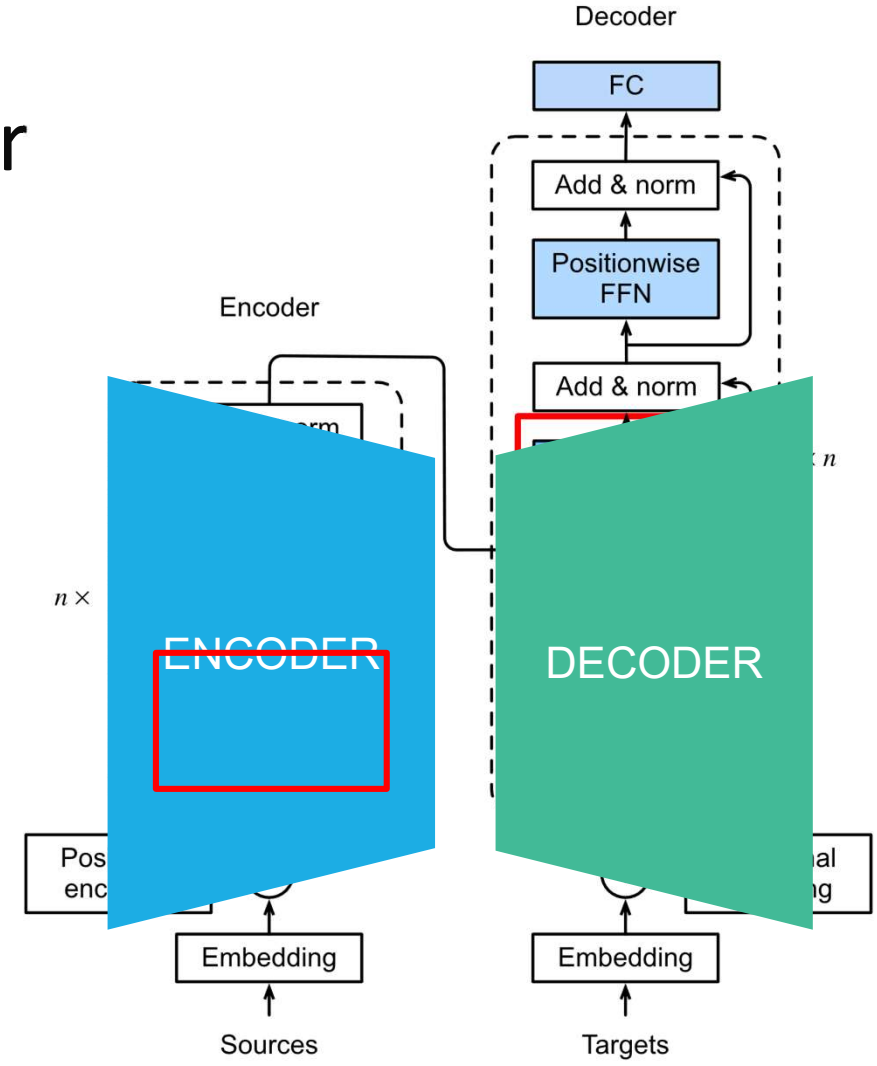


What is and End to End ASR?

	Classic ASR	End-to-End ASR (E2E ASR)
Architecture	Multiple-Components (AM, LM, Lexicon, Phonetic Dictionary)	(Mostly) Single Component, i.e. a Neural Network.
Decoding	Weighted Finite Stated Transducers	Connectionist Temporal Classification (CTC), Sequence-to-sequence with attention, Transducers, AED.
Input	Hand-crafted wave-based features (MFCC) + some learned features.	Raw waveform, hand-crafted wave-based features (MFCC), Spectrograms.
Output	Context based phones and then words.	Characters, Words or Word-parts.

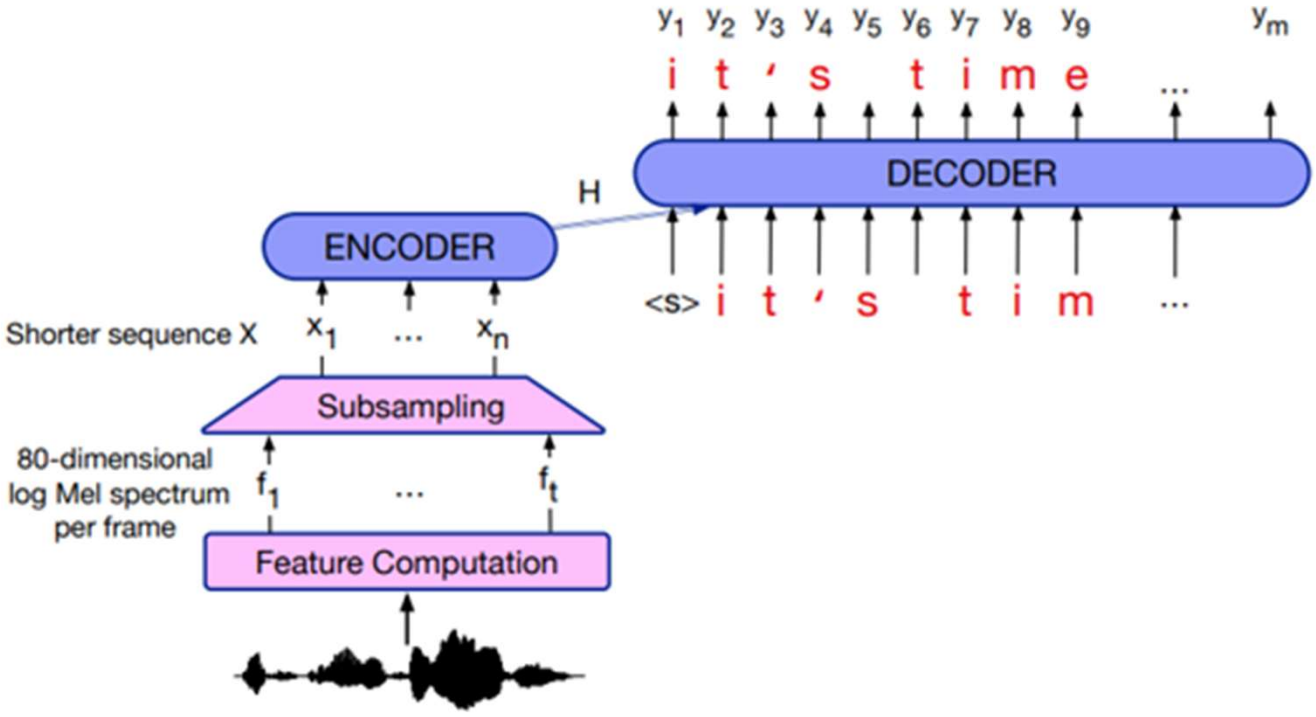
Autoencoder & Transformer





How an End to End ASR works?

1. An Encoder takes the Waveform and extracts a representation for each frame.
2. A Decoder associates a transcription to the compressed representation.



E2E ASR

Ranking

OpenASR Leaderboard

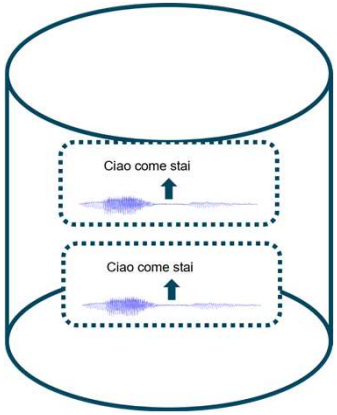
model	Average WER	RTFx	AMI	Earnings22	Gigaspeech	LS Clean
nvidia/canary-1b	6.5	235.34	13.9	12.19	10.12	1.48
nyrahealth/CrisperWhisper	6.67	84.05	8.71	12.89	10.24	1.82
nvidia/parakeet-tdt-1.1b	7.01	2390.61	15.87	14.49	9.52	1.4
nvidia/parakeet-rnnt-1.1b	7.12	2053.15	17.01	13.94	9.89	1.45
nvidia/parakeet-ctc-1.1b	7.4	2728.52	15.67	13.75	10.28	1.83
openai/whisper-large-v3	7.44	145.51	15.95	11.29	10.02	2.01
nvidia/parakeet-tdt_ctc-110m	7.49	5345.14	15.89	12.37	10.52	2.4
nvidia/parakeet-rnnt-0.6b	7.5	2815.72	17.4	14.66	10.01	1.62
distil-whisper/distil-large-v3	7.52	214.42	15.16	11.79	10.08	2.54
nvidia/parakeet-ctc-0.6b	7.69	4281.53	16.46	14.26	10.39	1.88
openai/whisper-large-v2	7.83	144.45	16.74	12.05	10.67	2.83

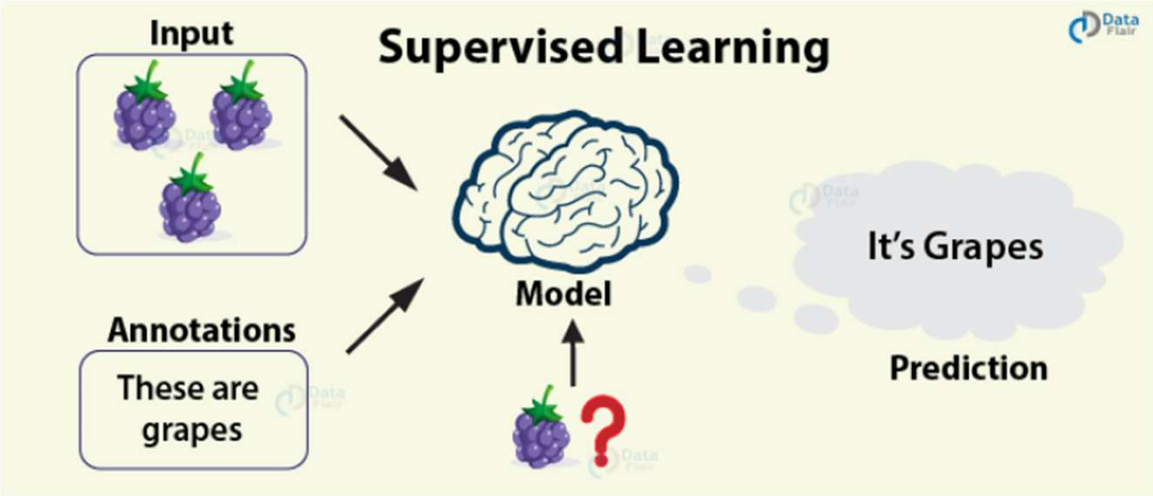
(Fast)Conformer

Conformer
Squeezeformer
FastConformer
Parakeet
Canary

Supervised Models

- Labeled data
- Task Oriented
- Models the relationships between two variables to predict a new outcome.





Supervised Learning - Source

Conformer

Based on Transformer

Input processing:

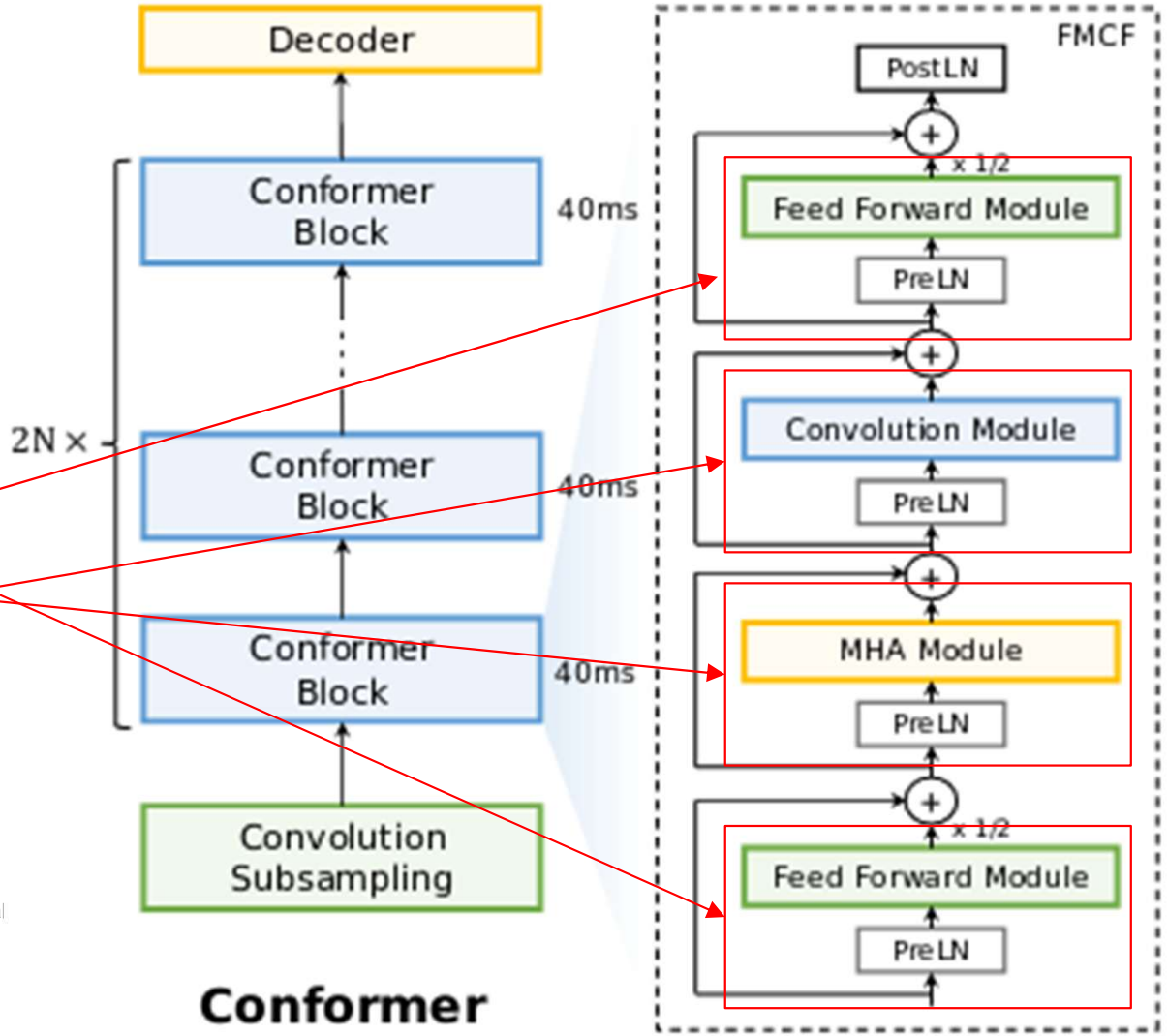
- Usually takes MFCC features

Conformer Blocks:

- Macaron Style FeedForward
- Multi-head self-attention
- Convolution Module

Advantages:

- Attention - Global phenomena
- Convolution - Local phenomena



Gulati, Anmol, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han et al. "Conformer: Convolution-augmented transformer for speech recognition." arXiv preprint arXiv:2005.08100 (2020).

Squeezeformer - why?

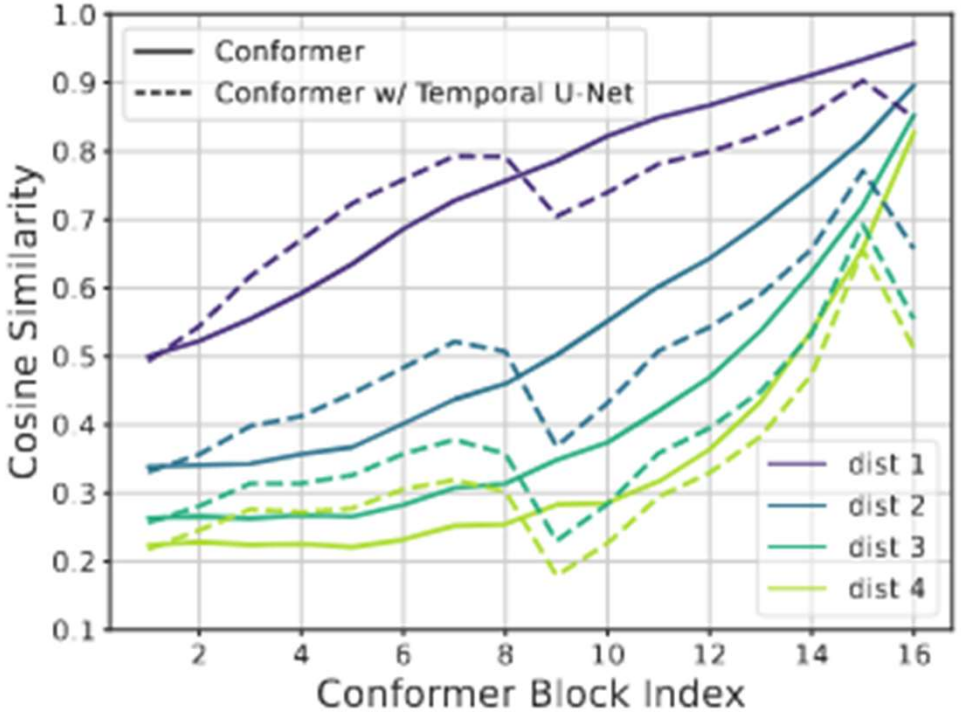
The intuition behind

Conformer attention on 40ms frames:

- Redundant information in intermediate blocks.
- Requires many resources and time.

So?

Compute attention on larger frames.

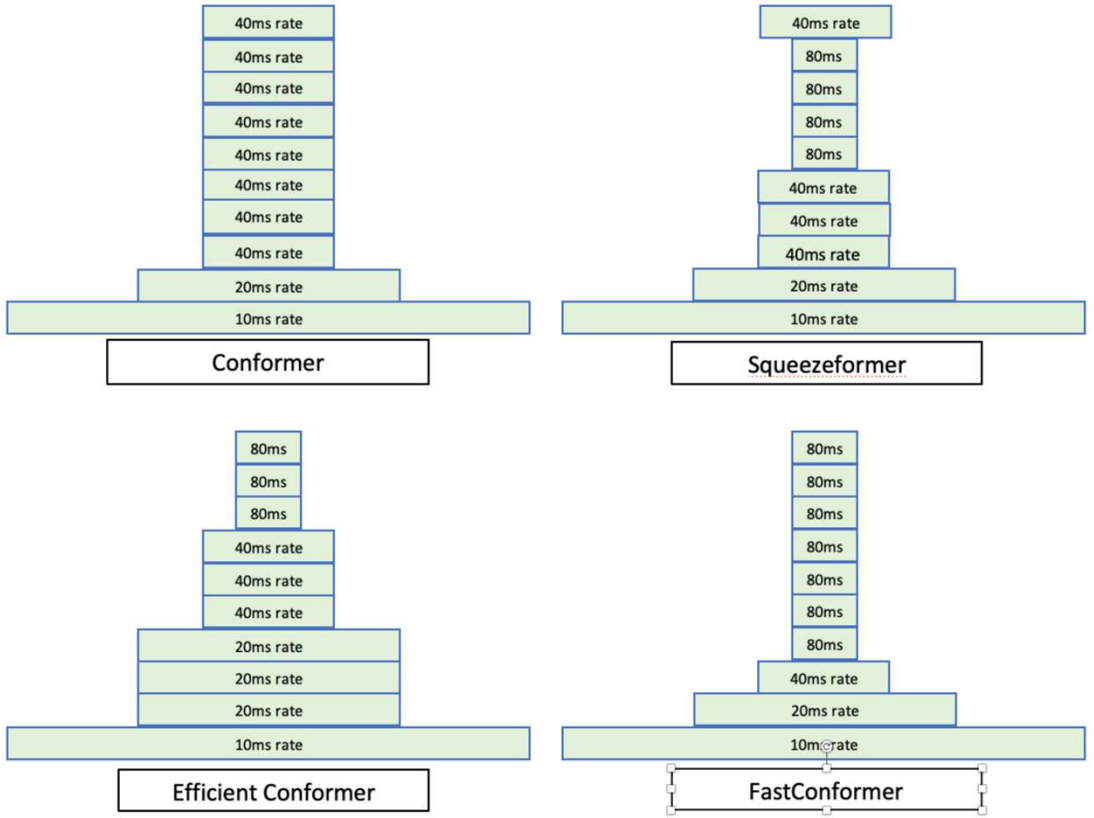


Conformer-based Encoders

The downsampling schemas for:

- Conformer
- SqueezeFormer
- Efficient Conformer
- Fast Conformer

The additional 2x reduction in the encoder output length versus Conformer yields further compute-memory savings in RNNT decoder.



Rekesh, D., Koluguri, N. R., Kriman, S., Majumdar, S., Noroozi, V., Huang, H., ... & Ginsburg, B. (2023, December). Fast conformer with linearly scalable attention for efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 1-8). IEEE.

Conformer & Squeezeformer vs Others

Model	dev-clean	dev-other	test-clean	test-other	Params (M)	GFLOPs	Thp (ex/s)
Conformer-CTC-S* [16]	4.21	10.54	4.06	10.58	8.7	26.2	613
QuartzNet 5x5 [27]	5.39	15.69	-	-	6.7	20.2	-
Citrinet 256 [36]	-	-	3.78	9.60	10.3	16.8	-
Squeezeformer-XS	**3.63**	**9.30**	**3.74**	**9.09**	9.0	15.8	763
Conformer-CTC-M* [16]	2.94	7.80	3.20	7.90	27.4	71.7	463
QuartzNet 5x10 [27]	4.14	12.33	-	-	12.8	38.5	-
QuartzNet 5x15 [27]	3.98	11.58	3.90	11.28	18.9	55.7	-
Citrinet 512 [36]	-	-	3.11	7.82	37.0	63.1	-
Eff. Conformer-CTC† [4]	-	-	3.57	8.99	13.2	26.0	-
Eff. Conformer-CTC‡ [4]	-	-	3.58	8.88	13.2	32.5	-
Squeezeformer-S	2.80	7.49	3.08	7.47	18.6	26.3	602
Squeezeformer-SM	**2.71**	**6.98**	**2.79**	**6.89**	28.2	42.7	558
Conformer-CTC-L* [16]	2.61	6.45	2.80	6.55	121.5	280.6	200
Citrinet 1024 [36]	-	-	**2.52**	6.22	143.1	246.3	-
Squeezeformer-M	2.43	6.51	2.56	6.50	55.6	72.0	431
Squeezeformer-ML	**2.34**	**6.08**	2.61	**6.05**	125.1	169.2	268
Transformer-CTC [31]	2.6	7.0	2.7	6.8	255.2	621.1	-
Squeezeformer-L	**2.27**	**5.77**	**2.47**	**5.97**	236.3	277.9	207

Canary

- Architecture:
  - ENCODER: FastConformer
  - DECODER: Transformer
- Multi task:
  - automatic speech-to-text recognition (ASR) in 4 languages (English, German, French, Spanish)
  - translation from English to German/French/Spanish
  - translation from German/French/Spanish to English
- With or without punctuation and capitalization (PnC).

OpenASR Leaderboard

model	Average WER	RTFx	AMI	Earnings22	Gigaspeech	LS Clean
nvidia/canary-1b	6.5	235.34	13.9	12.19	10.12	1.48
nyrahealth/CrisperWhisper	6.67	84.05	8.71	12.89	10.24	1.82
nvidia/parakeet-tdt-1.1b	7.01	2390.61	15.87	14.49	9.52	1.4
nvidia/parakeet-rnnt-1.1b	7.12	2053.15	17.01	13.94	9.89	1.45
nvidia/parakeet-ctc-1.1b	7.4	2728.52	15.67	13.75	10.28	1.83
openai/whisper-large-v3	7.44	145.51	15.95	11.29	10.02	2.01
nvidia/parakeet-tdt_ctc-110m	7.49	5345.14	15.89	12.37	10.52	2.4
nvidia/parakeet-rnnt-0.6b	7.5	2815.72	17.4	14.66	10.01	1.62
distil-whisper/distil-large-v3	7.52	214.42	15.16	11.79	10.08	2.54
nvidia/parakeet-ctc-0.6b	7.69	4281.53	16.46	14.26	10.39	1.88
openai/whisper-large-v2	7.83	144.45	16.74	12.05	10.67	2.83

Conformer Family

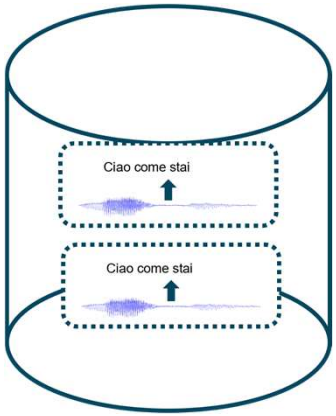
Model	Main Charachteristics	Advantages	Best for
Conformer (2020)	Combines self-attention & convolutions to capture both global and local dependencies.	✅ High accuracy ✅ Good latency management	**Balancing accuracy and latency**
Squeezeformer (2022)	Reduces Convolutions, uses Temporal U-Net, different versions of Attention to reduce complexity and keep quality.	✅ 2-4× faster than Conformer ✅ Lower memory footpring	**ASR on low-resource devices (edge, mobile)**
FastConformer (2022)	Optimized version of Conformer with linear attention and strided convolutions to improve efficiency.	✅ Faster ✅ Low memory usage in inferenfe	**Real-time applications at the edge and mobile**
Canary (2023)	Distilled model optimized for TPUs with pruning and quantization techniques.	✅ Highly scalable ✅ Lower computational cost in the cloud	**Large cloud deployments at scale**

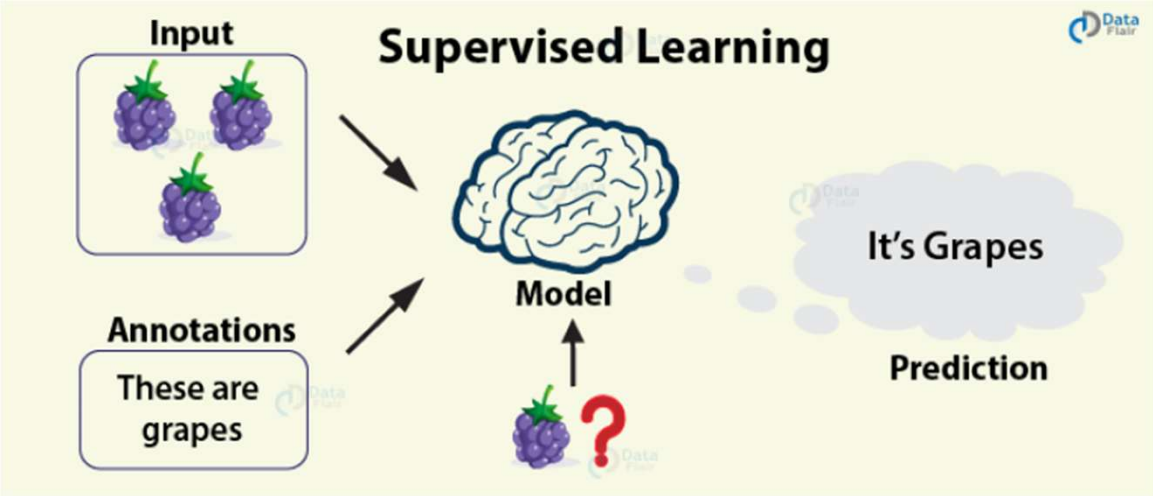
Self-Supervised

Wav2Vec2.0
HUBERT
WavLM

Supervised

- Labeled data
- Task Oriented
- Models the relationships between two variables to predict a new outcome.

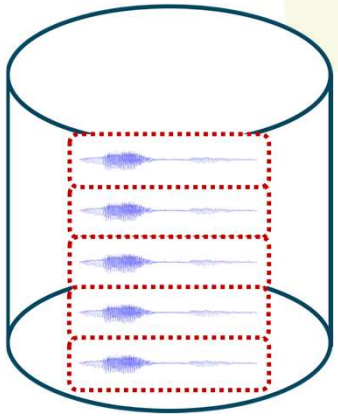


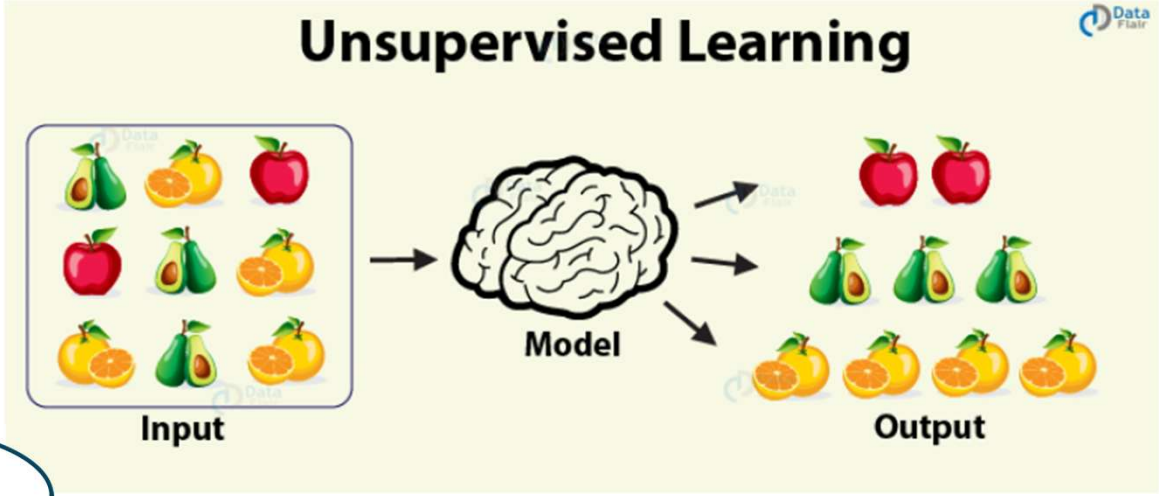


Supervised Learning - Source

Unsupervised

- Unlabeled data, i.e., not explicitly differentiated into classes.
- The model learns by finding implicit patterns.
- An Unsupervised Learning algorithm identifies the data based on:
  - Densitiy
  - Structures
  - Similar segments
  - Other similarities.





Unsupervised Learning - Source

Wav2Vec 2.0

Is the first Self-Supervised ASR model with the following characteristics:

1. Uses a self-supervised mechanism to learn speech representation.
2. Employs a masking mechanism similar to BERT.

Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in Neural Information Processing Systems* 33 (2020): 12449-12460.

Wav2Vec2.0 Model Training

Phase 1

Learn Sound Representation

Data: Unlabeled data

Training: Self-supervised learning

Learn: Speech representation

Phase 2

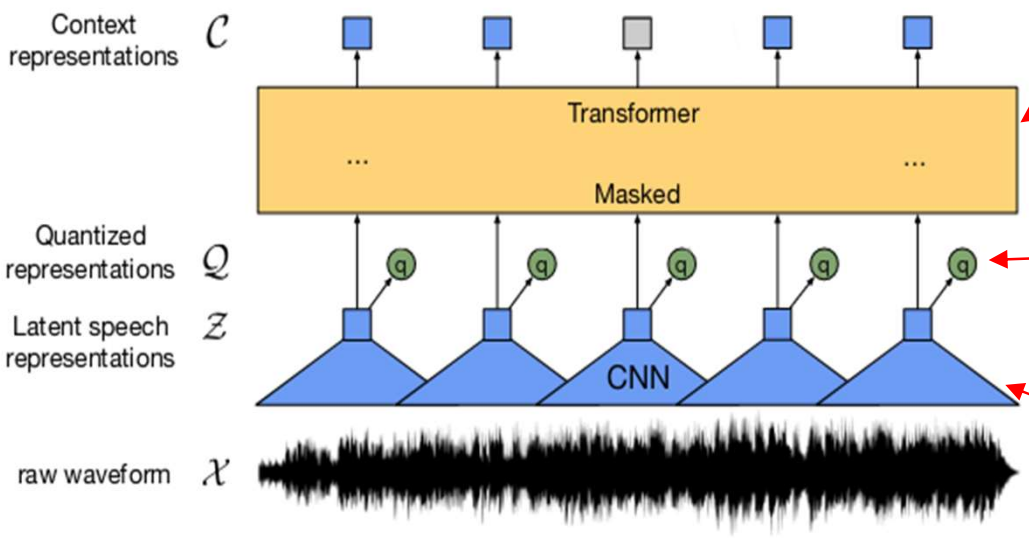
Learn Transcription

Data: Labeled data

Training: Supervised fine-tuning

Learn: Words/Phonemes

Phase 1



Context Network – Reconstruct sound seq.

1. Transformer-based
2. Relative positional embedding

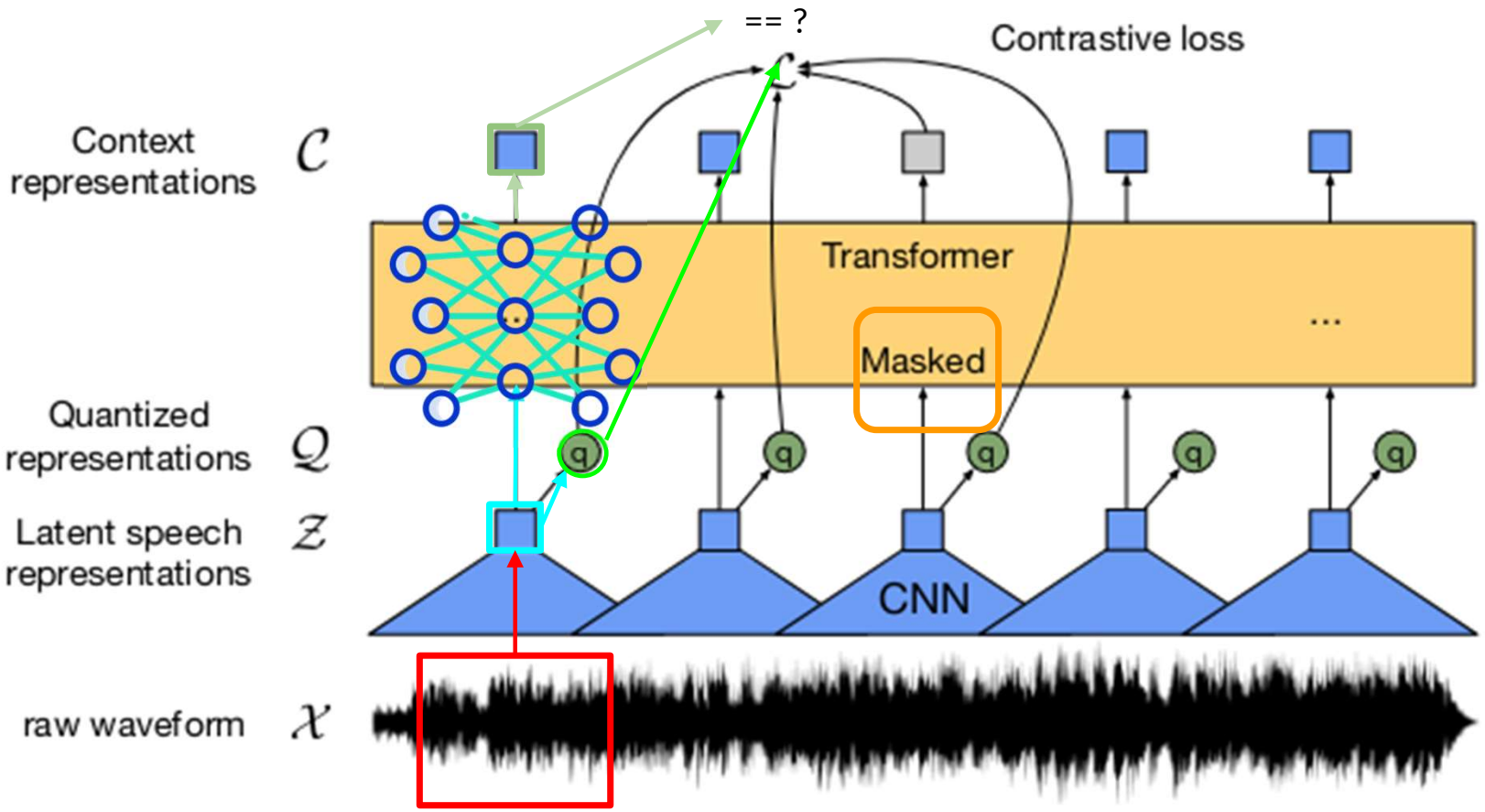
Quantization –Discretize sounds

1. Gumbel Softmax + Finite vocabularies
2. Creates a finite dictionary of sounds

Feature Encoder – Reduce size

1. Input 16Khz -> Output 49hz
2. Sample 20ms

Phase 1


== ?
Contrastive loss
Context representations
$\mathcal{C}$
Transformer
Masked
...
Quantized representations
$\mathcal{Q}$
q
Latent speech representations
$\mathcal{Z}$
CNN
raw waveform
$\mathcal{X}$

Obtain a Self Supervised Model

Phase 1

Learn Sound Representation

Data: Unlabeled data

Training: Self-supervised learning

Learn: Speech representation

Phase 2

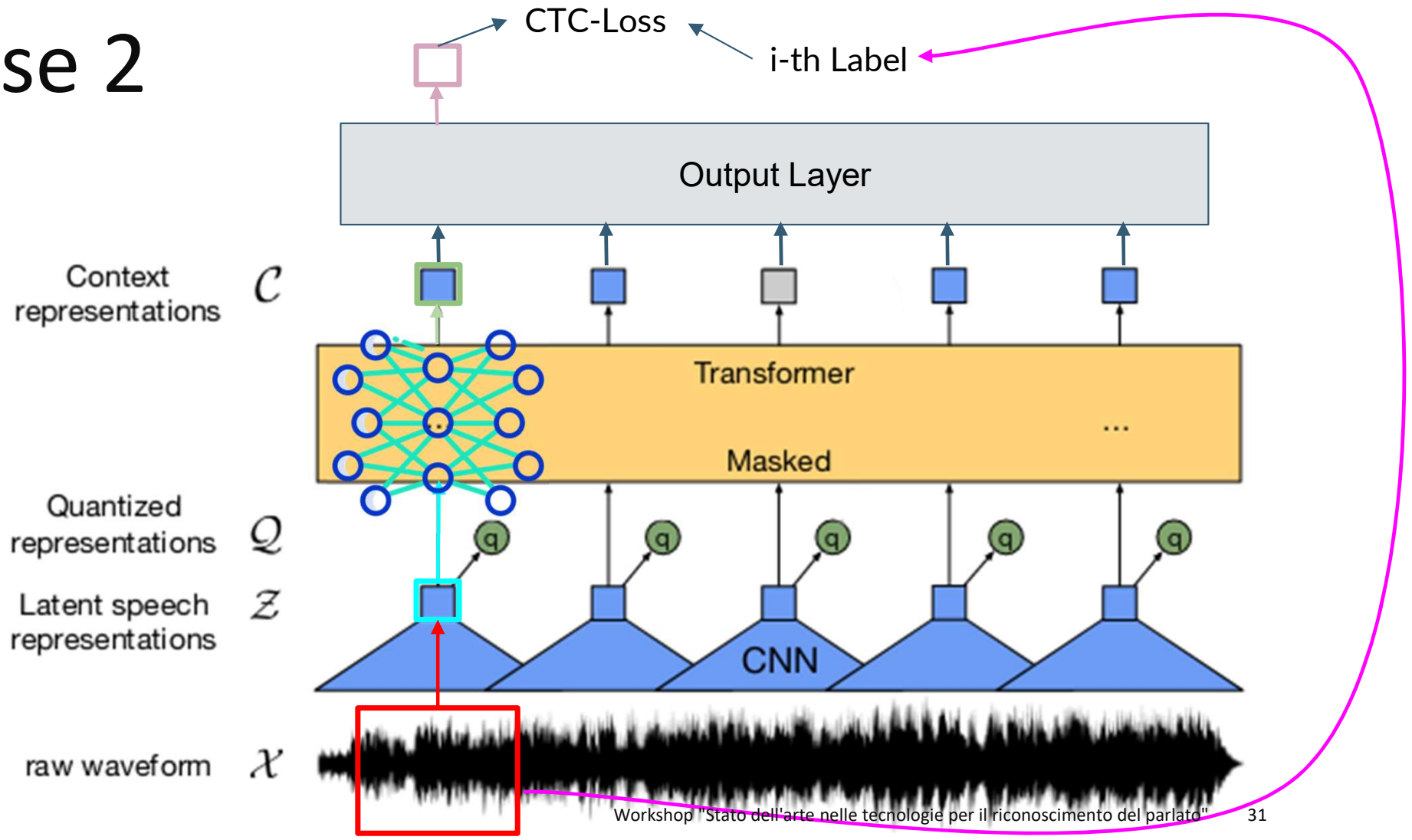
Learn Transcription

Data: Labeled data

Training: Supervised fine-tuning

Learn: Words/Phonemes

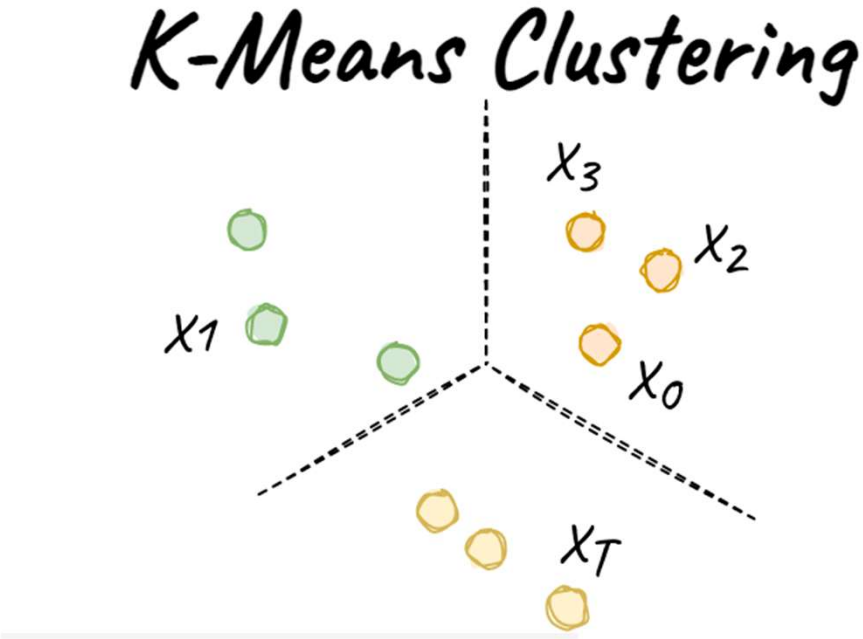
Phase 2


CTC-Loss
i-th Label
Output Layer
Context representations
$\mathcal{C}$
Transformer
...
Masked
Quantized representations
$\mathcal{Q}$
Latent speech representations
$\mathcal{Z}$
q
CNN
raw waveform
$\mathcal{X}$

HUBERT

- Hidden Units BERT
- Similar to Wav2Vec2.0
- Finds Hidden Unit through clustering: K-means
- Units number (i.e., sounds) Fixed to 100 (at first step)

Linguists



Self-Supervised Performance

High Dimensions:

- Wav2Vec2.0: 317Mln param
- HuBert: 316 Mln param
- WavlM: 316Mln param

Computing resources.

Training Time.

Conformer Large has 120Mln!

Model	Unlabled Data	LM	test-clean	test-other
	100-hour labeled			
wav2vec 2.0 Base	LS-960	None	6.1	13.3
WavLM Base	LS-960	None	5.7	12.0
WavLM Base+	MIX-94k	None	4.6	10.1
DeCoAR 2.0	LS-960	4-gram	5.0	12.1
DiscreteBERT	LS-960	4-gram	4.5	12.1
wav2vec 2.0 Base	LS-960	4-gram	3.4	8.0
HuBERT Base	LS-960	4-gram	3.4	8.1
WavLM Base	LS-960	4-gram	3.4	7.7
WavLM Base+	MIX-94k	4-gram	2.9	6.8
wav2vec 2.0 Large	LL-60k	4-gram	2.3	4.6
WavLM Large	MIX-94k	4-gram	2.3	4.6
wav2vec 2.0 Large	LL-60k	Transformer	2.0	4.0
HuBERT Large	LL-60k	Transformer	2.1	3.9
WavLM Large	MIX-94k	Transformer	2.1	4.0

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., ... & Wei, F. (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1505-1518.

Self Supervised Comparison

Model	Pre-training	Learning objective	Advantages	Limitations
Wav2Vec2.0	Contrastive Learning on quantized features	Predict features missing from masked audio	✅ Excellent for ASR ✅ Works with little labeled data	❌ Doesn't leverage advanced semantic clusters
HuBERT	Masked Prediction on features obtained from clustering k-means	Predict features missing from masked audio	✅ Better robustness than Wav2Vec2 ✅ Requires less labeled data	❌ Clustering phase depends on hyperparameters
WavLM	Masked Prediction plus Noisy Student Training	Predict missing features + improve generalization	✅ Best for ASR, speaker verification and diarization ✅ Resistant to noise and voice overlap	❌ Computationally heavier

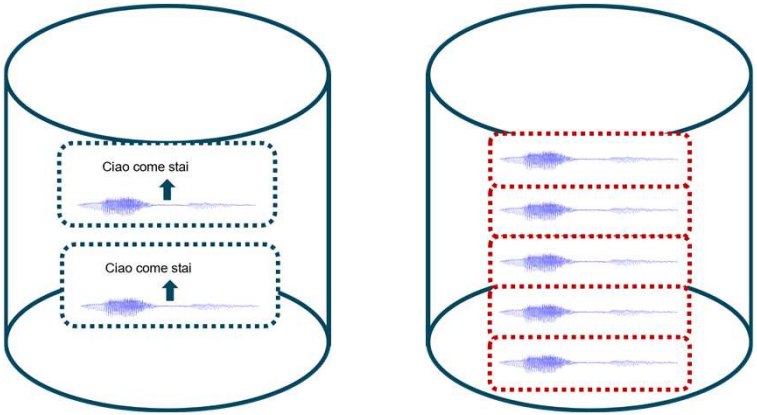
Whisper

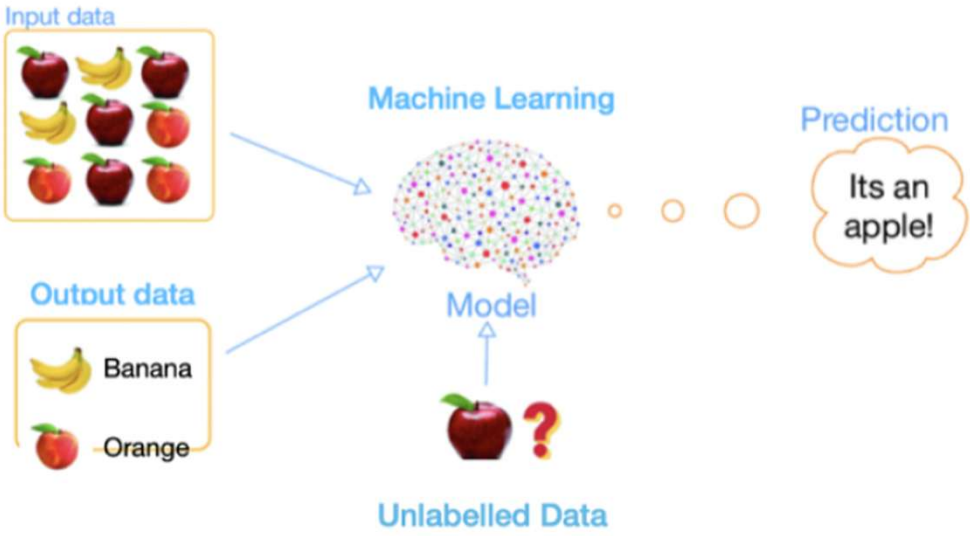
Foundational models

Semi-Supervised

Labeled data and un-labeled data.

1. Uses known data and unsupervised learning to label unknown data.
2. Trains the model on a **pseudo-labeled** dataset.

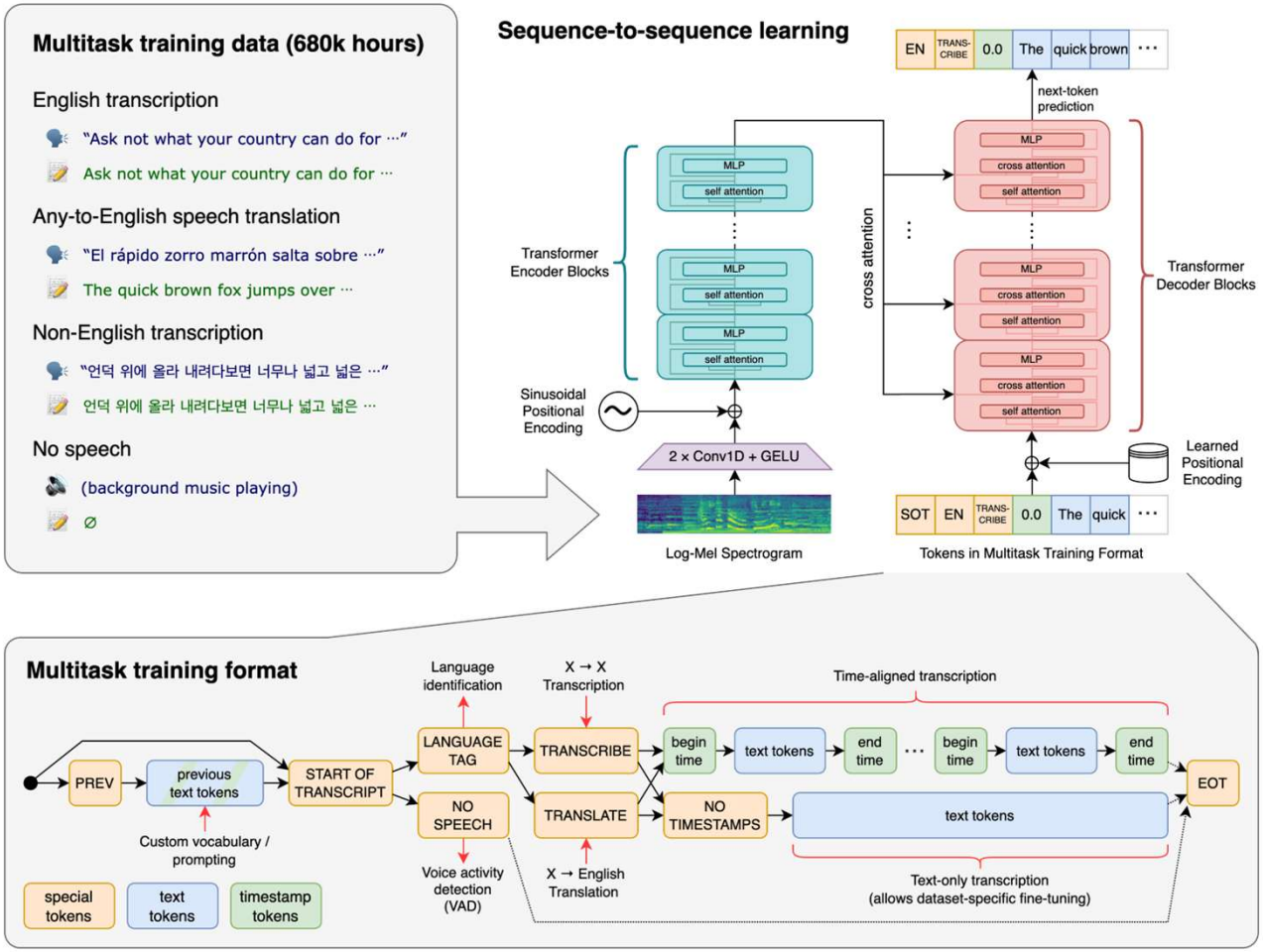




Semi-Supervised Learning - Source

Foundational Models

- Multi-task
- Whisper, Llama, GPT, GEMINI



OpenASR Leaderboard

model	Average WER	RTFx	AMI	Earnings22	Gigaspeech	LS Clean
nvidia/canary-1b	6.5	235.34	13.9	12.19	10.12	1.48
nyrahealth/CrisperWhisper	6.67	84.05	8.71	12.89	10.24	1.82
nvidia/parakeet-tdt-1.1b	7.01	2390.61	15.87	14.49	9.52	1.4
nvidia/parakeet-rnnt-1.1b	7.12	2053.15	17.01	13.94	9.89	1.45
nvidia/parakeet-ctc-1.1b	7.4	2728.52	15.67	13.75	10.28	1.83
openai/whisper-large-v3	7.44	145.51	15.95	11.29	10.02	2.01
nvidia/parakeet-tdt_ctc-110m	7.49	5345.14	15.89	12.37	10.52	2.4
nvidia/parakeet-rnnt-0.6b	7.5	2815.72	17.4	14.66	10.01	1.62
distil-whisper/distil-large-v3	7.52	214.42	15.16	11.79	10.08	2.54
nvidia/parakeet-ctc-0.6b	7.69	4281.53	16.46	14.26	10.39	1.88
openai/whisper-large-v2	7.83	144.45	16.74	12.05	10.67	2.83

Whisper vs Self Supervised

Model	Pre-training	Learning objective	Advantages	Limitations
Wav2Vec2.0	Contrastive Learning on quantized features	Predicting the future frame from discrete features	✅ Excellent for ASR ✅ Works with little labeled data	❌ Doesn't leverage advanced semantic clusters
HuBERT	Masked Prediction on features obtained from clustering k-means	Predict features missing from masked audio	✅ Better robustness than Wav2Vec2 ✅ Requires less labeled data	❌ Clustering phase depends on hyperparameters
WavLM	Masked Prediction plus Noisy Student Training	Predict missing features + improve generalization	✅ Best for ASR, speaker verification and diarization ✅ Resistant to noise and voice overlap	❌ Computationally heavier
Whisper	Semi-Supervised (Pretraining Unlabeled + Fine-Tuning Supervised)	High amounts of data for multi-lingual multi-objective	✅ Multiple languages ✅ Robus wrt noise and accents ✅ Text-aware output	❌ Computationally heavier and less-flexible

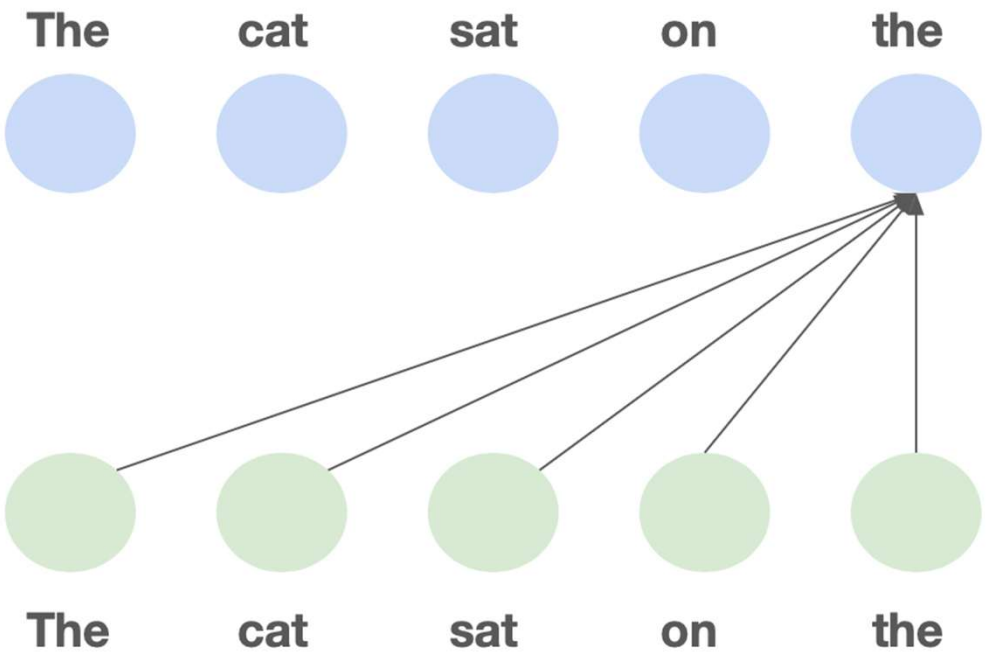
Mamba-based model

A promising alternative to transformers?

Transformer the Core Problem

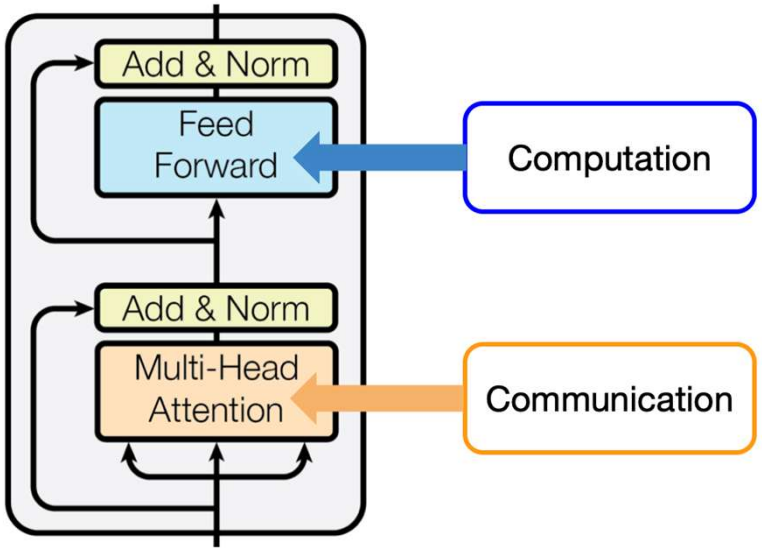
- Every token can look back at every previous token when making predictions.
- A forward pass is $O(n^2)$ time complexity in training (the dreaded quadratic bottleneck)
- Each new token generated autoregressively takes O(n) time.

Do we need to look back at **every** previous token?

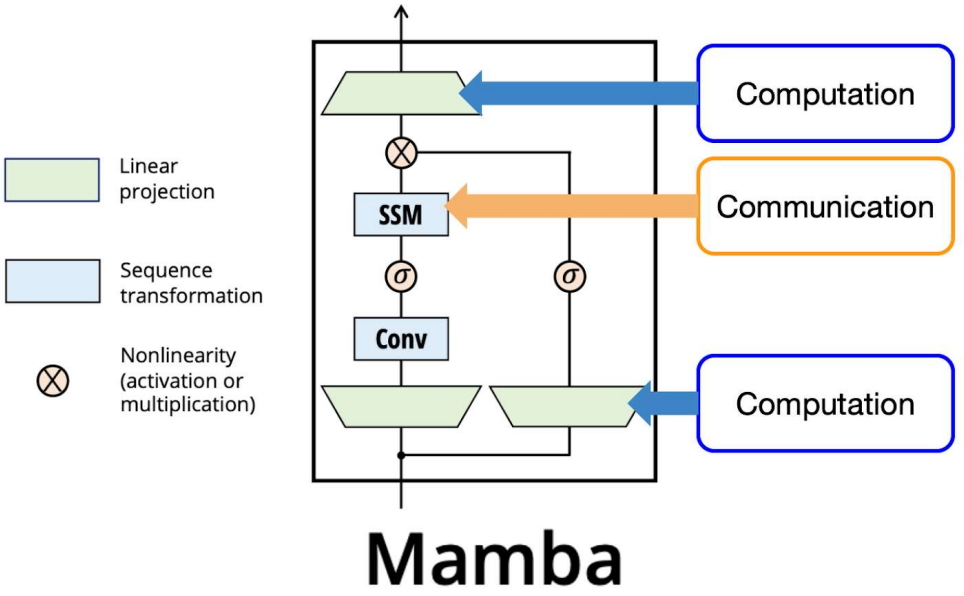


Transformer vs Mamba

Transformer



Mamba



SAMBA-ASR

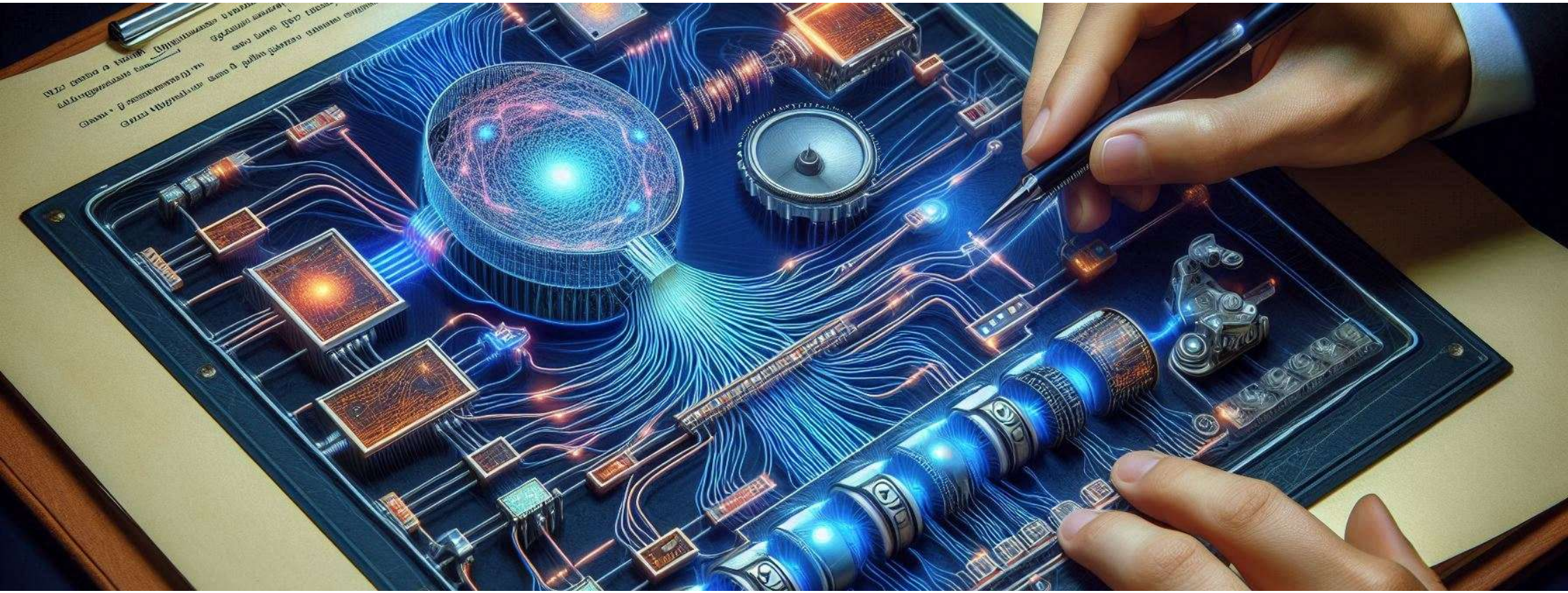
- Based on Mamba
- Modeling local and global temporal dependencies with efficient state dynamics (State-Space Modeling)

Model	Average WER	Gigaspeech	LS Clean	LS Other	SPGISpeech
Samba-ASR (SandLogic)	**3.65**	**9.12**	**1.17**	**2.48**	**1.84**
nvidia/canary-1b	4.15	10.12	1.48	2.93	2.06
nyrahealth/CrisperWhisper	4.69	10.24	1.82	4.00	2.7
nvidia/parakeet-tdt-1.1b	7.01	9.52	1.40	2.60	3.16
openai/whisper-large-v3	7.44	10.02	2.01	3.91	2.94

Shakhadri, S. A. G., KR, K., & Angadi, K. B. (2025). Samba-asr state-of-the-art speech recognition leveraging structured state-space models. *arXiv preprint arXiv:2501.02832*.

Whisper vs WavLM vs Samba

Model	Pre-training	Learning objective	Advantages	Limitations
WavLM	Masked Prediction plus Noisy Student Training	Predict missing features + improve generalization	✅ Best for ASR, speaker verification and diarization ✅ Resistant to noise and voice overlap	❌ Computationally heavier
Whisper	Semi-Supervised (Pretraining Unlabeled + Fine-Tuning Supervised)	High amounts of data for multi-lingual multi-objective	✅ Multiple languages ✅ Robus wrt noise and accents ✅ Text-aware output	❌ Computationally heavier and less-flexible
Samba	Supervised (Fully Supervised Learning)	Modeling local and global temporal dependencies with efficient state dynamics (State-Space Modeling)	✅ Greater computational efficiency ✅ Superior in ASR compared to Transformers	❌ New, requires further validation

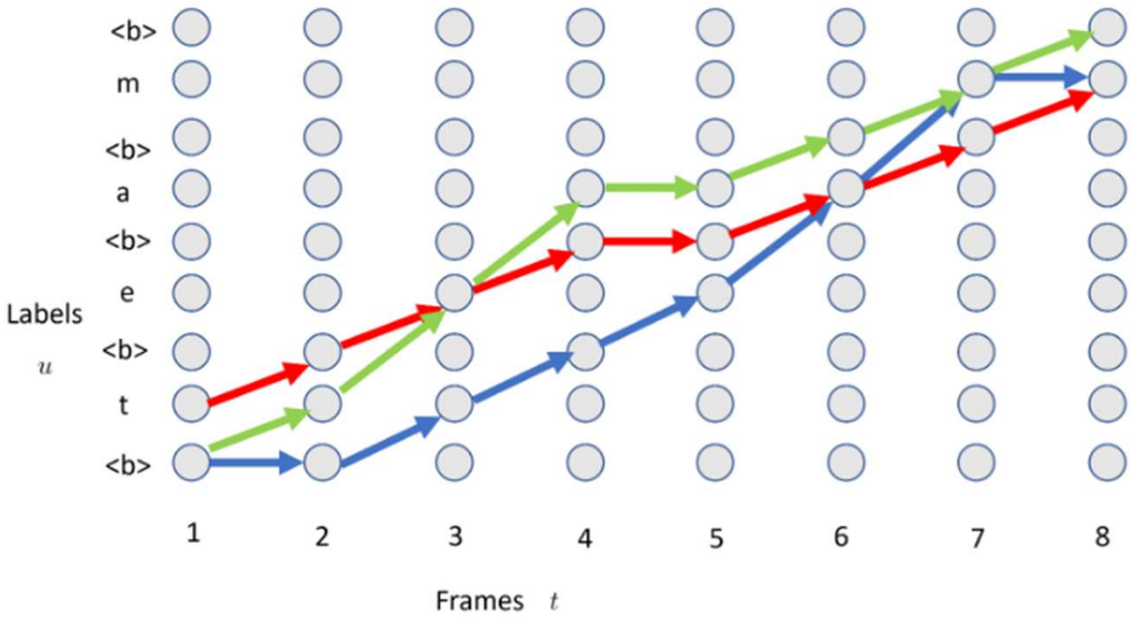


Decoders

CTC
Transducers
Transformer

Connectionist Temporal Classification (CTC)

- ADDS a special "_" character for alignments
- X length >= Y length
- Multiple alignment for same X-Y
- **Non-autoregressive** Outputs assumed to be independent - $Y_{t=1}$ independent from $Y_{t=0}$



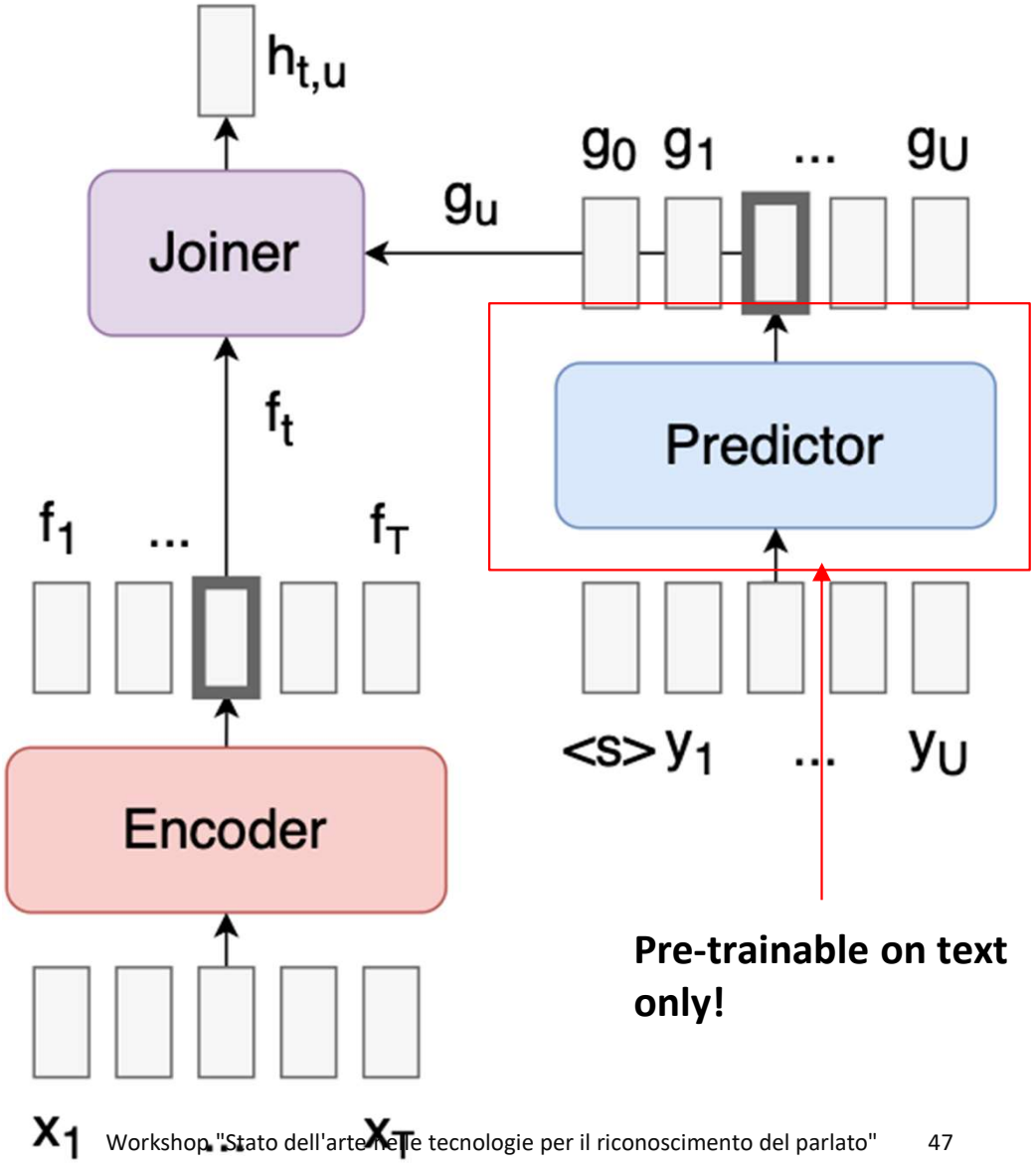
Transducer Architecture

Solves CTC problems

1. Multiple outputs for each input.
2. Adds Predictor and Joiner networks.

Components

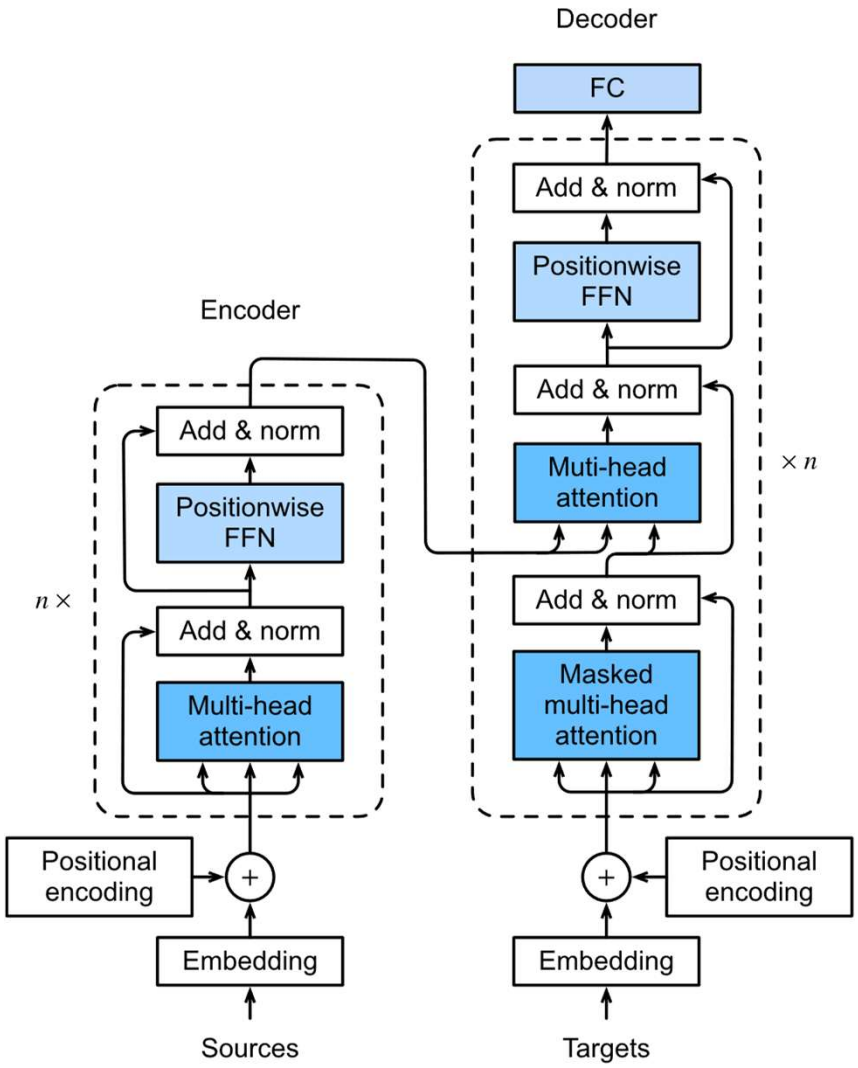
- Predictor is **autoregressive** and works like a standard language model.
- Joiner a simple NN combining inputs.

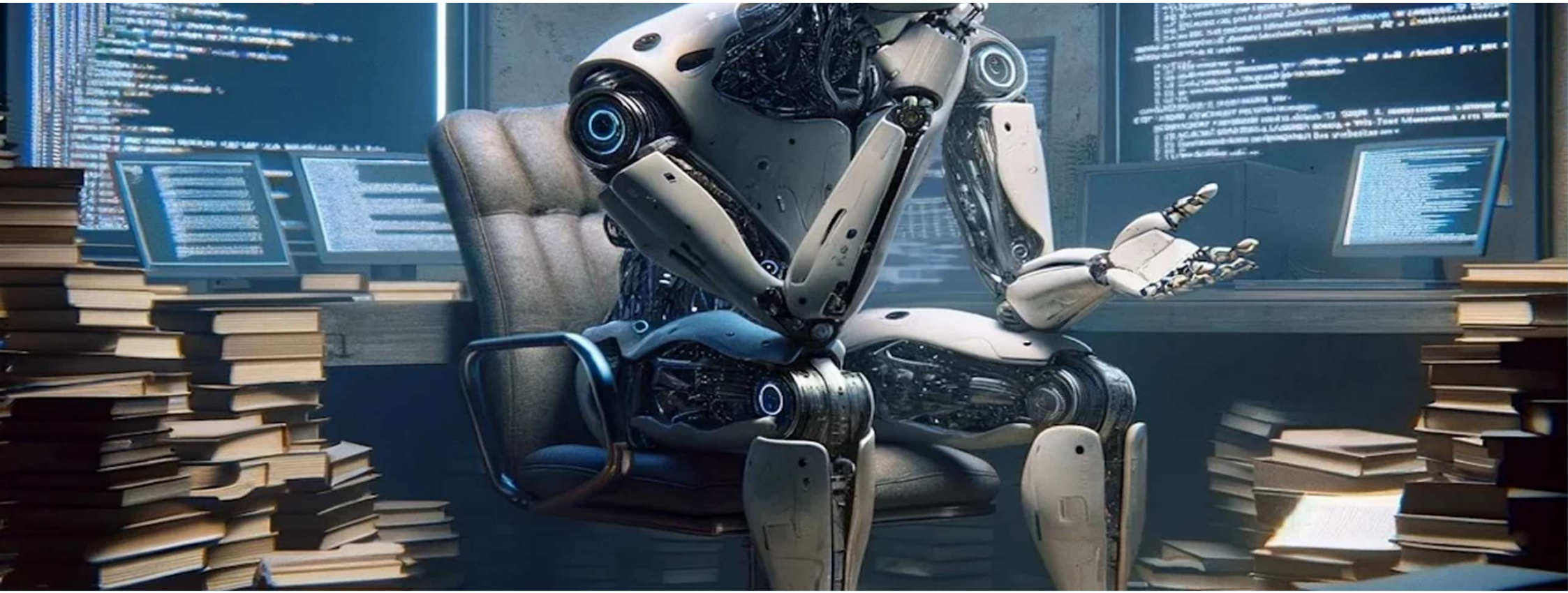


Transformer

Disadvantages WRT CTC and Transducer:

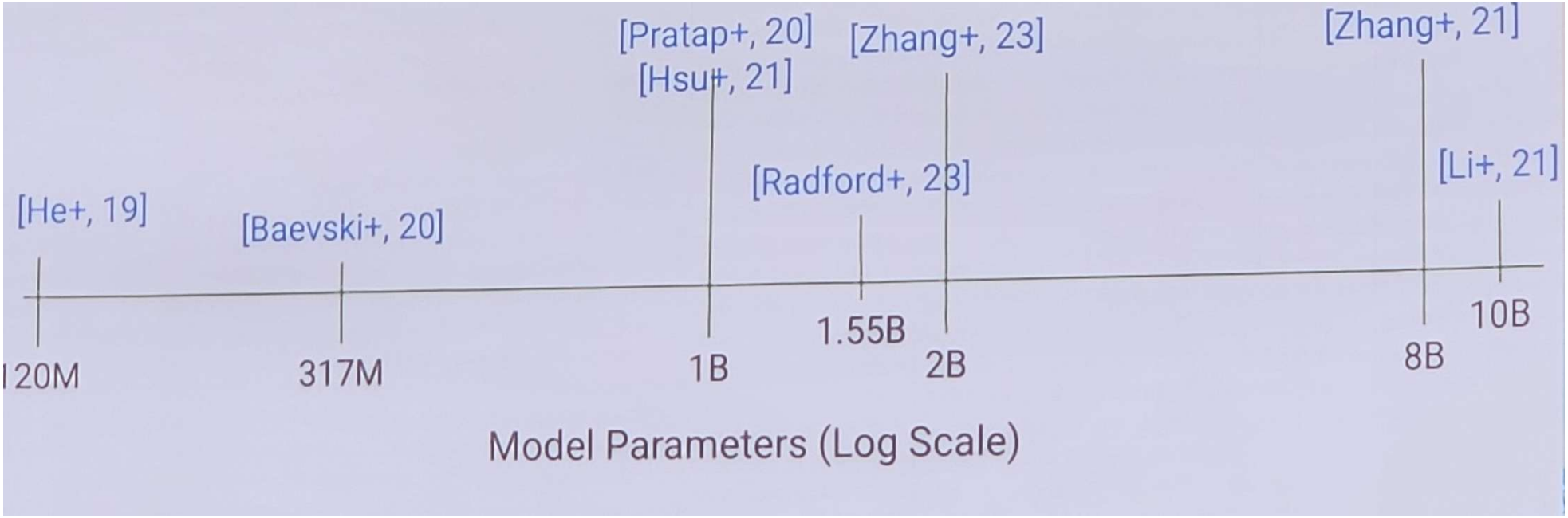
- High computational requirements
- Latency
- Limitations regarding OOV words





Limitations and Challenges

Size of Models



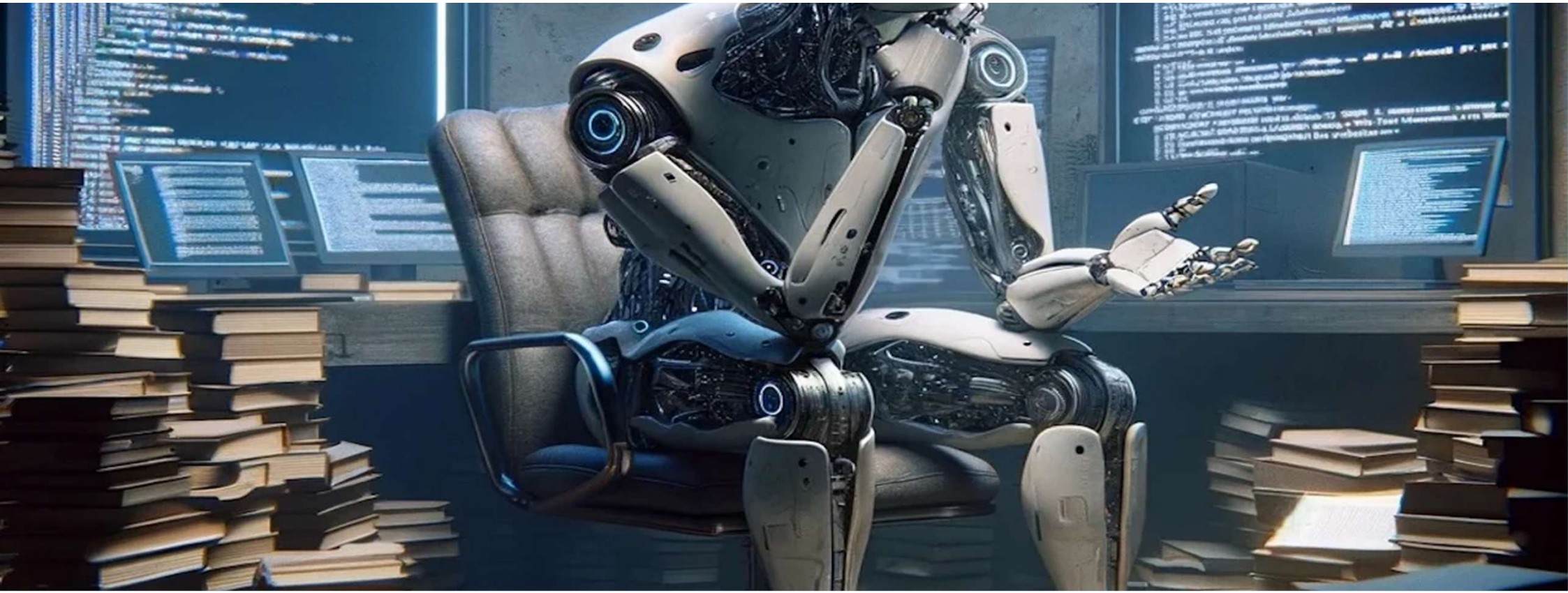
Math meets Language

Improvements

- Improve base unit identification (now 20ms)
- Increase linguistical knowledge
- Use and refinement of ASR as data preparation tool:
  - Wav2Vec2.0 is really wrong?

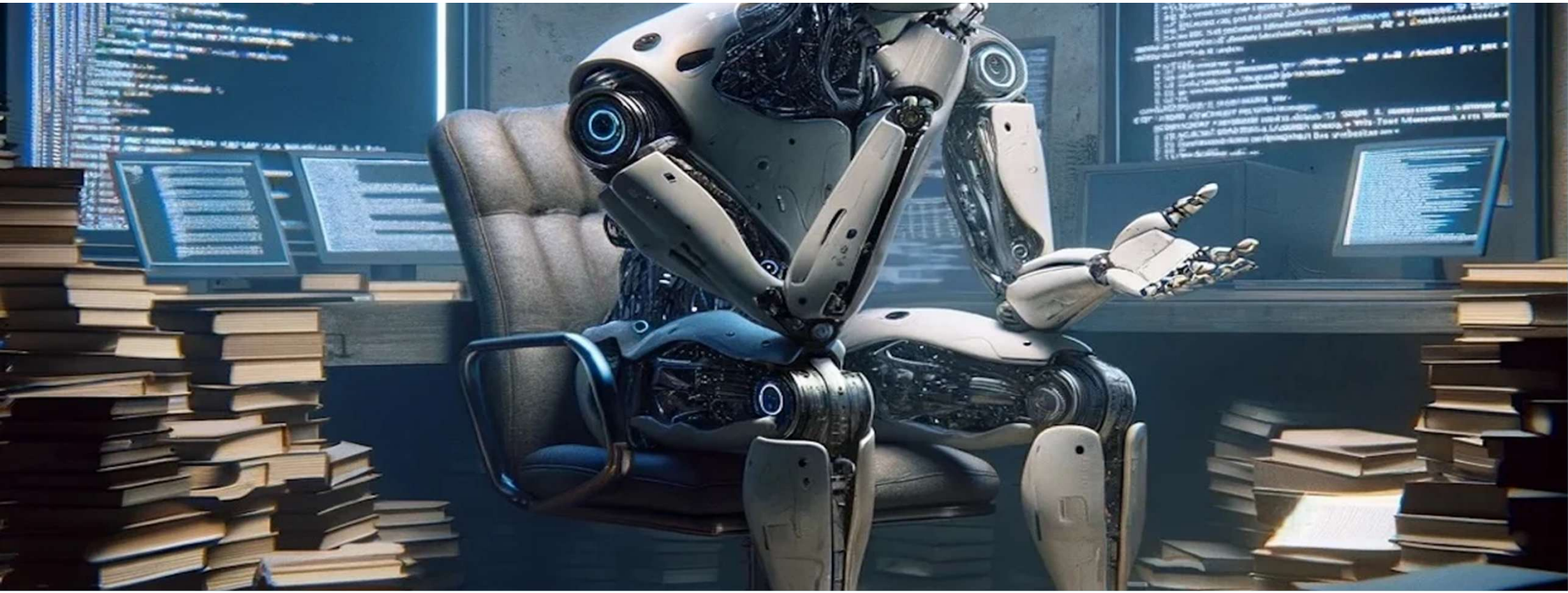
Interpretability

- Crisper Whisper
  - Fine tuning and prompt
  - Not Annotated data
- RichSpeech:
  - Probing
  - Annotated data



Hands-on

See the notebook



Questions?

Thank you for your attention!

Questions?