高等机器学习

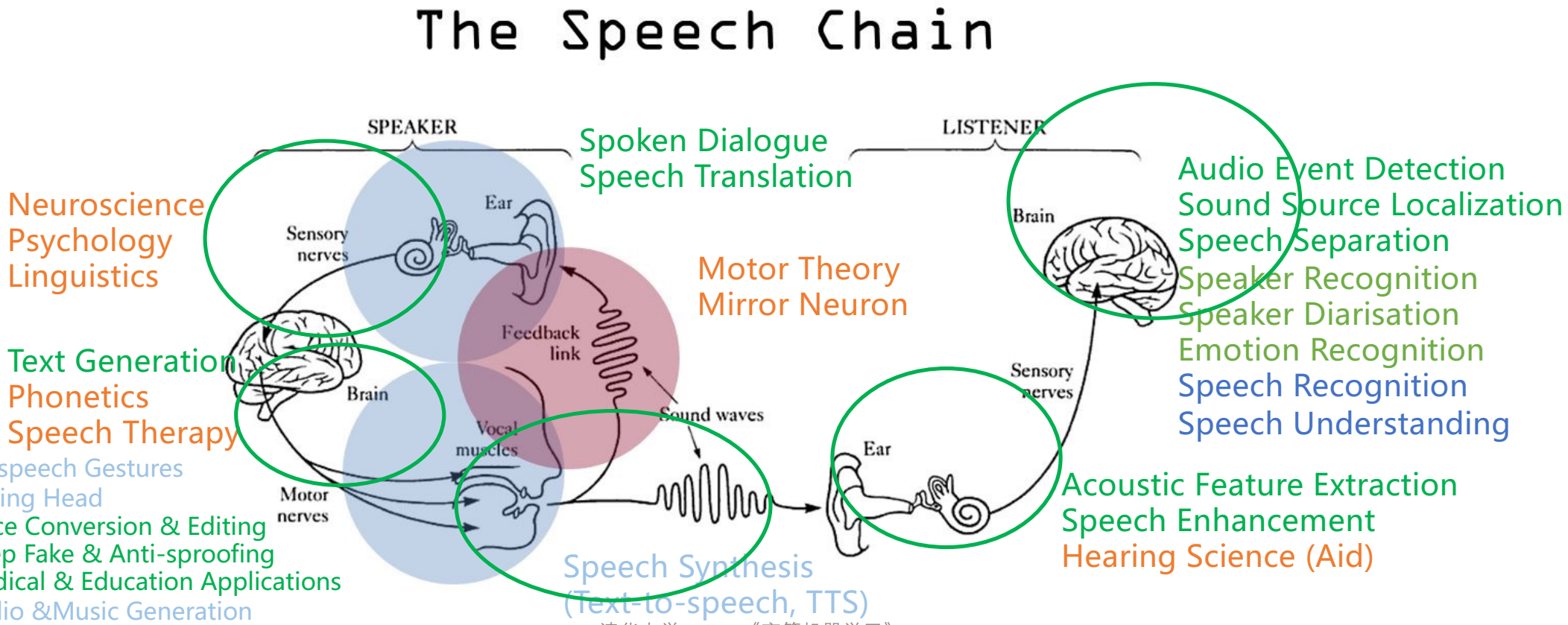
语音处理(下)

张超（图信所）
清华大学 电子工程系

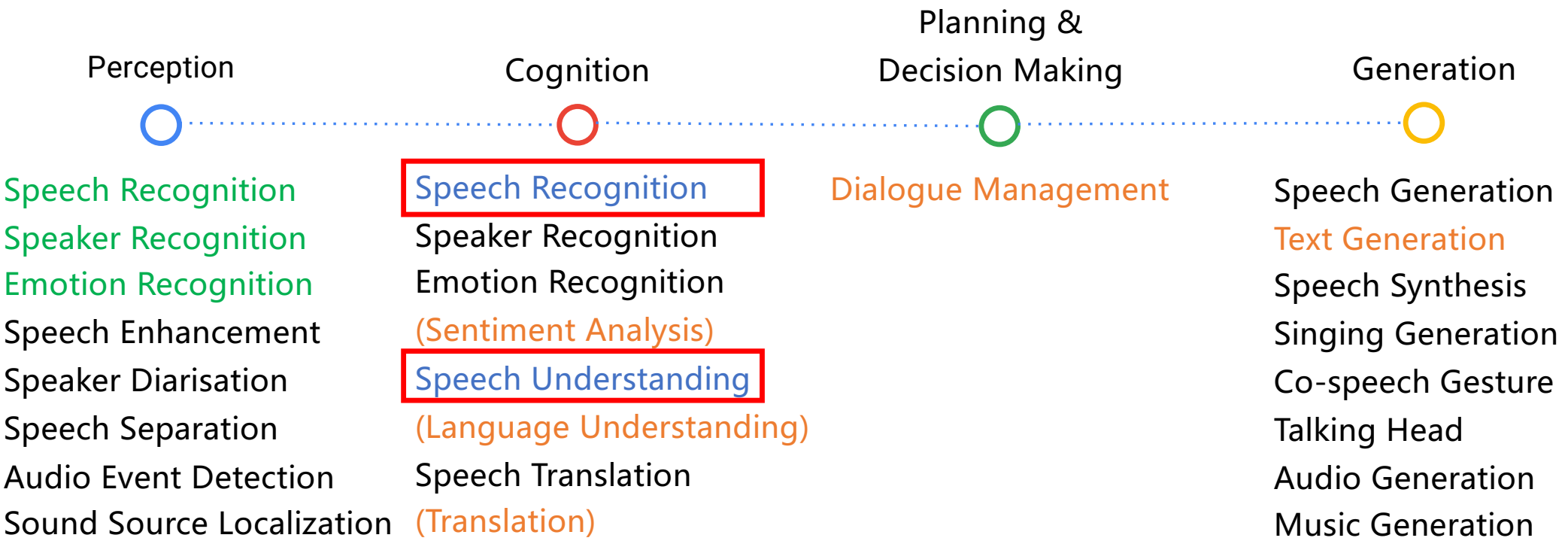
Microsoft

清華大學
Tsinghua University

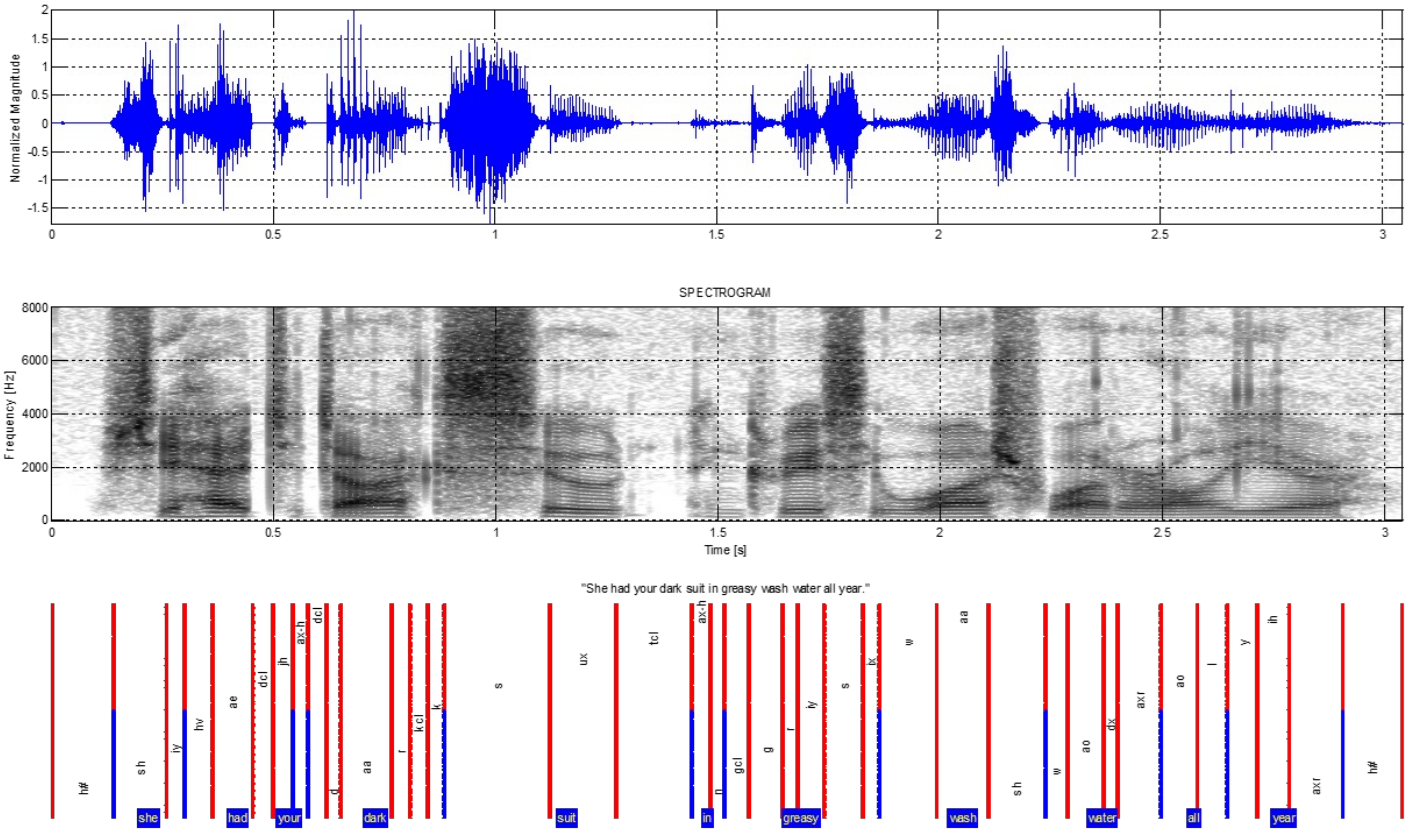
Speech Science & Technology



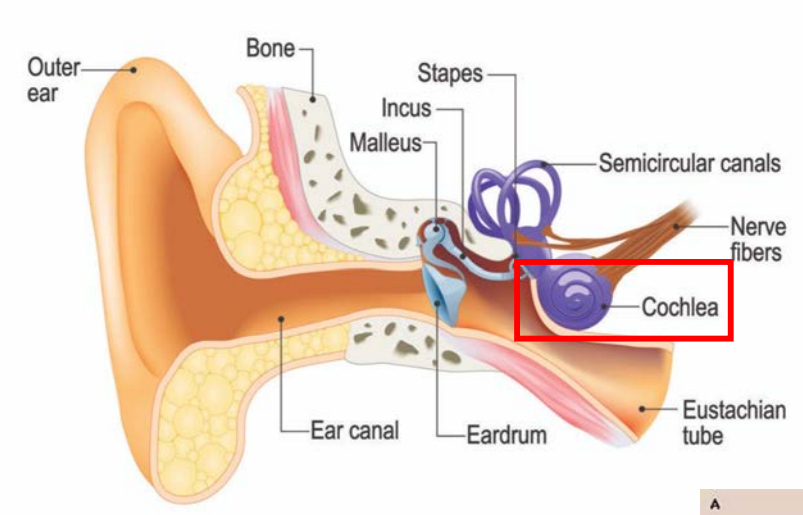
Speech vs Text: An AI Perspective

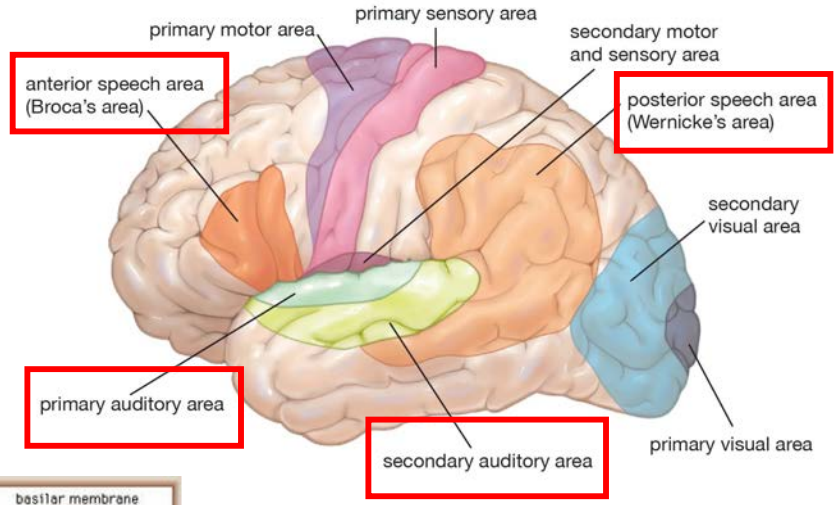


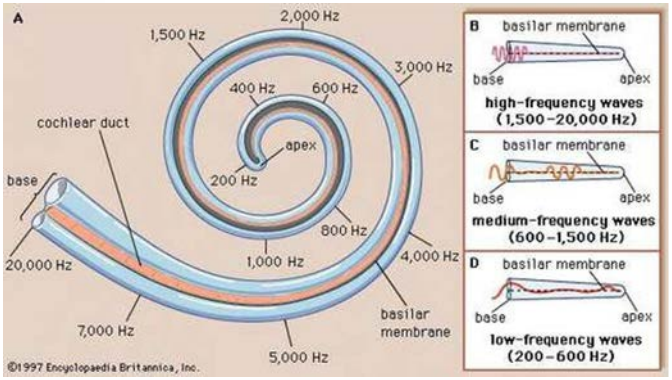
Speech Representations: Spectrograms



Speech Recognition: The Human Solution







Automatic Speech Recognition(ASR)

Automatic Speech Recognition

- Speech waveform in, text (words for speech content) out
- Evaluation: Word Error Rate
- Metrics: Word Error Rate
- History
  - 1975-1990: Statistical method, HMMs, n-gram language models (IBM)
  - 1985-2000: GMM-HMMs (Bell Lab, CMUSphinx)
  - 1990-2020: Large vocabulary continuous speech recognition (Cambridge, HTK)
  - 2010-2020: RNN LM, DNN-HMMs, CTC (Deep learning, Kaldi)
  - 2014-now: RNN-T, AED (Google, ESPnet, SpeechBrain, FairSeq, TorchAudio…)
  - Large language models & Future?

Statistical ASR Method

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} P(\mathbf{W}|\mathbf{O})$$

$$= \arg\max_{\mathbf{W}} P(\mathbf{O}|\mathbf{W})P(\mathbf{W})/p(\mathbf{O})$$

$$= \arg\max_{\mathbf{W}} P(\mathbf{O}|\mathbf{W})P(\mathbf{W})$$

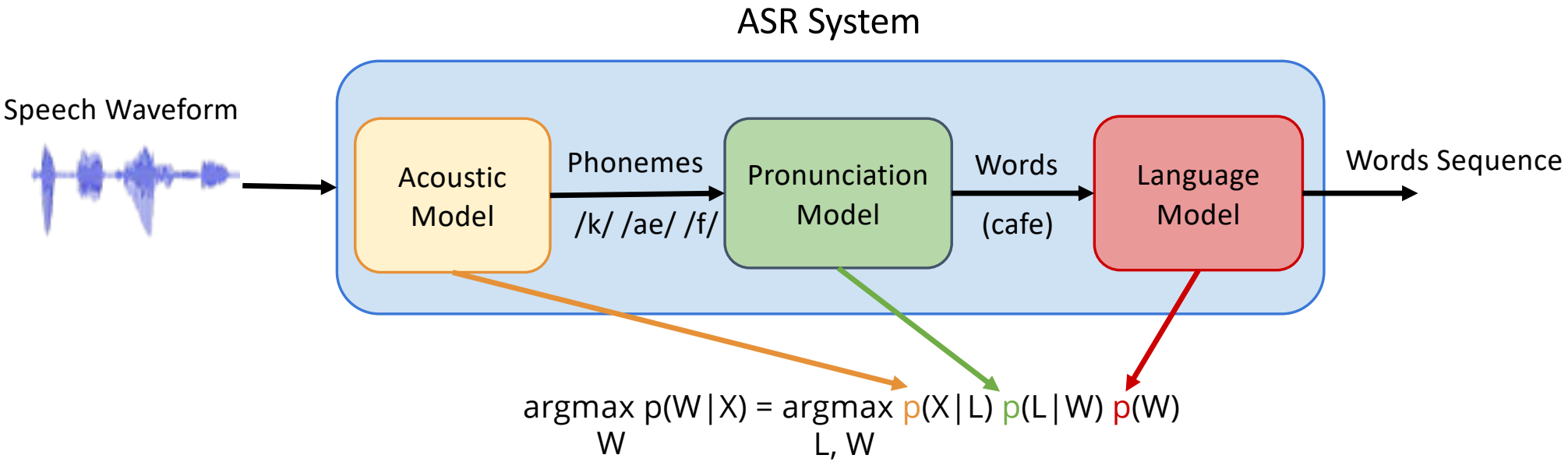
Discriminative model

Generative model

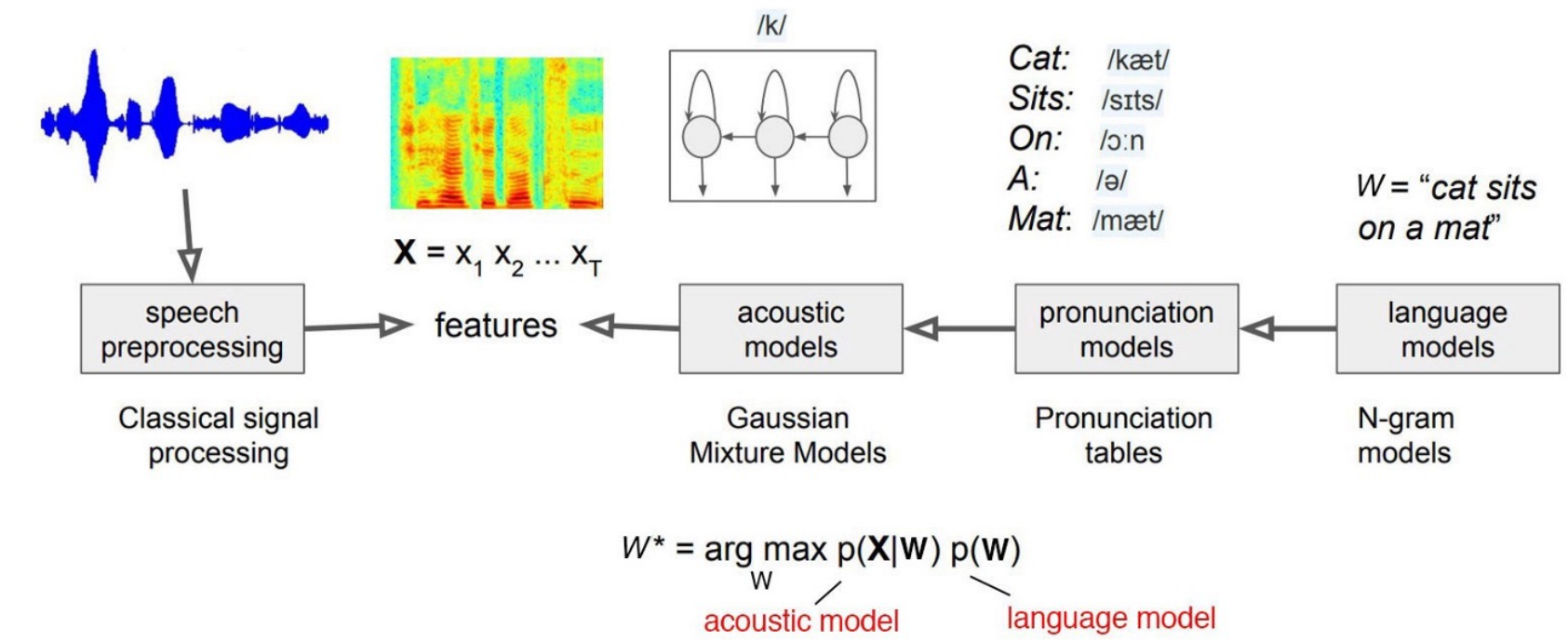
Acoustic model

Language model

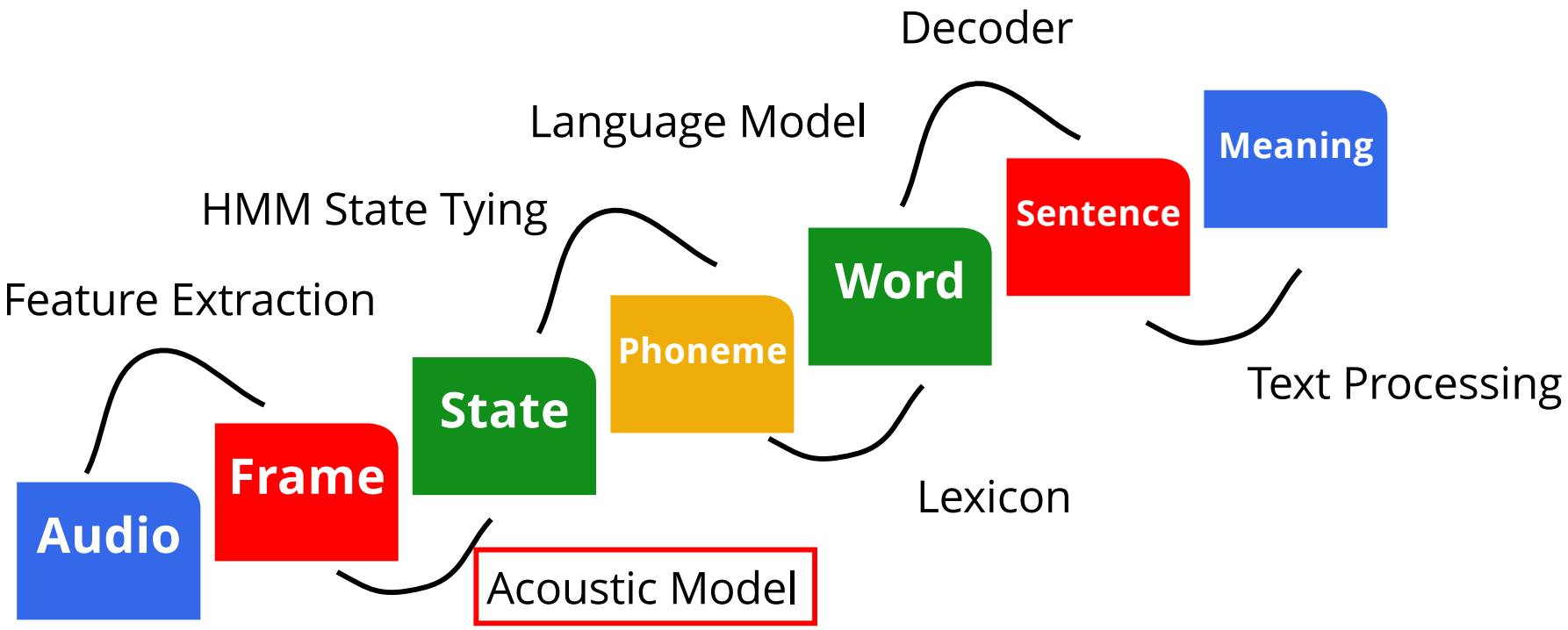
Statistical ASR Method



Modularised System



Bottom-Up Probabilistic Transduction

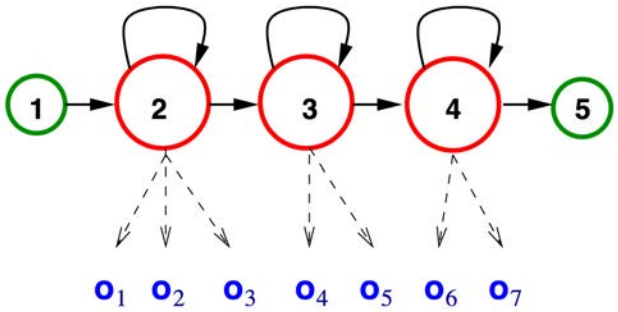


Hidden Markov Model (HMM)

- Align between the input/output sequences (unequal lengths)
- Hidden: Unknown state sequence (alignment)
  - Sum over all possible state sequences (forward-backward / alpha-beta procedure)
  - Calculate the most likely state sequence (Viterbi algorithm)

$$p(\mathbf{O}, \mathbf{Q}|\lambda) = a_{12} b_2(\mathbf{o}_1) a_{22} b_2(\mathbf{o}_2) a_{22} b_2(\mathbf{o}_3) a_{23} \ldots b_4(\mathbf{o}_7) a_{45}$$

$$p(\mathbf{O}|\lambda) = \sum_{\mathbf{Q}} p(\mathbf{O}, \mathbf{Q}|\lambda)$$



Forward Procedure

$$\alpha_j(t) = p(\mathbf{o}_{1:t}, q_t = j|\lambda)$$

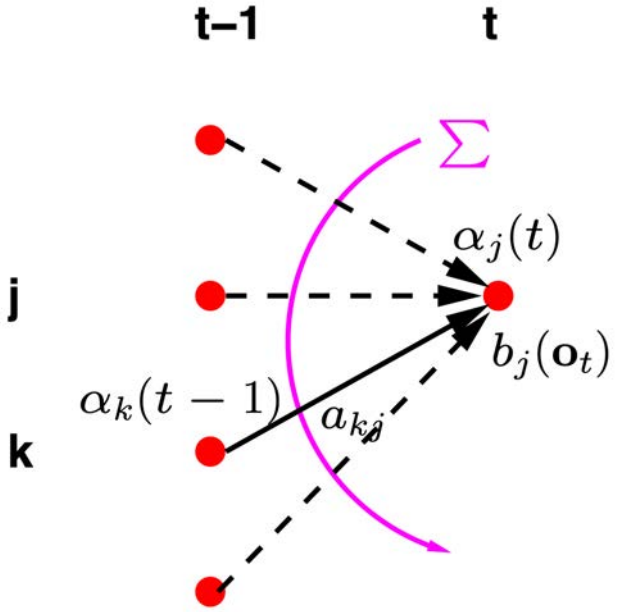
$$p(\mathbf{o}_{1:t}|\lambda) = \sum_{j=1}^{N} \alpha_j(t)$$

$$p(\mathbf{o}_{1:t}, q_{t-1} = k, q_t = j|\lambda) = \alpha_k(t-1)a_{kj}b_j(\mathbf{o}_t)$$

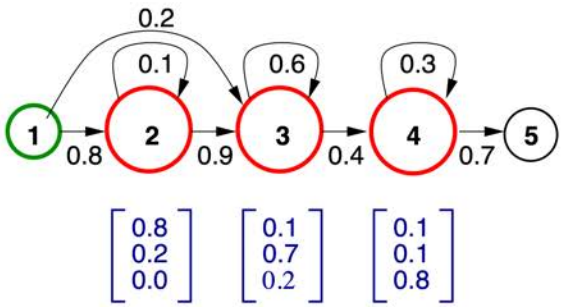
$$\alpha_j(t) = \sum_{k=1}^{N} p(\mathbf{o}_{1:t}, q_{t-1} = k, q_t = j|\lambda)$$

$$p(\mathbf{O}|\lambda) = \sum_{k=2}^{N=1} \alpha_k(T)a_{kN}$$

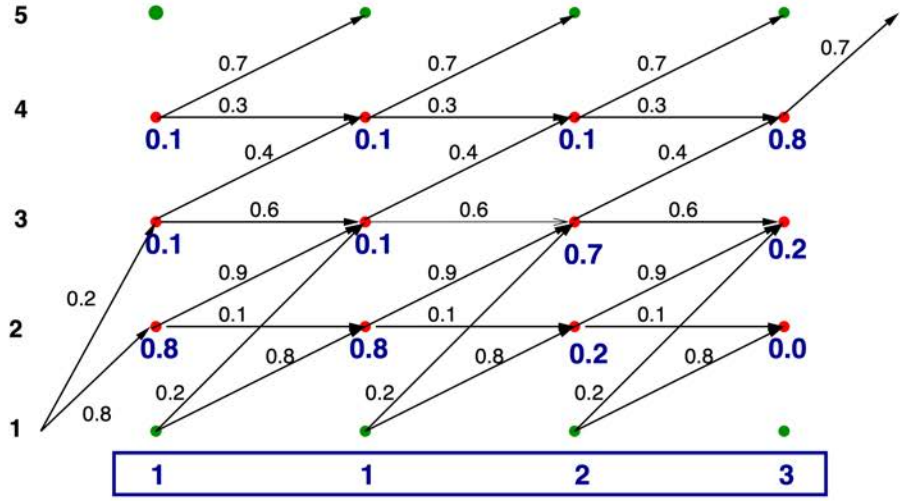
$$\alpha_j(t) = b_j(\mathbf{o}_t)\left[\sum_{k=1}^{N-1} \alpha_k(t-1)a_{kj}\right]$$



Forward Procedure



Given the HMM with discrete output distributions and the observed sequence $\mathbf{O} = [1, 1, 2, 3]$:



5	-	-	-	-	-	0.013156
4	0.0	0.0	0.0008	0.002376	0.018795	-
3	0.0	0.02	0.0588	0.056952	0.0070186	-
2	0.0	0.64	0.0512	0.001024	0.0	-
1	1.0	0.0	0.0	0.0	0.0	-
	-	1	1	2	3	-

Backward Procedure

$$\beta_j(t) = \sum_{k=2}^{N-1} a_{jk} b_k(\mathbf{o}_{t+1}) \beta_k(t+1)$$

$$p(\mathbf{O}|\lambda) = \sum_{k=2}^{N-1} a_{1k} b_k(\mathbf{o}_1) \beta_k(1)$$

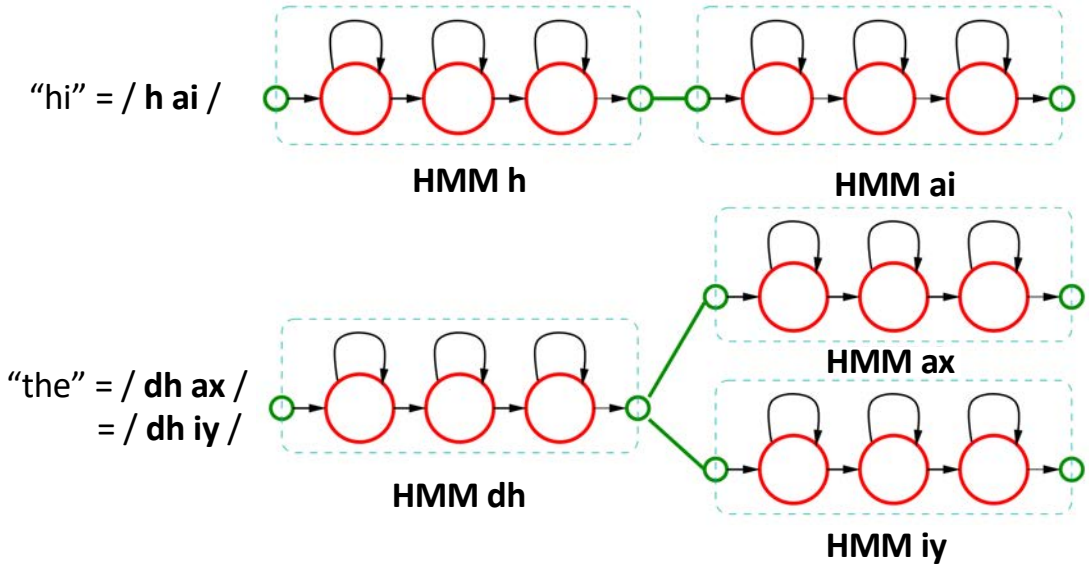
$$\alpha_i(t)\beta_i(t) = p(\mathbf{O}, q_t = i|\lambda)$$

$$p(\mathbf{O}|\lambda) = \sum_{i=1}^{N} \alpha_i(t)\beta_i(t)$$

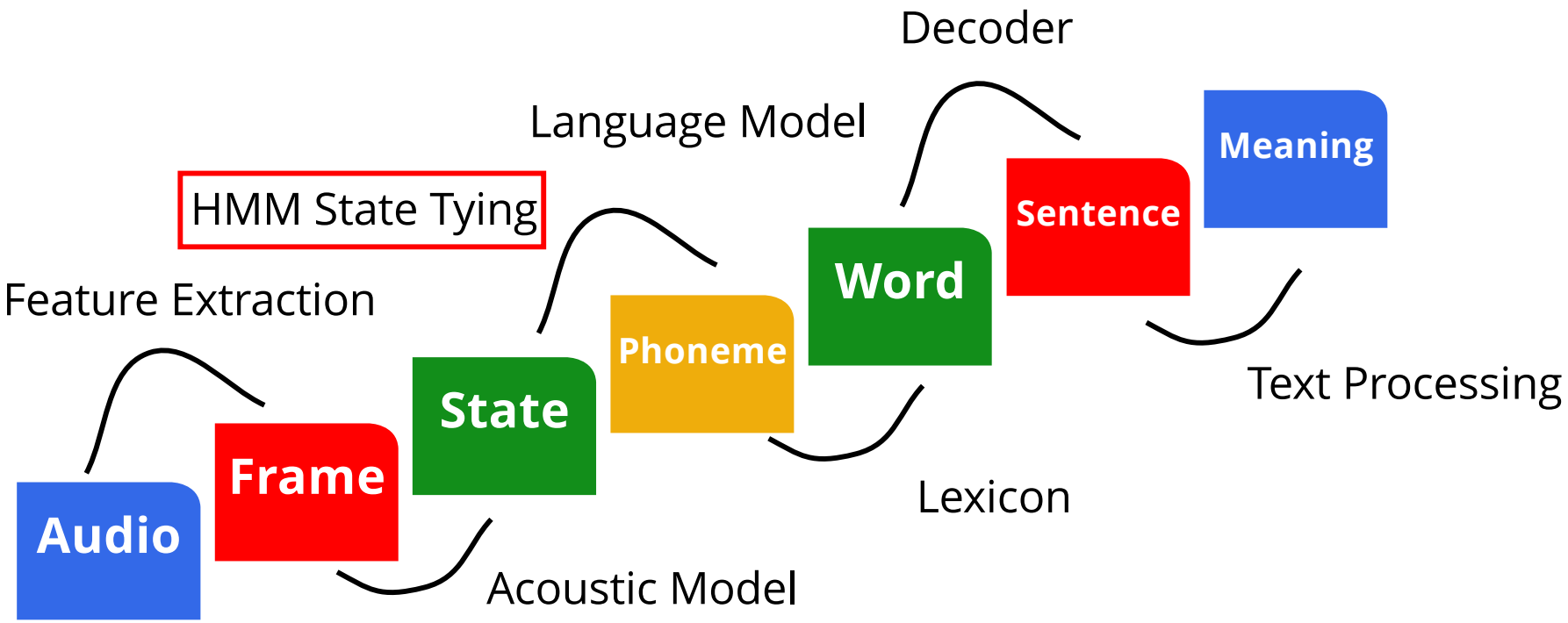
State occupancy (soft alignment)

$$\gamma_i(t) = P(q_t = i|\mathbf{O}, \lambda) = \frac{p(\mathbf{O}, q_t = i|\lambda)}{p(\mathbf{O}|\lambda)}$$

- Can build larger HMMs (composite HMMs) using smaller HMMs.
- Allow to build word/sentence models from subword models (for open vocabulary).



Bottom-Up Probabilistic Transduction



HMM State Tying

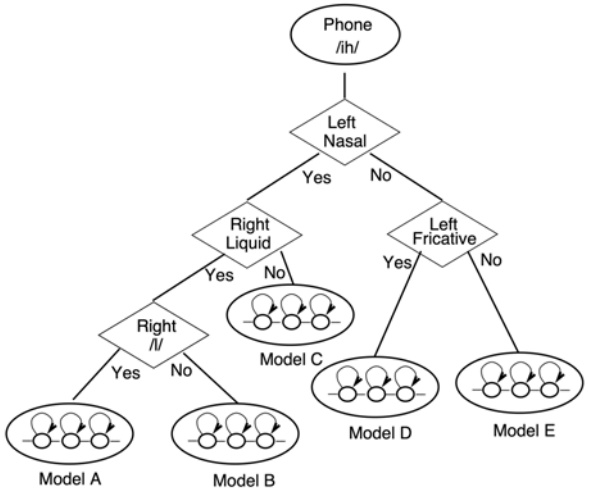
- Context-dependent phones are used for HMMs due to co-articulation.
  - "speech"= **Monophone** / s p iy ch /
    Biphones (L) / sil-s s-p p-iy iy-ch /
    Biphones (R) / s+p p+iy iy+ch ch+sil /
    Triphones / sil-s+p s-p+iy p-iy+ch iy-ch+sil /
  - "speech task"=

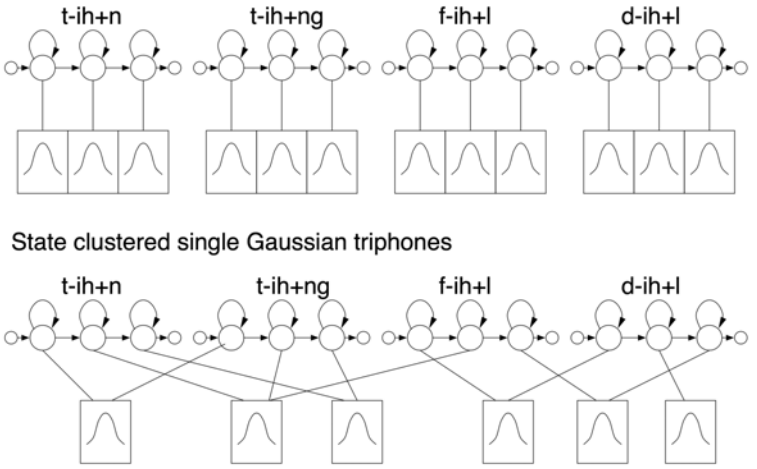
Word-internal

/ sil s+p s-p+iy p-iy+ch iy-ch
t+ae t-ae+s ae-s+k s-k sil /

Cross-word

/ sil-s+p s-p+iy p-iy+ch iy-ch+t
ch-t+ae t-ae+s ae-s+k s-k+sil /





Other (Subword) Models

- Graphemes
  - Building acoustic models using characters & context-dependent characters
- UTF-8 Bytes:
  - Break each multilingual character into 1-4 bytes
- Word-piece model (WPM) & Byte-pair encoding (BPE)
  - Segmenting words into most common pieces
  - aaabdaaabac
  - ZabdZabac, Z=aa
  - ZYdZYac, Y=ab, Z=aa
  - XdXac, X=ZY Y=ab, Z=aa

hello world = he llo_ wor ld_

- Sentence-piece model (SPM)

  Segmenting sentences into cross-word pieces

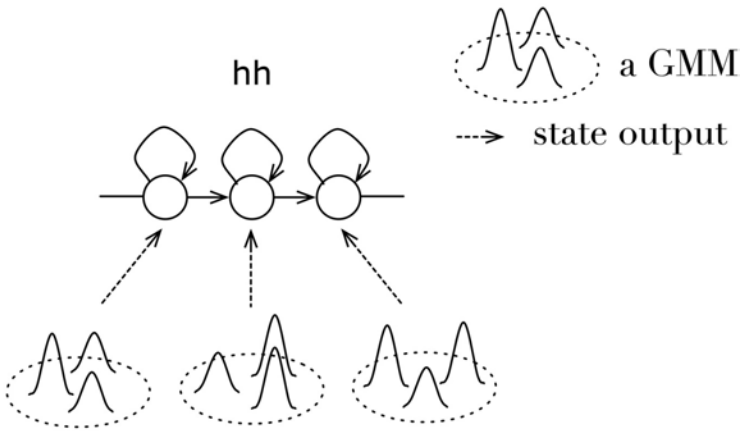
Observation Density Functions

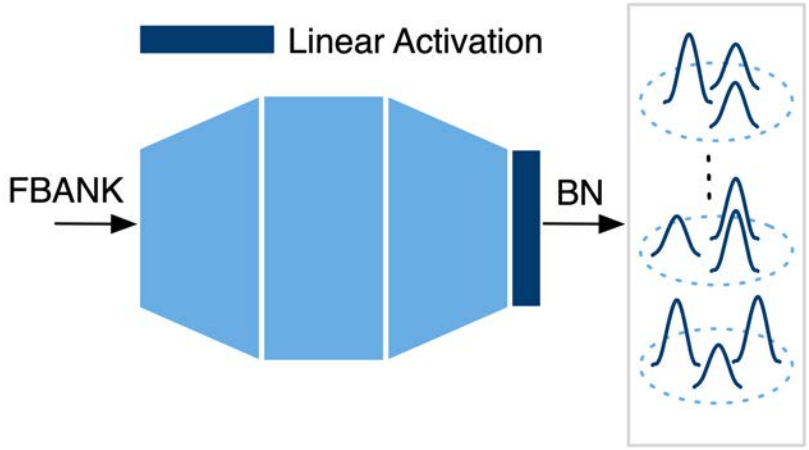
- GMM-HMMs

B.-H. Juang, Maximum-likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains, 1985.

H. Hermansky, D. Ellis, S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems", 2000.

  - GMMs can form smooth approximations to arbitrarily shaped densities.
  - A GMM is built for each HMM state corresponding to a subword unit.
  - Tandem system: GMM-HMMs with DNN extracted features.





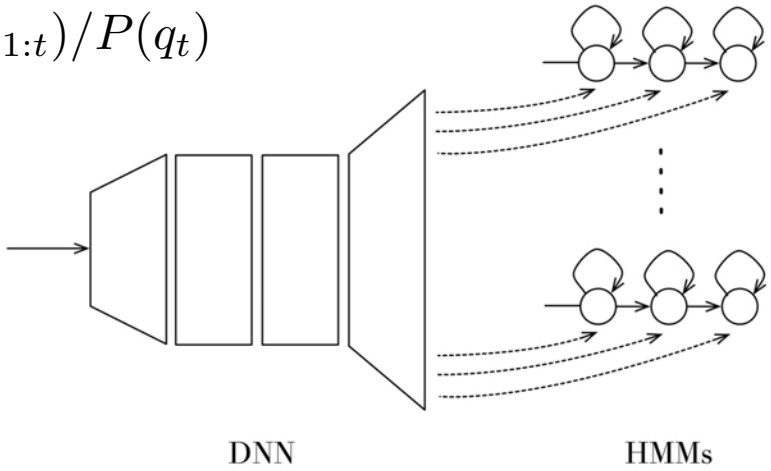
Observation Density Functions

- ANN-HMMs (Hybrid system)

H. Bourlard, N. Morgan, Connectionist Speech Recognition: A Hybrid Approach, 1993.

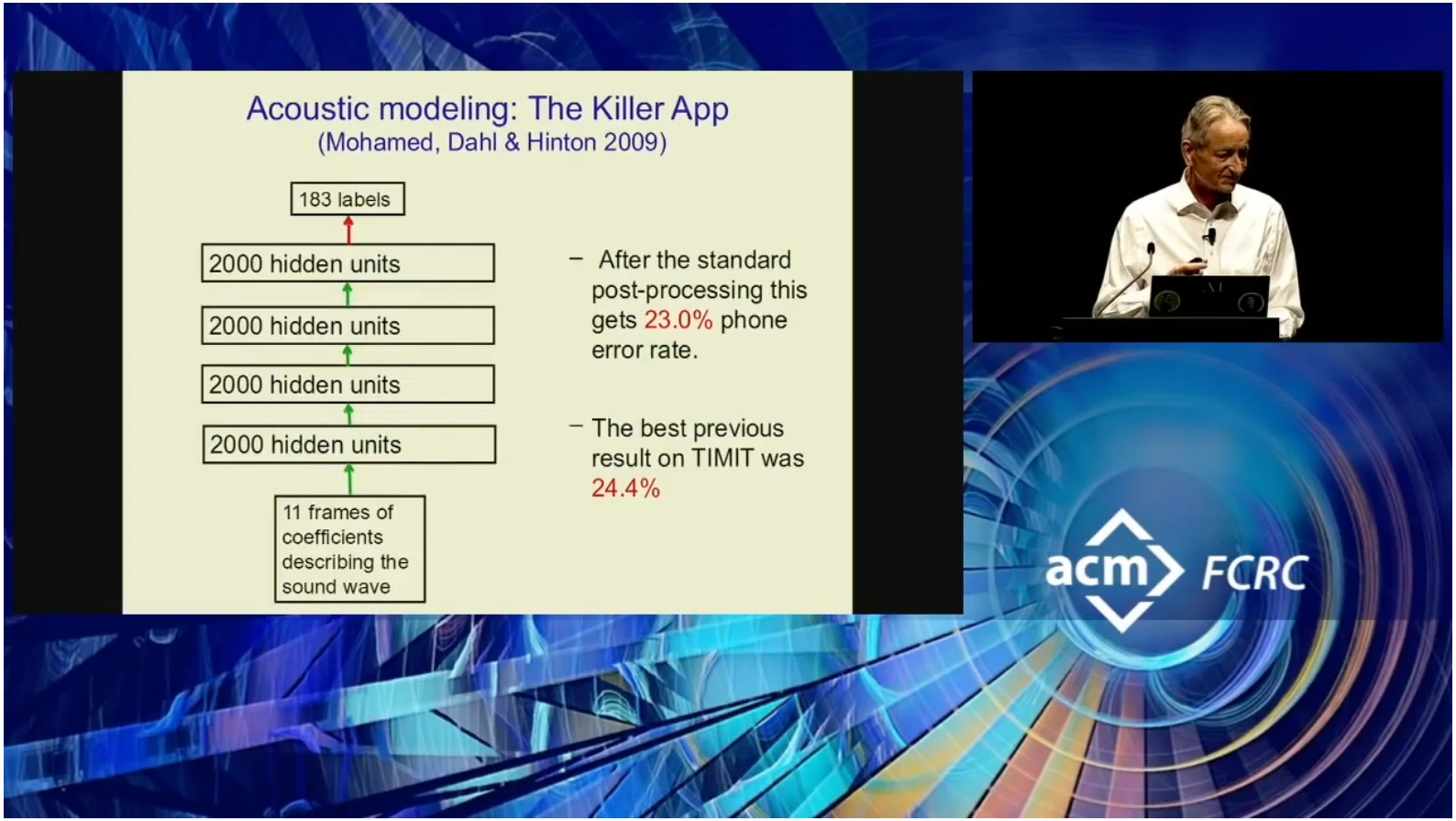
  - ANNs are universal function approximators w/o any constraint on input features.
  - A three-layer multi-layer perceptron can represent any continuous mapping function arbitrarily accurately.
  - ANN-HMMs often outperform GMM-HMMs due wider input windows.

$$p(\mathbf{o}_{1:t}|q_t) \propto P(q_t|\mathbf{o}_{1:t})/P(q_t)$$



DNN-HMMs

https://www.youtube.com/watch?v=-cv0ddgcImk



DNN-HMMs

Deep Belief Networks for phone r

Abdel-rahman Mohamed, George Dahl, and Geof
Department of Computer Science
University of Toronto
{asamir,gdahl,hinton}@cs.toronto

Abstract

Hidden Markov Models (HMMs) have been the state-of-th
acoustic modeling despite their unrealistic independence assu
limited representational capacity of their hidden states. There
in the research community for deeper models that are capa
many types of variability present in the speech generation p
Networks (DBNs) have recently proved to be very effective
chine learning problems and this paper applies DBNs to acc
the standard TIMIT corpus, DBNs consistently outperform c
the best DBN achieves a phone error rate (PER) of 23.0% on
set.

INTERSPEECH 2011

**Conversational Speech Transcript
Using Context-Dependent Deep Neural**

Frank Seide[1], *Gang Li,*[1] *and Dong Yu*[2]

[1]Microsoft Research Asia, Beijing, P.R.C
[2]Microsoft Research, Redmond, USA
{fseide,ganl,dongyu}@microsoft.com

Abstract

We apply the recently proposed Context-Dependent Deep-Neural-Network HMMs, or CD-DNN-HMMs, to speech-to-text transcription. For single-pass speaker-independent recognition on the RT03S Fisher portion of phone-call transcription benchmark (Switchboard), the word-error rate is reduced from 27.4%, obtained by discriminatively trained Gaussian-mixture HMMs, to 18.5%—a 33% relative improvement.

CD-DNN-HMMs combine classic artificial-neural-network HMMs with traditional tied-state triphones and deep-belief-network pre-training. They had previously been shown to reduce errors by 16% relatively when trained on tens of hours of data using hundreds of tied states. This paper takes CD-DNN-HMMs further and applies them to transcription using over 300 hours of training data, over 9000 tied states, and up to 9 hidden layers, and demonstrates how sparseness can be exploited.

On four less well-matched transcription tasks, we observe relative error reductions of 22–28%.

ers), and task (from
tion). This is demons
the Switchboard phon
and RT03S sets). W
trained on up to 2000
fewer output units (m
Second, we adva
weight sparseness and
strate that this can red
Third, we present
ceptron (MLP) and
understanding which
provements achieved

**2. The
Deep N**

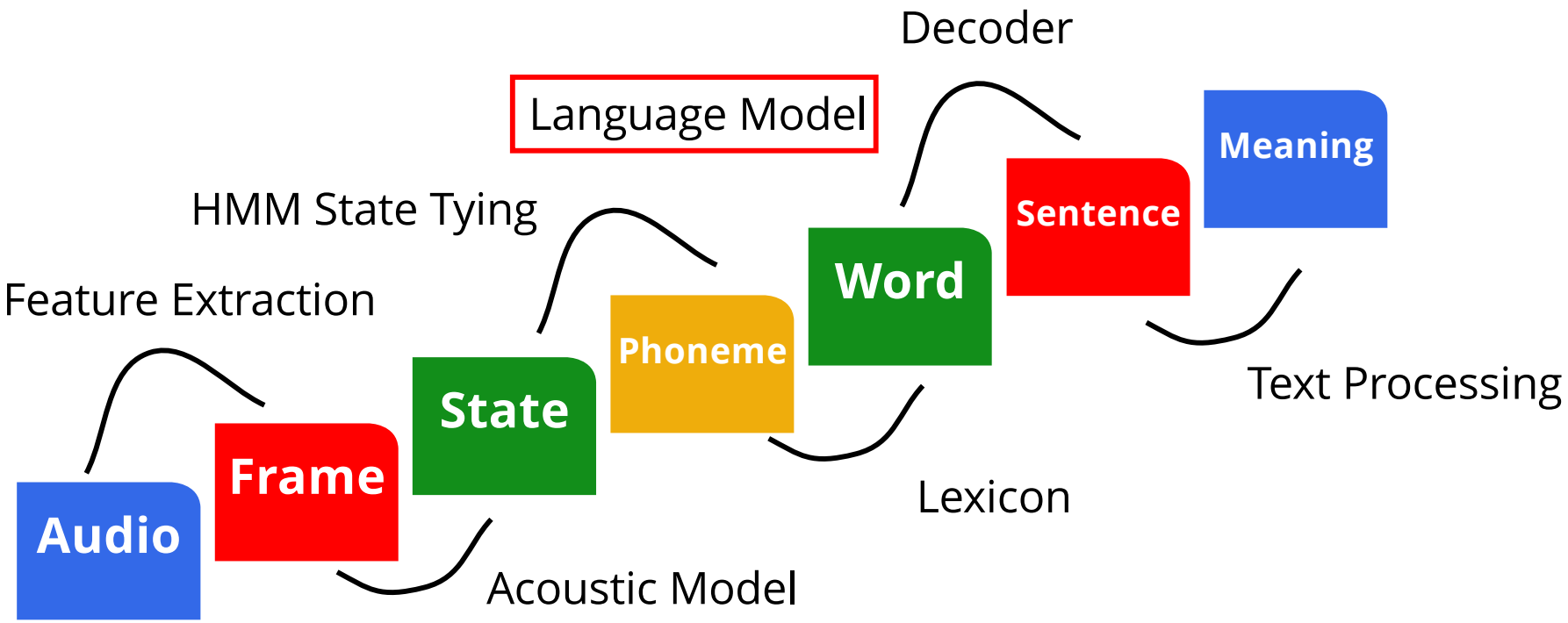
A deep neural network
ceptron (MLP, [8]) w

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury

Deep Neural Networks for Acoustic Modeling in Speech Recognition

The shared views of four research groups

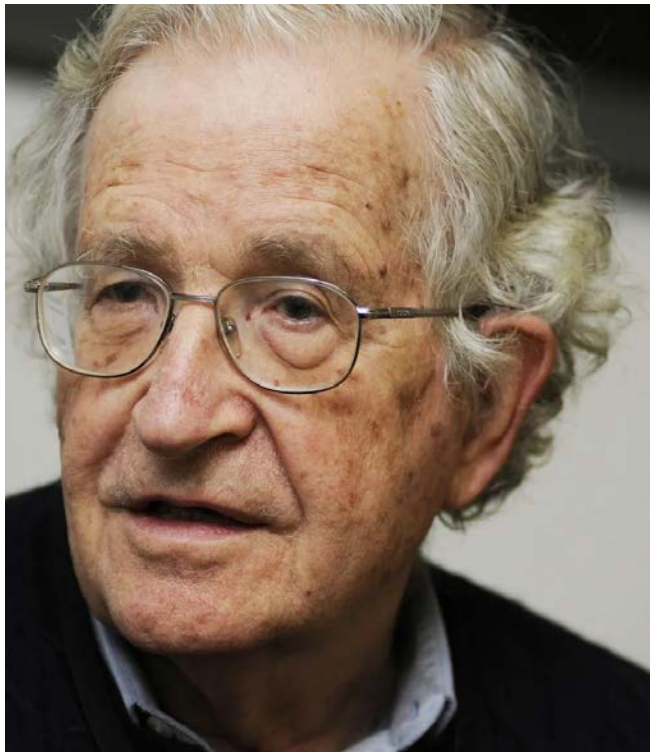
Bottom-Up Probabilistic Transduction



Jelinek & Chomsky

“Every time I fire a linguist, the performance of the speech recognizer goes up.” (1998)

“Machine learning will degrade our science and debase our ethics by incorporating into our technology a fundamentally flawed conception of language and knowledge.” (2023)



Unidirectional Language Model

- Language model (LM) $P(\mathbf{W}) = P(w_{1:L})$
- Unidirectional LM
  - $P(w_{1:L}) = P(w1) \prod_{l=2}^{L} P(w_l | w_{1:l-1})$
  - Neural network LM: Predicting the next word (piece) using previous ones
  - Left-to-right, right-to-left, & any other order that makes the production
  - e.g. n-gram, feedforward LM, RNN/LSTM LM, Transformer LM, GPT…
- Perplexity
  - Can be defined as the inverse probability of the test set

$$\log_2 \mathrm{PPL}(w_{1:L}) = \frac{1}{L} \log_2 P(w_{1:L})$$

Bidirectional Language Model

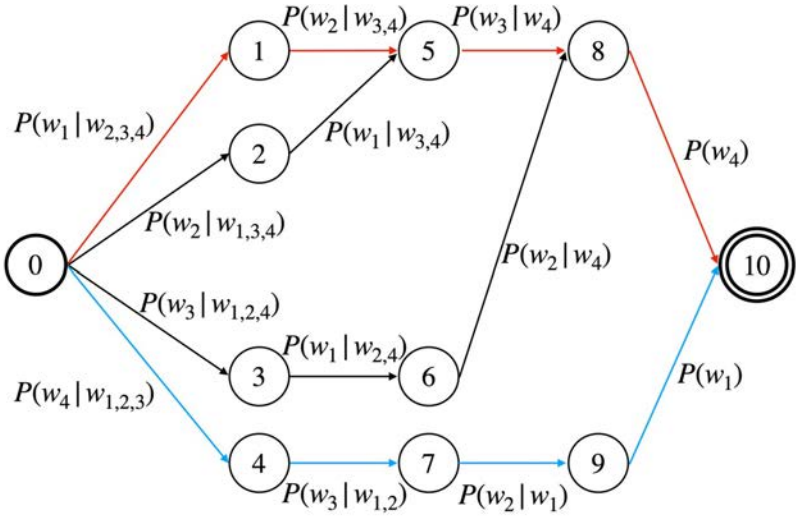
- Bidirectional LM
  - Mask LM: Predicting the current word (piece) using the past & future ones
  - $\phi = P(w_1|w_{2:L})P(w_2|w_1, w_{3:L})...P(w_L|w_{1:L-1})$
  - e.g. ELMo, BERT (“encoder”)
- Can’t compute Perplexity directly (“pseudo”)
  - To compute $P(\mathbf{W})$ (L^2 complexity)

$$P(w_{1:L}) = P(w_1|w_{2:L})P(w_{2:L})$$

$$\vdots$$

$$P(w_{1:L}) = P(w_T|w_{1:L-1})P(w_{1:L-1}).$$

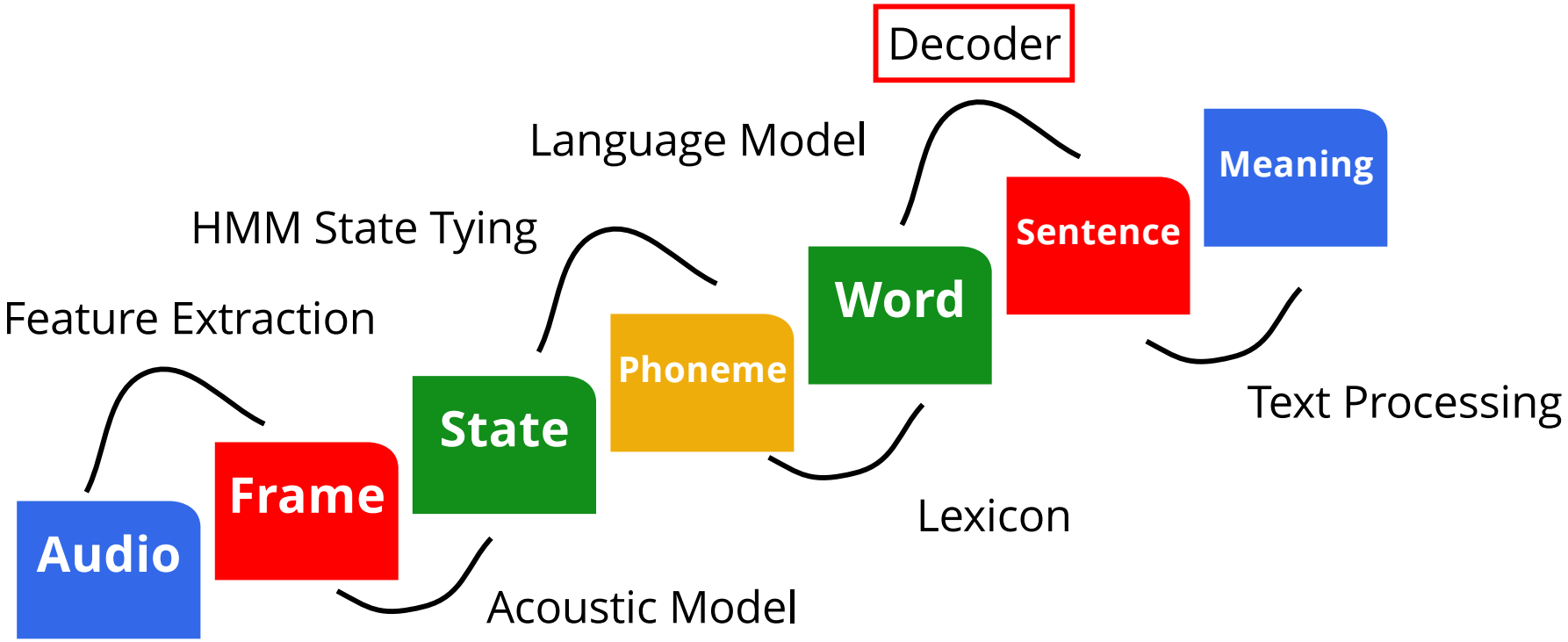
  - $P(w_{1:L}) = [\phi P(w_{2:L})P(w_1, w_{3:L})...P(w_{1:L-1})]^{\frac{1}{L}}$



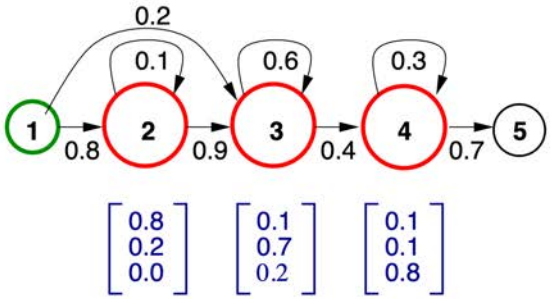
Language Models in Production

- Class-based LM
  - An overly large set of words in practical word-based LM (e.g. over a million).
  - Can use classes to speed up softmax (NN LM), or reduce data sparsity (n-gram)
  - Let $C(w)$ be the mapping between the word and its class,
    $P(w_l|w_{1:l-1}) = P(w_l|C(w_l))P(C(w_l)|C(w_1), \ldots, C(w_{l-1}))$
  - Word clustering can be achieved in many ways (e.g. unigram frequencies).
  - Not so much a problem with subword-level LMs.
- LM with class-based contextual biasing (e.g. contact book)
  - Used in almost all production-level LM (requires class annotation).
    $P(w_l|w_{1:l-1}) = P(C(w_l)|w_{1:l-1})P(w_l|w_{1:l-1}, C(w_l))$

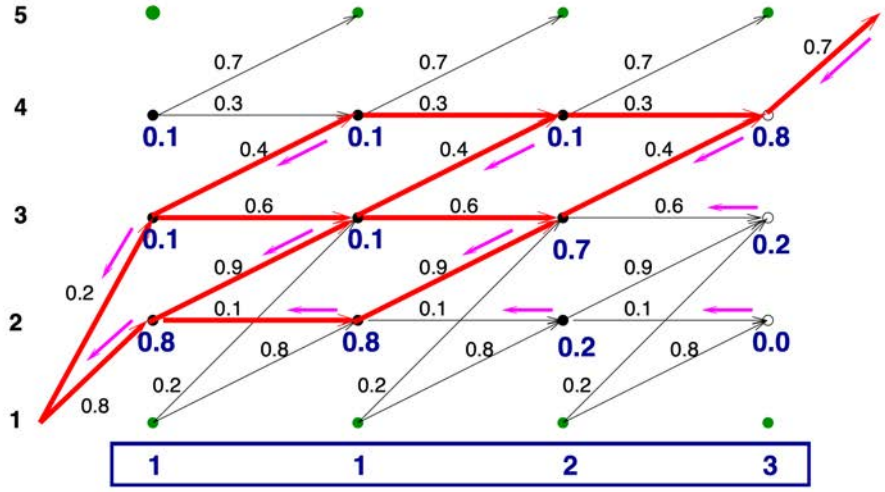
Bottom-Up Probabilistic Transduction



Viterbi Algorithm



Given the HMM with discrete output distributions and the observed sequence $\mathbf{O} = [1, 1, 2, 3]$:



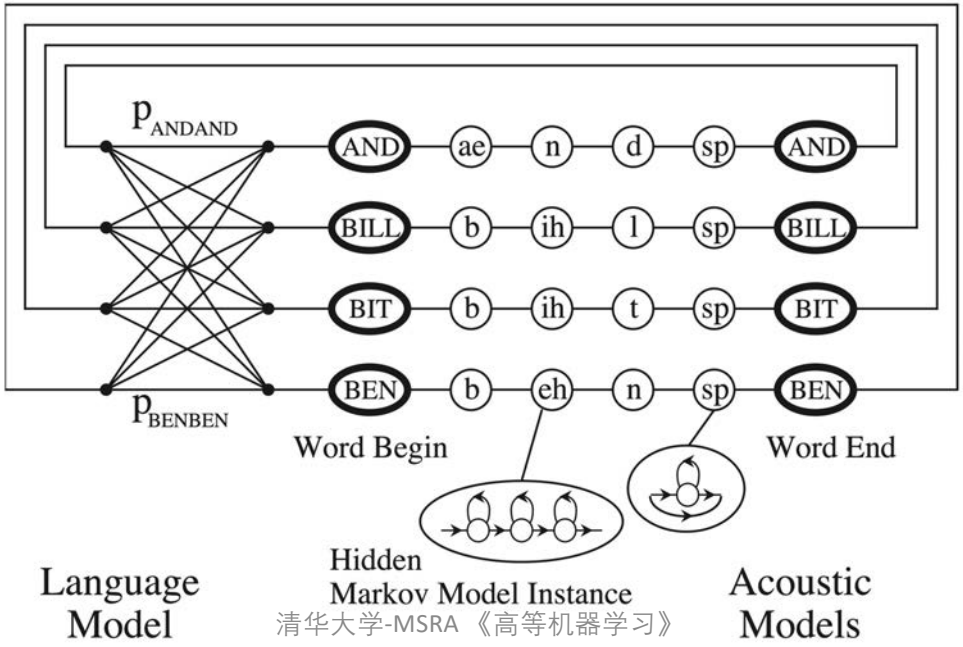
Can be applied to training as well.

Other search algorithms will also work.

5	-	-	-	-	-	0.0072253
4	0.0	0.0	0.00080	0.002304	0.0103220	-
3	0.0	0.02	0.05760	0.032256	0.0038707	-
2	0.0	0.64	0.0512	0.001024	0.0	-
1	1.0	0.0	0.0	0.0	0.0	-
	-	1	1	2	3	-

Decoding Graph

- In decoding, a static/dynamic/hybrid search graph is often built.
  - A monophone HMMs, bigram LM.
  - Weighted finite state transducer (WFST) can be used with a reduced complexity.



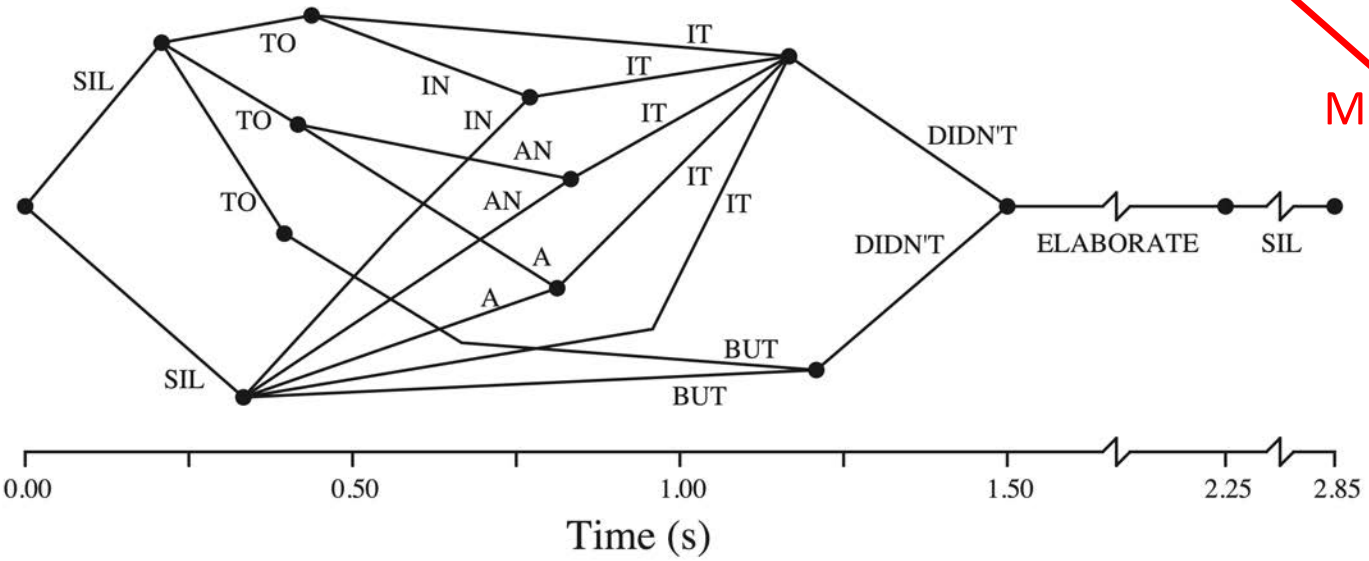
Forced Alignment

- Using Viterbi algorithm to decode based on reference words.
  - A constrained decoding.
  - To find the optimal phone sequence.
  - To find the optimal timing of HMM states (time-to-state alignment).
- P2FA (HTK) and MFA (Montreal)

```
"AMIXXX-00000-0EN2001a-XXXXXX-00_XXXXXXX_0001109_0001553.lab"
0 100000 sil[2] -134.278931 <s>
100000 200000 sil[4] -135.271545
200000 400000 sil-d+ah[2] -256.152130 DOES
400000 500000 sil-d+ah[3] -128.286621
500000 600000 sil-d+ah[4] -127.528603
600000 800000 d-ah+z[2] -254.171280
800000 900000 d-ah+z[3] -128.092651
900000 1100000 d-ah+z[4] -253.537064
1100000 1300000 ah-z+eh[2] -258.717560
1300000 1400000 ah-z+eh[3] -129.799911
1400000 1800000 ah-z+eh[4] -506.590973
1800000 1900000 z-eh+n[2] -126.738007 ANYONE
1900000 2100000 z-eh+n[3] -254.378723
2100000 2200000 z-eh+n[4] -128.702545
```

Lattice

- Lattice is a **reduced static** representation of the dynamic search space.
  - Notes correspond to time, and arcs correspond to word/phone identities.
  - Can be generated by simple AMs & LMs, and rescored by complex ones.



Multi-pass decoding

Discriminative Sequence Loss

Ref: Launch Super Mario
Hyp: To Lunch Mario
[INS] [SUB] [DEL] [CORR]

- Word error rate (WER)
  - WER = (C – I) / **N** x 100% = (**N** – S – D) / **N** x 100%
  - Can be computed by dynamic programming with Levenshtein Distance.
- Minimum Bayesian Risk (sequence-level)
  - Risk: WER (MWE), phone error rate (MPE), HMM state error rate (sMBR)…

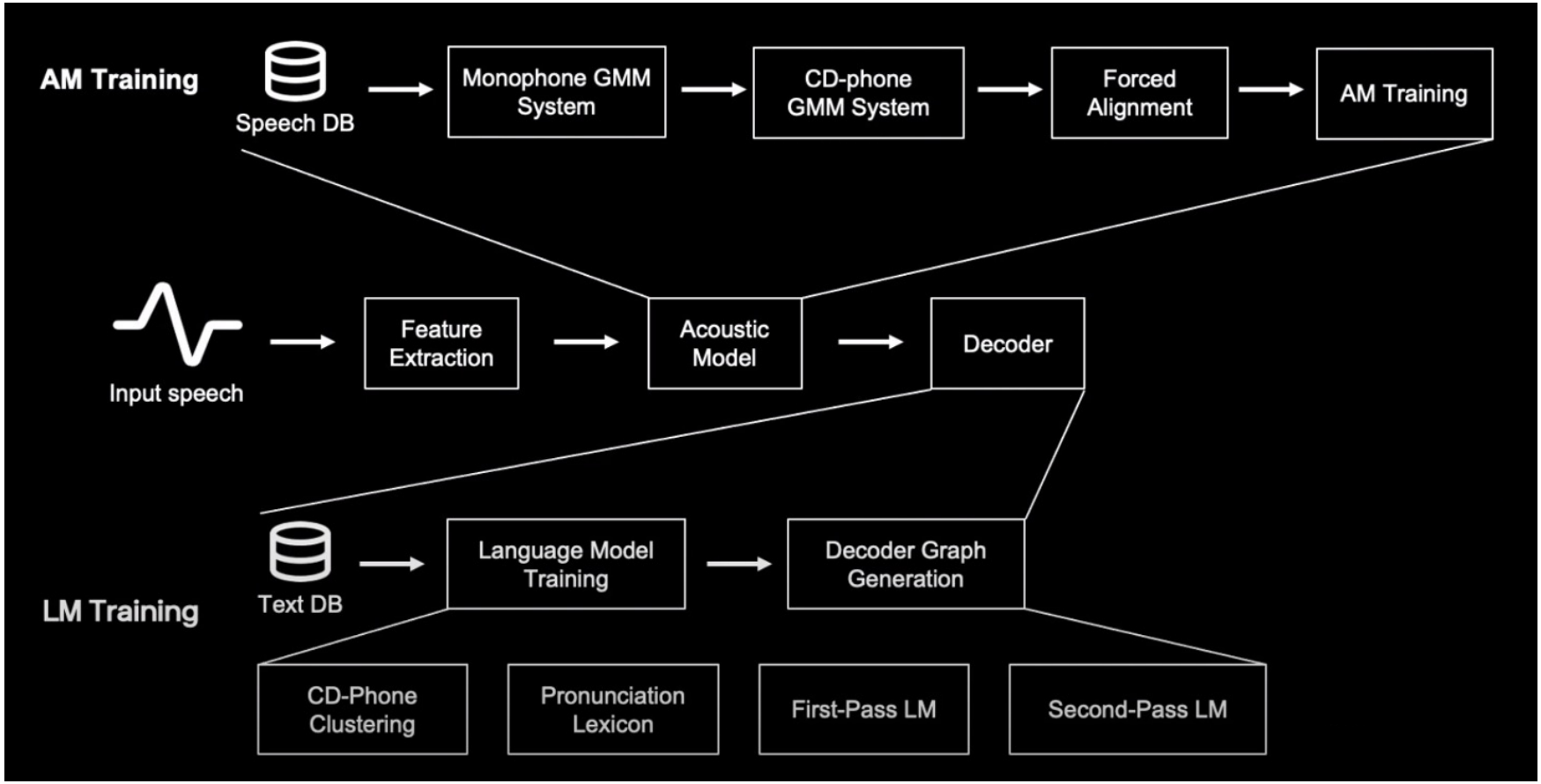
$$\mathbb{E}_{P(\mathbf{W}|\mathbf{O})}[\mathcal{R}(\mathbf{W}^{\text{ref}}, \mathbf{W})] = \frac{\sum_{\mathbf{W}} p(\mathbf{O}|\mathbf{W})P(\mathbf{W})\mathcal{R}(\mathbf{W}^{\text{ref}}, \mathbf{W})}{\sum_{\mathbf{W}} p(\mathbf{O}|\mathbf{W})P(\mathbf{W})}$$

- Maximum mutual information (MMI), lattice-free MMI

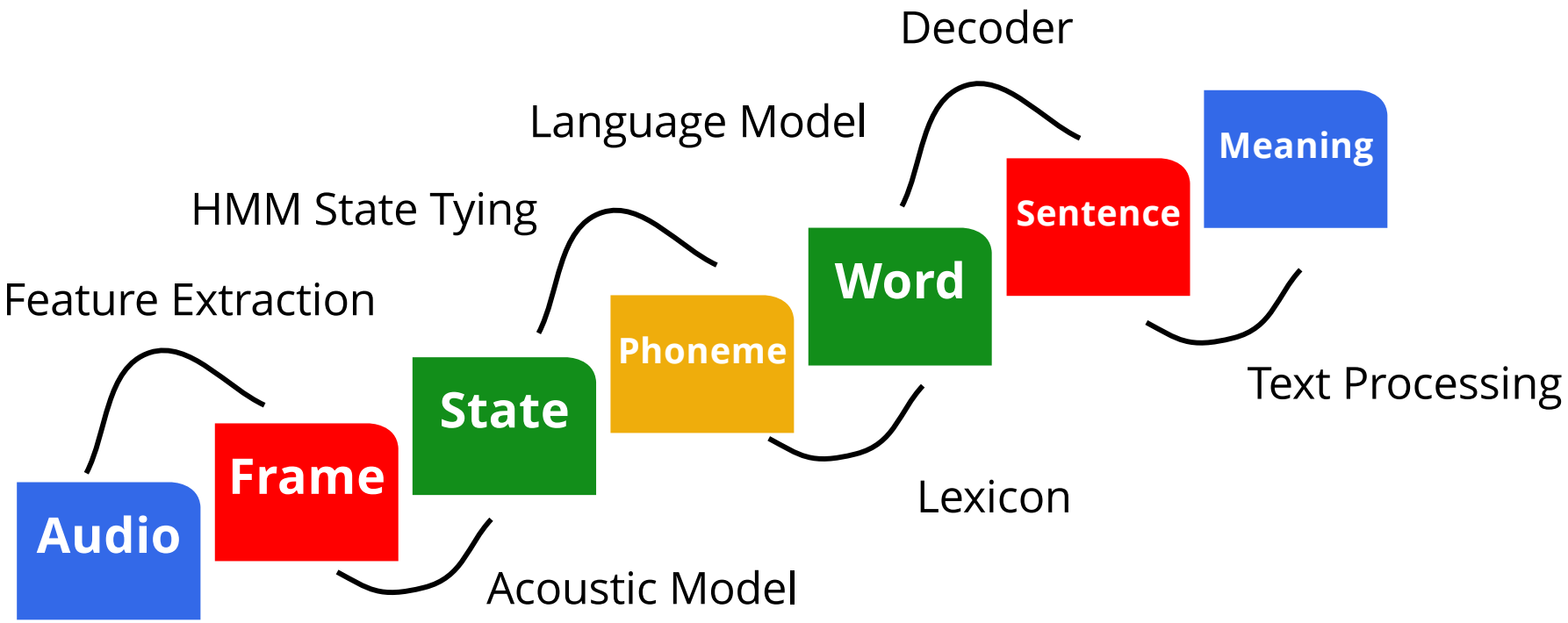
$$P(\mathbf{W}|\mathbf{O}) = \frac{p(\mathbf{O}|\mathbf{W}^{\text{ref}})P(\mathbf{W}^{\text{ref}})}{\sum_{\mathbf{W}} p(\mathbf{O}|\mathbf{W})P(\mathbf{W})}$$

- Lattices are often used to reduce the (repeated) decoding cost.

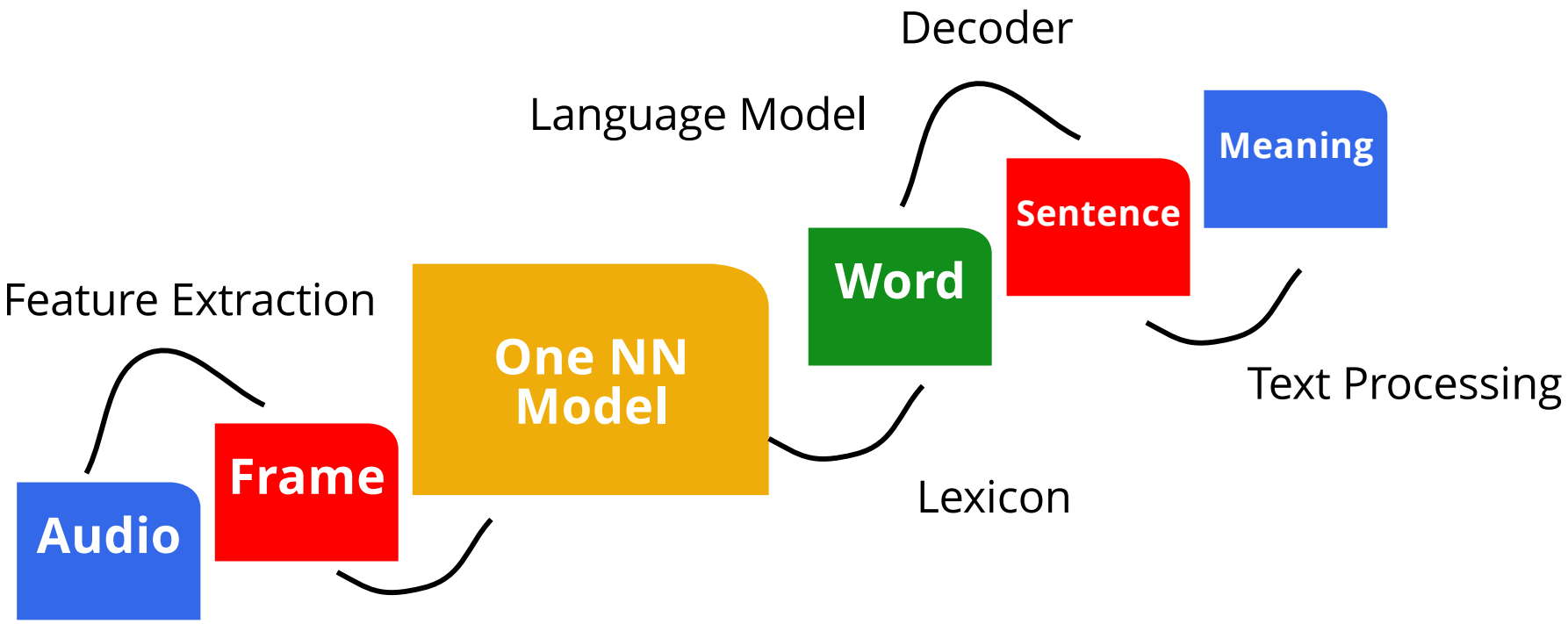
The Complexity of Traditional ASR Pipeline



Bottom-Up Probabilistic Transduction



Bottom-Up Probabilistic Transduction



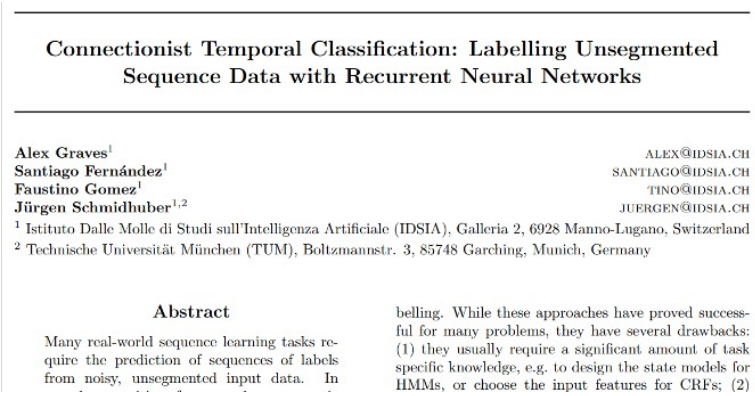
A Brief History of Pure NN ASR

- D. Rumelhart, J. McClelland and the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1986.
- A. Robinson, F. Fallside, “The Utility Driven Dynamic Error Propagation Network”, 1987.
- A. Robinson, Dynamic Error Propagation Networks, 1989.
- A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang, “Phoneme Recognition using Time-delay Neural Networks”, 1989.
- J. Bridle, “Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition”, 1990.
- J. Bridle, “Alpha-nets: A Recurrent Neural Network Architecture with A Hidden Markov Model Interpretation”, 1990.
- Y. Bengio, Artificial Neural Networks and Their Application to Sequence Recognition, 1991.
- A. Senior, Off-line Cursive Handwriting Recognition using Recurrent Neural Networks, 1994.

Connectionist Temporal Classification (CTC)

- CTC is a way to train an acoustic model without requiring frame-level alignments
- Early work used CTC with phoneme output targets.
- CD-phoneme based CTC models achieve state-of-the art performance for conventional, word-level lagged behind ASR.

$$\mathcal{L}_{\mathrm{CTC}} = \ln \sum_{\mathbf{Q}} \prod_{t=1}^{T} P(q_t|\mathbf{O})$$



Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks

Alex Graves[1] ALEX@IDSIA.CH
Santiago Fernández[1] SANTIAGO@IDSIA.CH
Faustino Gomez[1] TINO@IDSIA.CH
Jürgen Schmidhuber[1,2] JUERGEN@IDSIA.CH
[1] Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Galleria 2, 6928 Manno-Lugano, Switzerland
[2] Technische Universität München (TUM), Boltzmannstr. 3, 85748 Garching, Munich, Germany

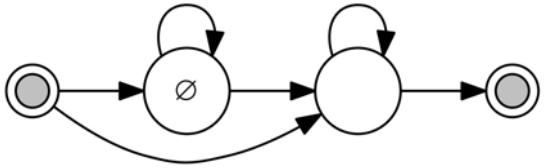
Abstract

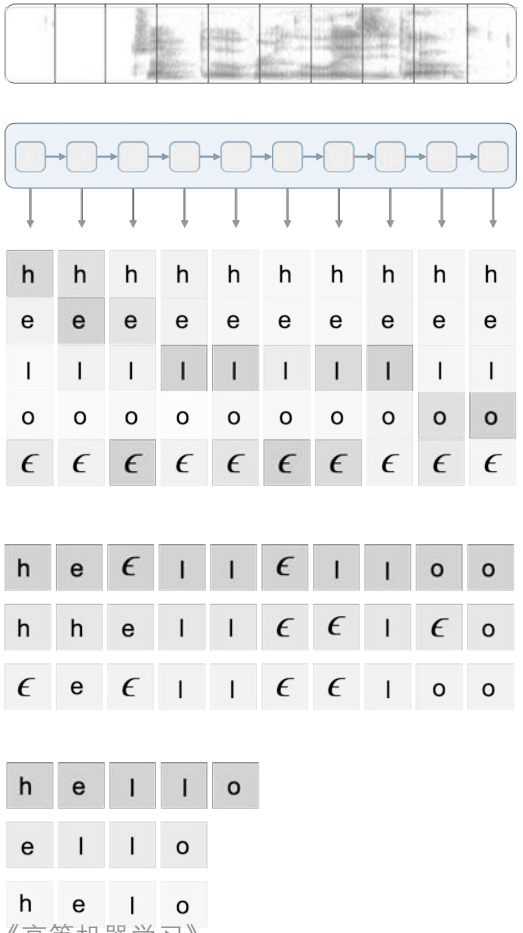
Many real-world sequence learning tasks require the prediction of sequences of labels from noisy, unsegmented input data. In belling. While these approaches have proved successful for many problems, they have several drawbacks: (1) they usually require a significant amount of task specific knowledge, e.g. to design the state models for HMMs, or choose the input features for CRFs; (2)

[Graves et al., 2006]

Connectionist Temporal Classification (CTC)

- CTC is trained by forward-backward procedure and decoded by Viterbi algorithm.
- CTC equals to special case of NN-HMMs with a special blank HMM state.





We start with an input sequence, like a spectrogram of audio.

The input is fed into an RNN, for example.

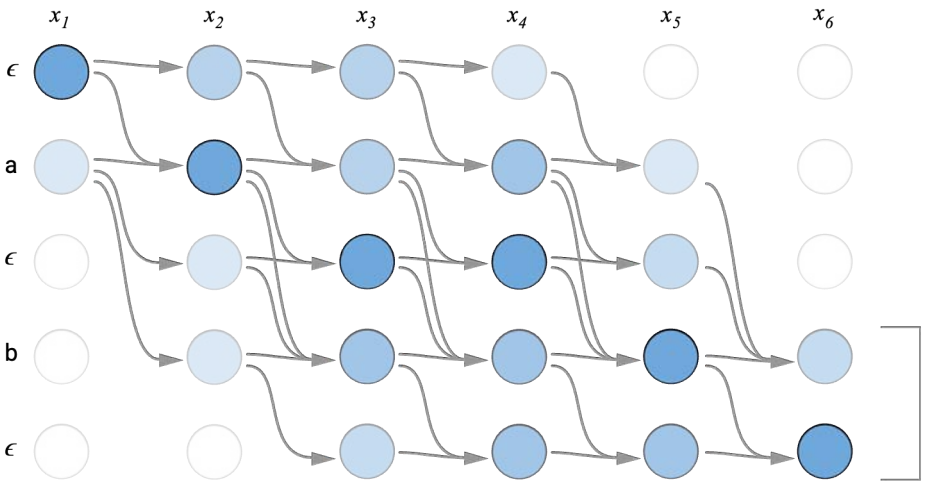
The network gives $p_t(a \mid X)$, a distribution over the outputs {h, e, l, o, ϵ} for each input step.

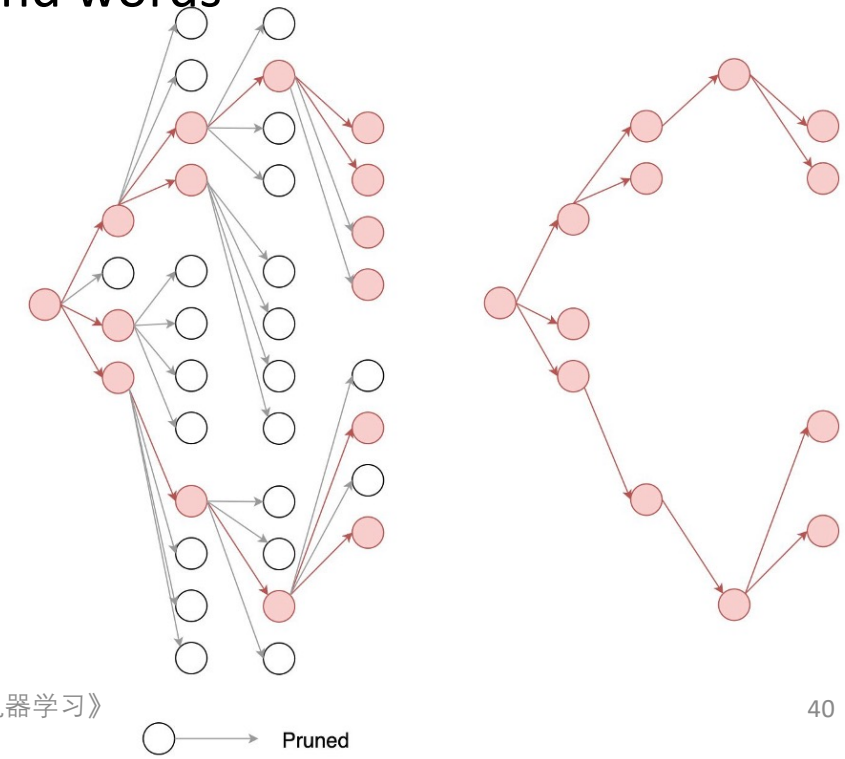
With the per time-step output distribution, we compute the probability of different sequences

By marginalizing over alignments, we get a distribution over outputs.

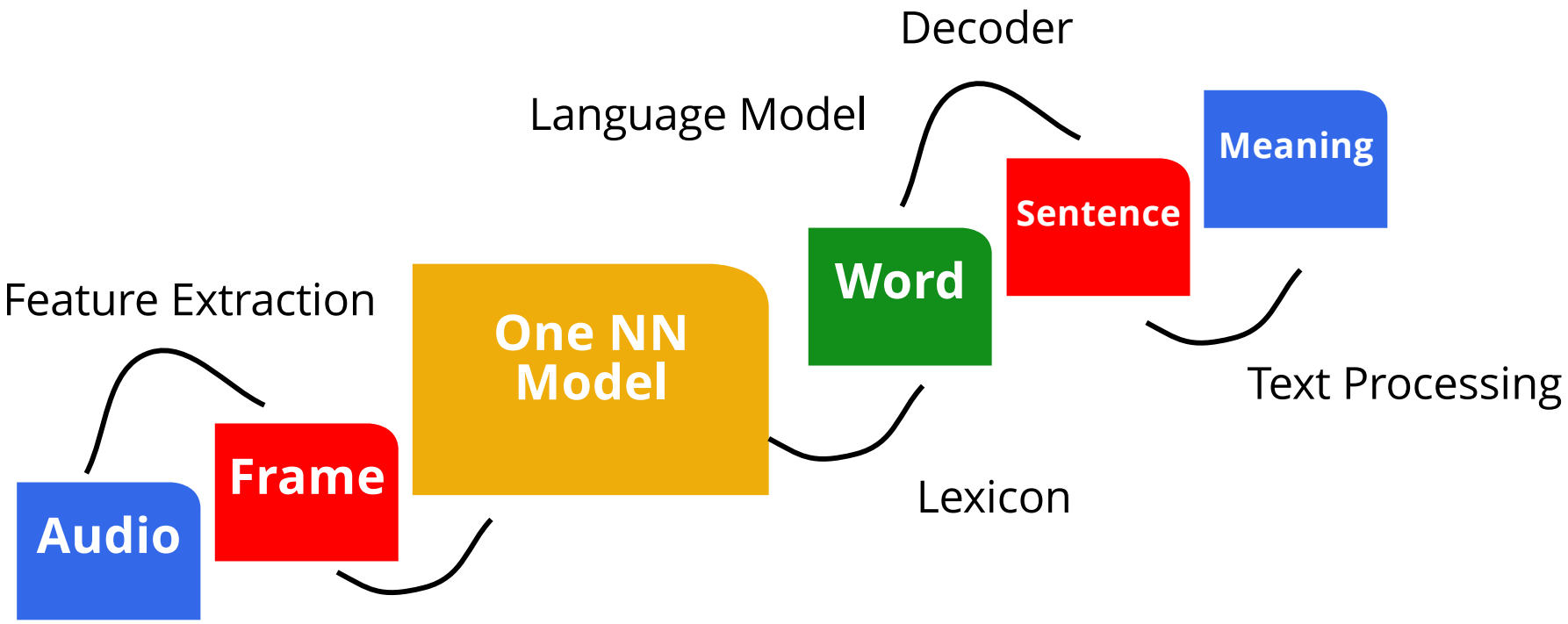
Connectionist Temporal Classification (CTC)

- A good tutorial about CTC can be found at https://distill.pub/2017/ctc/
- CTC can be built on word-piece models and words





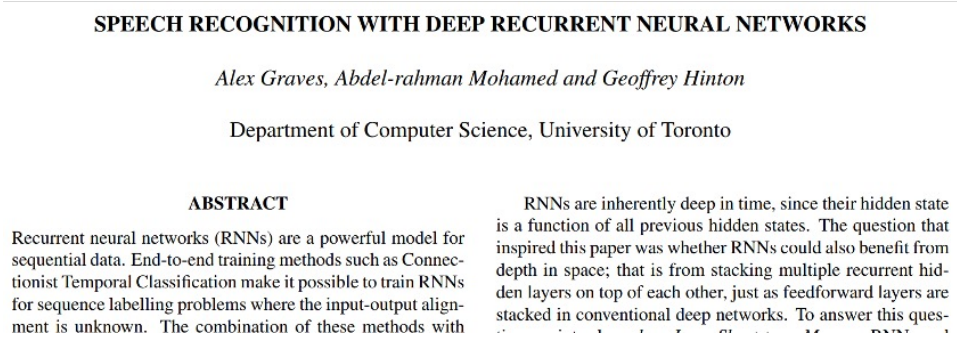
Bottom-Up Probabilistic Transduction



Bottom-Up Probabilistic Transduction

Recurrent Neural Network Transducer (RNN-T)

- Proposed by Graves et al., RNN-T augments a CTC-based model with a recurrent LM component.
- Both components are trained jointly on the available acoustic data
- As with CTC, the method does not require aligned training data.



SPEECH RECOGNITION WITH DEEP RECURRENT NEURAL NETWORKS

Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton

Department of Computer Science, University of Toronto

ABSTRACT

Recurrent neural networks (RNNs) are a powerful model for sequential data. End-to-end training methods such as Connectionist Temporal Classification make it possible to train RNNs for sequence labelling problems where the input-output alignment is unknown. The combination of these methods with

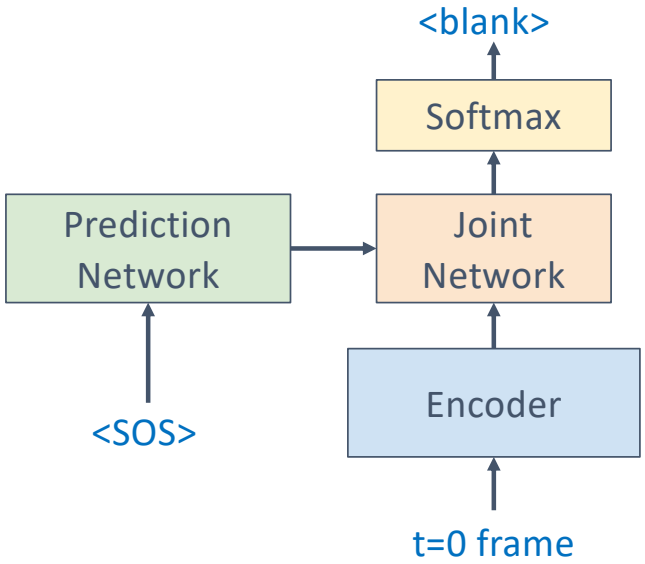
RNNs are inherently deep in time, since their hidden state is a function of all previous hidden states. The question that inspired this paper was whether RNNs could also benefit from depth in space; that is from stacking multiple recurrent hidden layers on top of each other, just as feedforward layers are stacked in conventional deep networks. To answer this ques-

[Graves et al., 2013] ICASSP;
[Graves, 2012] ICML Workshop

Recurrent Neural Network Transducer (RNN-T)

- Softmax over n + 1 labels, includes a blank like CTC
- <blank> → advance in Encoder, retain prediction network state

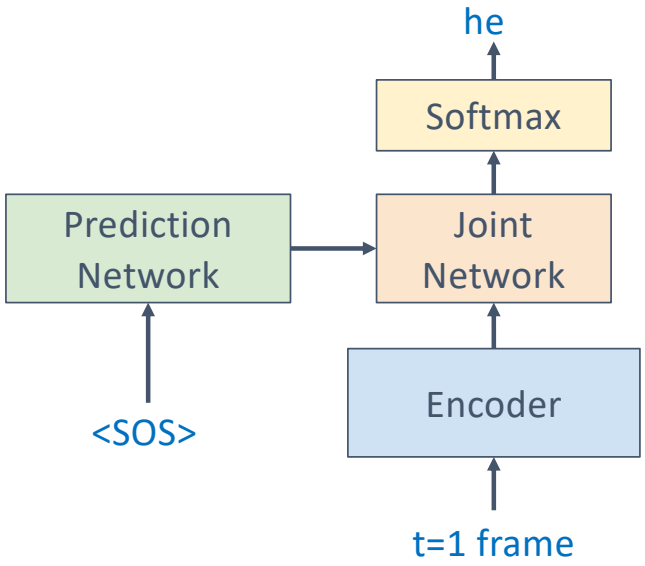
Output:



Recurrent Neural Network Transducer (RNN-T)

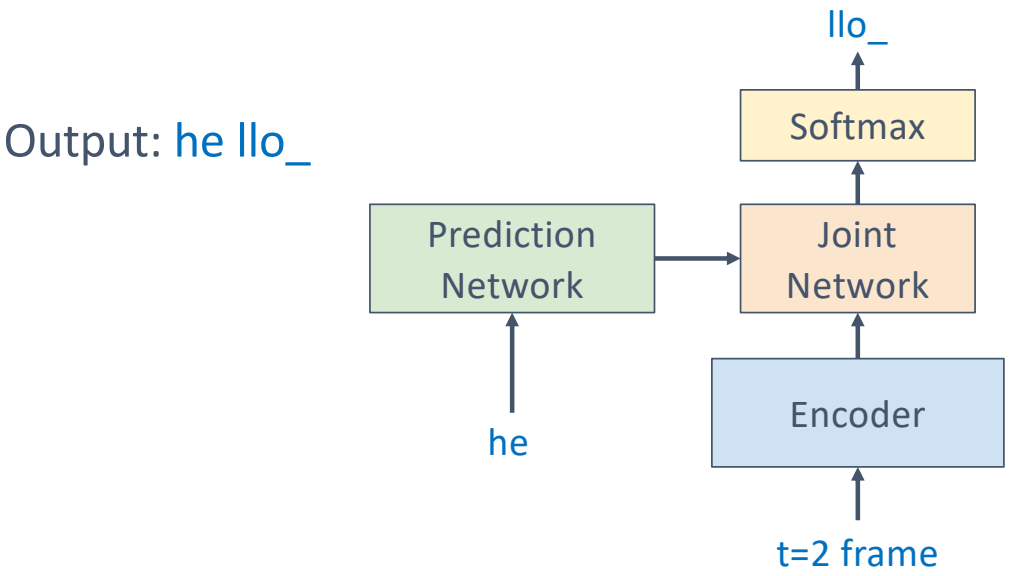
- Softmax over n + 1 labels, includes a blank like CTC
- <blank> → advance in Encoder, retain prediction network state

Output: he



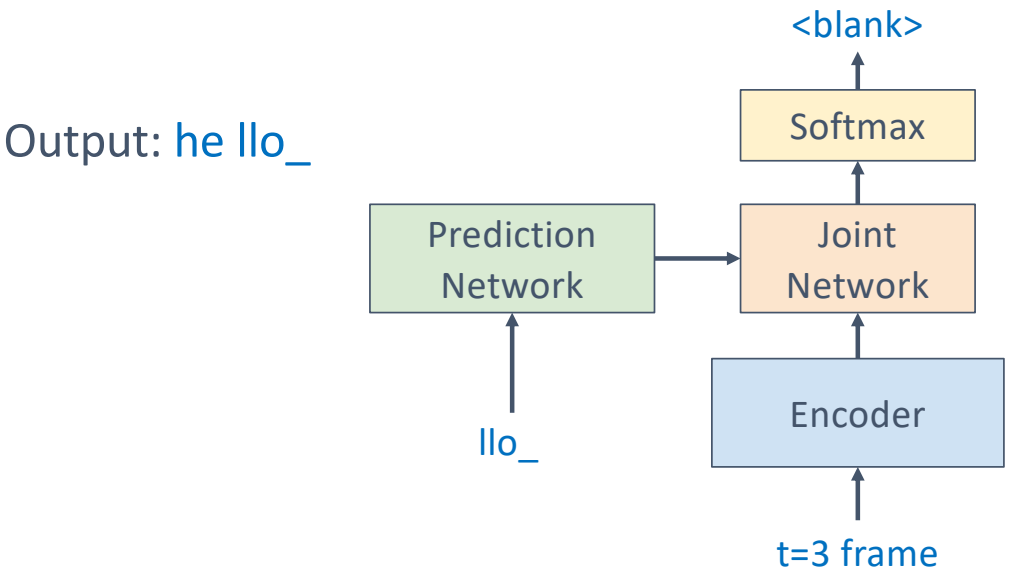
Recurrent Neural Network Transducer (RNN-T)

- Softmax over n + 1 labels, includes a blank like CTC
- <blank> → advance in Encoder, retain prediction network state



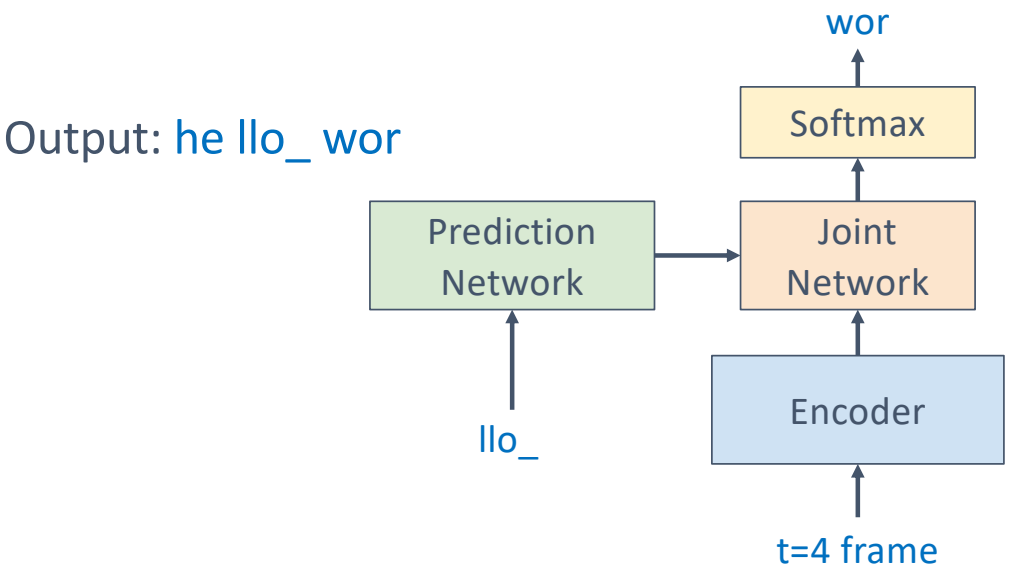
Recurrent Neural Network Transducer (RNN-T)

- Softmax over n + 1 labels, includes a blank like CTC
- <blank> → advance in Encoder, retain prediction network state



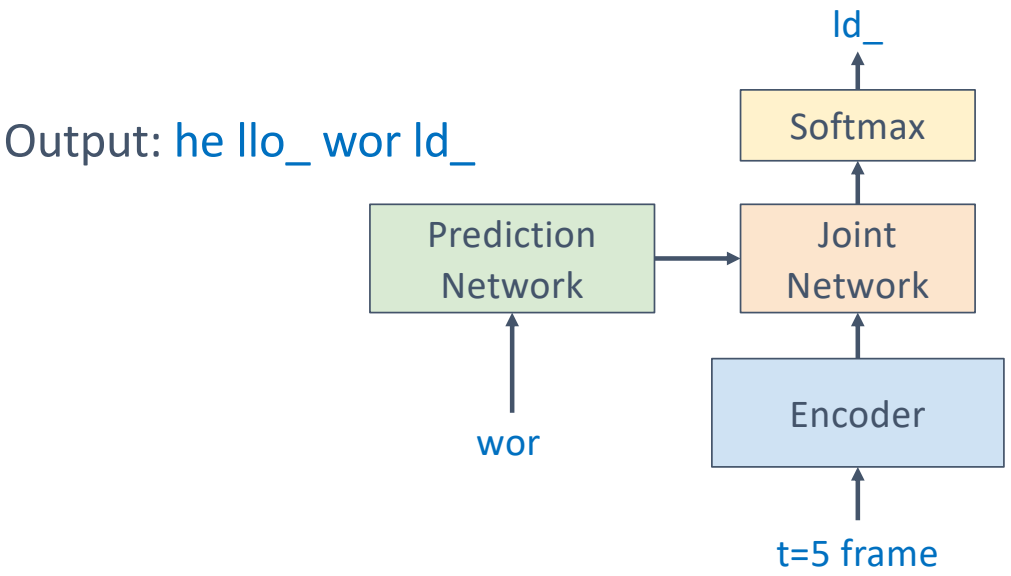
Recurrent Neural Network Transducer (RNN-T)

- Softmax over n + 1 labels, includes a blank like CTC
- <blank> → advance in Encoder, retain prediction network state



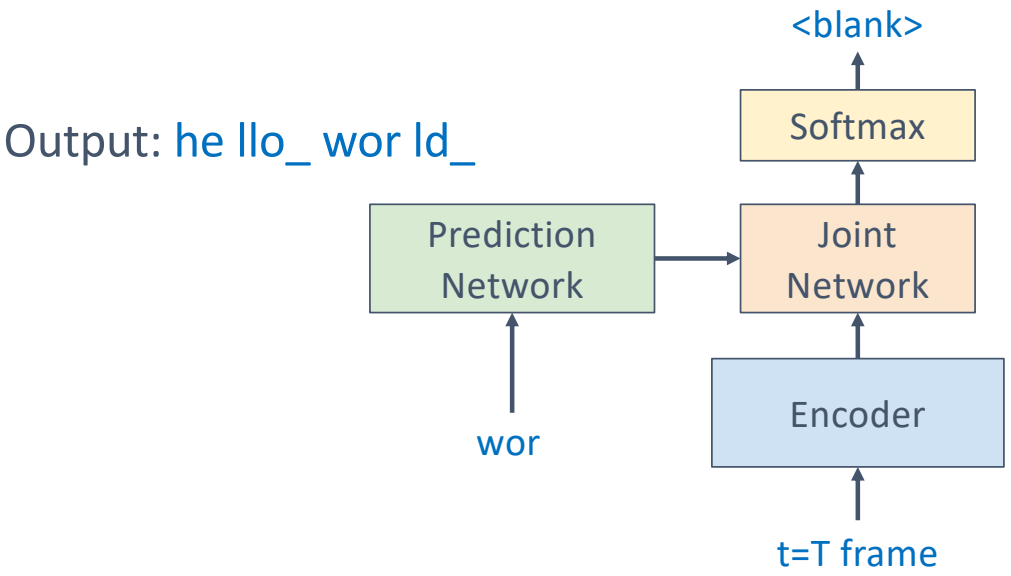
Recurrent Neural Network Transducer (RNN-T)

- Softmax over n + 1 labels, includes a blank like CTC
- <blank> → advance in Encoder, retain prediction network state



Recurrent Neural Network Transducer (RNN-T)

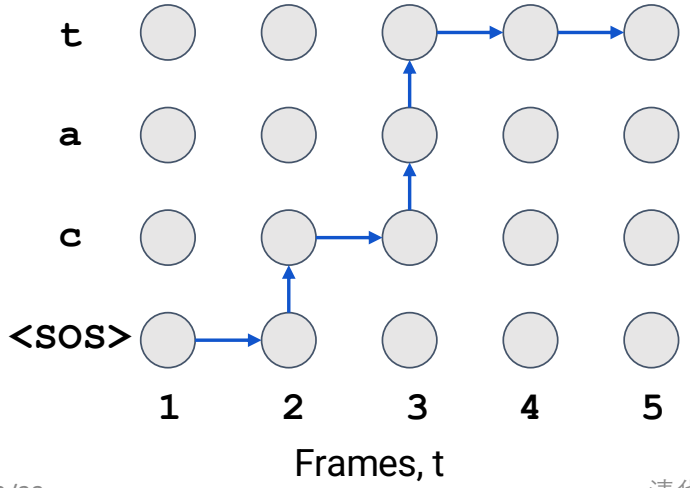
- Softmax over n + 1 labels, includes a blank like CTC
- <blank> → advance in Encoder, retain prediction network state

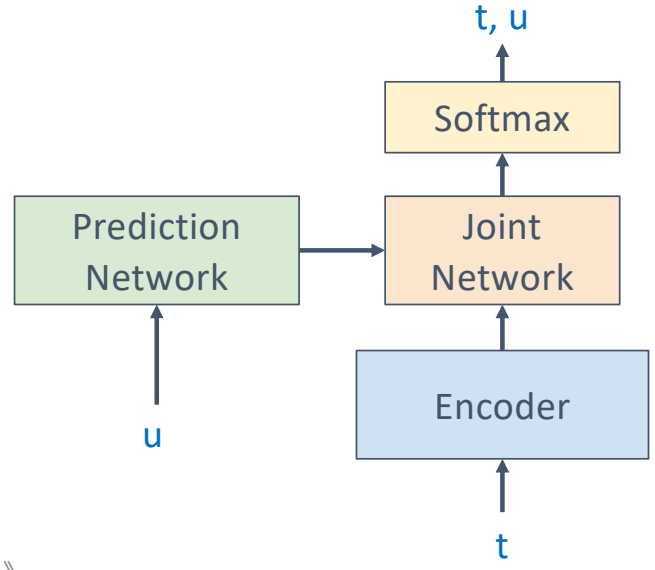


Inference terminates when all input frames have been consumed

Recurrent Neural Network Transducer (RNN-T)

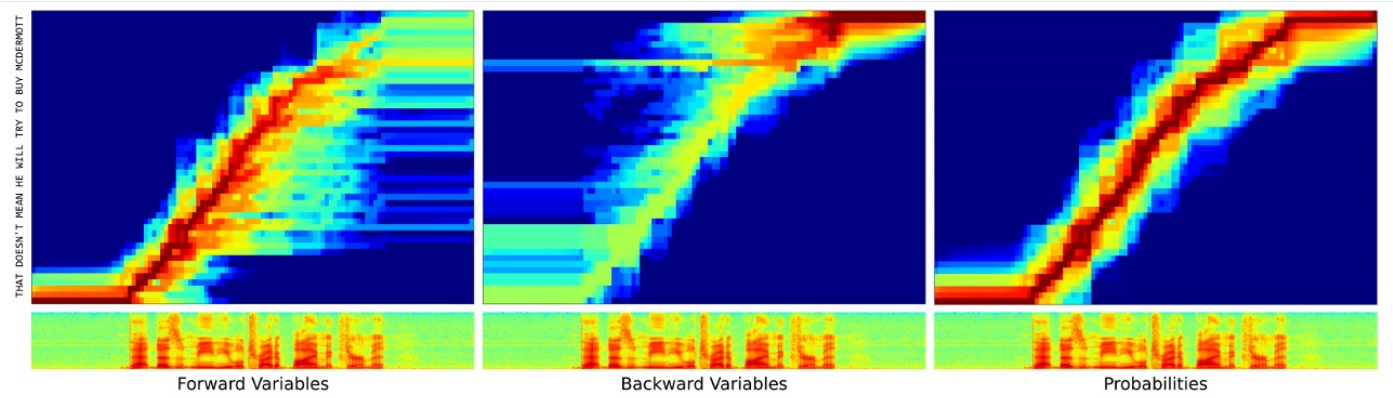
- During training, feed the true label sequence to the LM and use CTC-like forward-backward procedure on all possible alignments.
- Given a target sequence of length U and T acoustic frames we generate UxT softmax





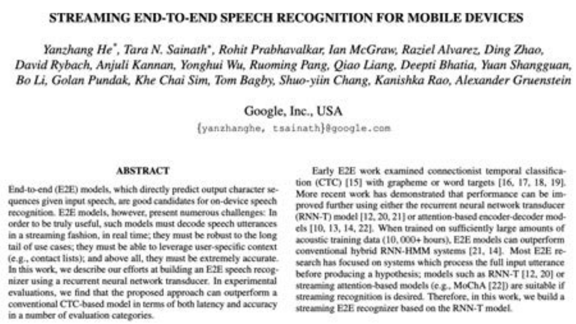
Recurrent Neural Network Transducer (RNN-T)

- During training, feed the true label sequence to the LM and use CTC-like forward-backward procedure on all possible alignments.
- Given a target sequence of length U and T acoustic frames we generate UxT softmax



Recurrent Neural Network Transducer (RNN-T)

- It is natural for RNN-T to handle streaming ASR outputs, and thus RNN-T was more widely used so far.
- An example of streaming grapheme ASR is provided here.



STREAMING END-TO-END SPEECH RECOGNITION FOR MOBILE DEVICES

Yanzhang He, Tara N. Sainath*, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, Qiao Liang, Deepti Bhatia, Yuan Shangguan, Bo Li, Golan Pundak, Khe Chai Sim, Tom Bagby, Shuo-yiin Chang, Kanishka Rao, Alexander Gruenstein*

Google, Inc., USA

{yanzhanghe, tsainath}@google.com

ABSTRACT

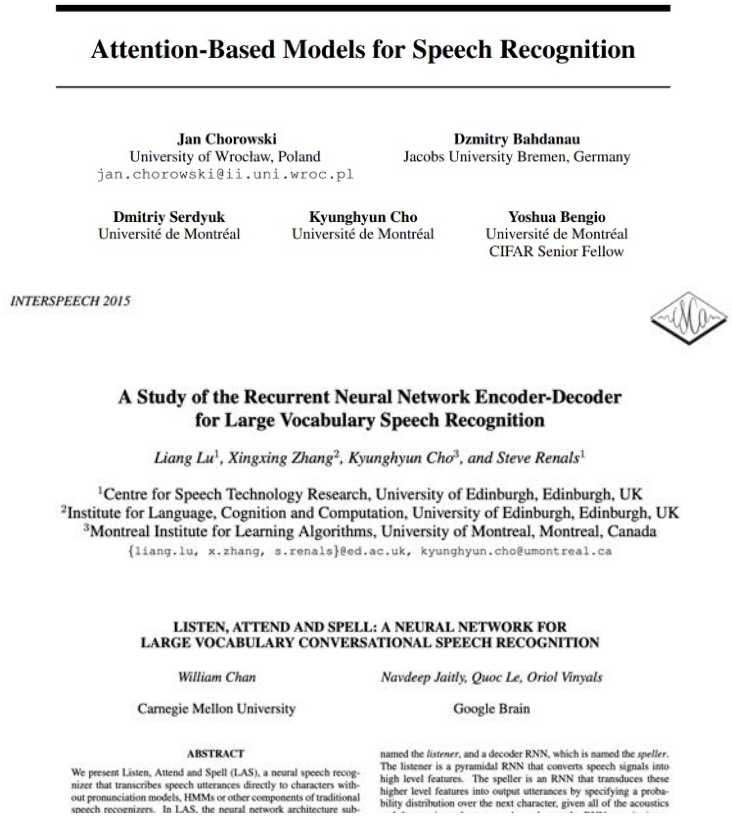
End-to-end (E2E) models, which directly predict output character sequences given input speech, are good candidates for on-device speech recognition. E2E models, however, present numerous challenges: In order to be truly useful, such models must decode speech utterances in a streaming fashion, in real time; they must be robust to the long tail of use cases; they must be able to leverage user-specific context (e.g., contact lists); and above all, they must be extremely accurate. In this work, we describe our efforts at building an E2E speech recognizer using a recurrent neural network transducer. In experimental evaluations, we find that the proposed approach can outperform a conventional CTC-based model in terms of both latency and accuracy in a number of evaluation categories.

Early E2E work examined connectionist temporal classification (CTC) [15] with grapheme or word targets [16, 17, 18, 19]. More recent work has demonstrated that performance can be improved further using either the recurrent neural network transducer (RNN-T) model [12, 20, 21] or attention-based encoder-decoder models [10, 13, 14, 22]. When trained on sufficiently large amounts of acoustic training data (10, 000+ hours), E2E models can outperform conventional hybrid RNN-HMM systems [21, 14]. Most E2E research has focused on systems which process the full input utterance before producing a hypothesis; models such as RNN-T [12, 20] or streaming attention-based models (e.g., MoChA [22]) are suitable if streaming recognition is desired. Therefore, in this work, we build a streaming E2E recognizer based on the RNN-T model.

[Y. He and T.N. Sainath et al., 2018]

Attention-based Encoder-Decoder (AED)

- AED emerged first in the context of neural machine translation. [Chorowski et al., 2015]
- First applied to ASR by [Chorowski et al., 2015] [Lu et al., 2015] [Chan et al., 2016]. [Lu et al., 2015]
- Use attention to align input and output sequences. [Chan et al., 2016]



Attention-Based Models for Speech Recognition

Jan Chorowski
University of Wrocław, Poland
jan.chorowski@ii.uni.wroc.pl

Dzmitry Bahdanau
Jacobs University Bremen, Germany

Dmitriy Serdyuk
Université de Montréal

Kyunghyun Cho
Université de Montréal

Yoshua Bengio
Université de Montréal
CIFAR Senior Fellow

INTERSPEECH 2015

A Study of the Recurrent Neural Network Encoder-Decoder for Large Vocabulary Speech Recognition

Liang Lu[1], Xingxing Zhang[2], Kyunghyun Cho[3], and Steve Renals[1]

[1]Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
[2]Institute for Language, Cognition and Computation, University of Edinburgh, Edinburgh, UK
[3]Montreal Institute for Learning Algorithms, University of Montreal, Montreal, Canada
{liang.lu, x.zhang, s.renals}@ed.ac.uk, kyunghyun.cho@umontreal.ca

LISTEN, ATTEND AND SPELL: A NEURAL NETWORK FOR LARGE VOCABULARY CONVERSATIONAL SPEECH RECOGNITION

William Chan
Carnegie Mellon University

Navdeep Jaitly, Quoc Le, Oriol Vinyals
Google Brain

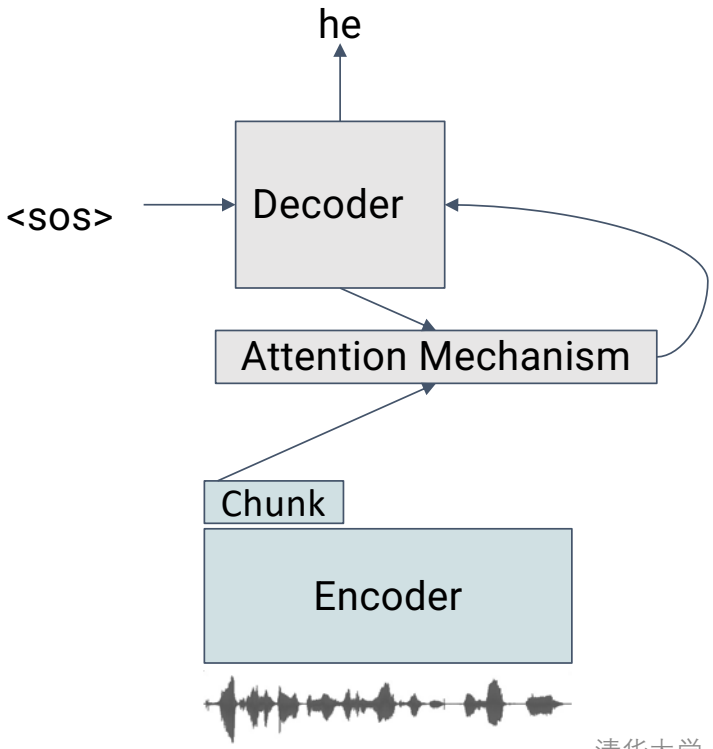
ABSTRACT

We present Listen, Attend and Spell (LAS), a neural speech recognizer that transcribes speech utterances directly to characters without pronunciation models, HMMs or other components of traditional speech recognizers. In LAS, the neural network architecture sub-

named the *listener*, and a decoder RNN, which is named the *speller*. The listener is a pyramidal RNN that converts speech signals into high level features. The speller is an RNN that transduces these higher level features into output utterances by specifying a probability distribution over the next character, given all of the acoustics

Attention-based Encoder-Decoder (AED)

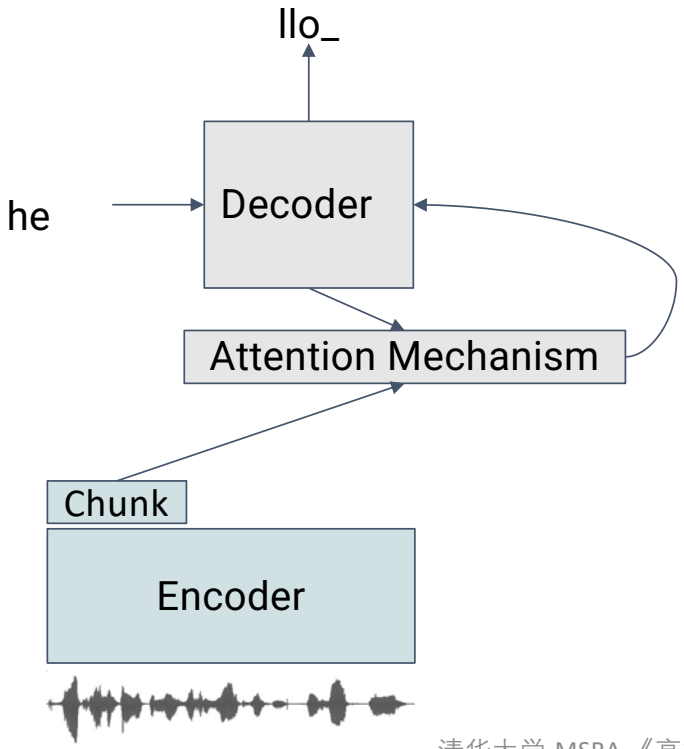
- Decoder runs once per-label (e.g. a world-piece model)



Output: he

Attention-based Encoder-Decoder (AED)

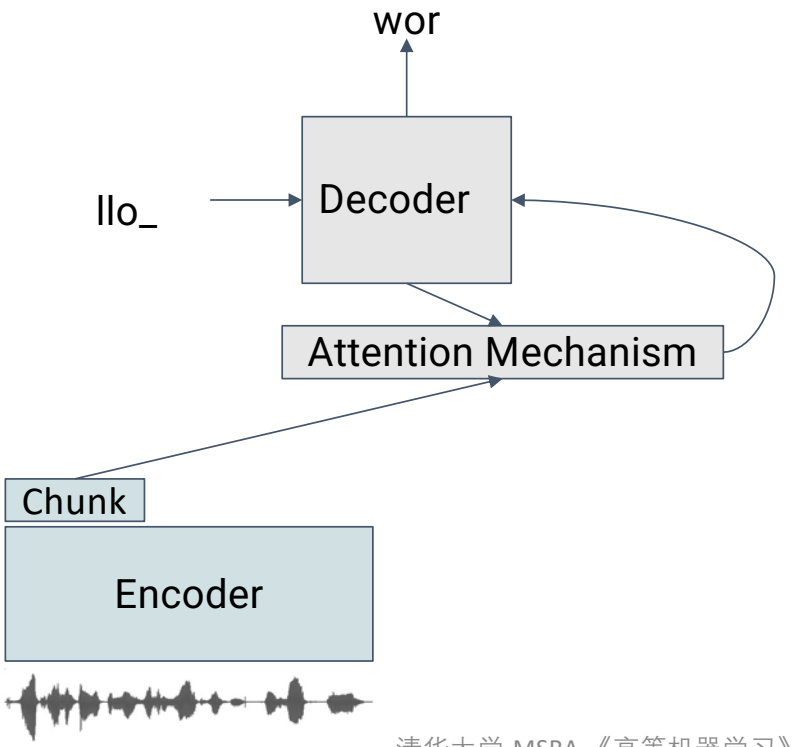
- Decoder runs once per-label (e.g. a world-piece model)



Output: he llo_

Attention-based Encoder-Decoder (AED)

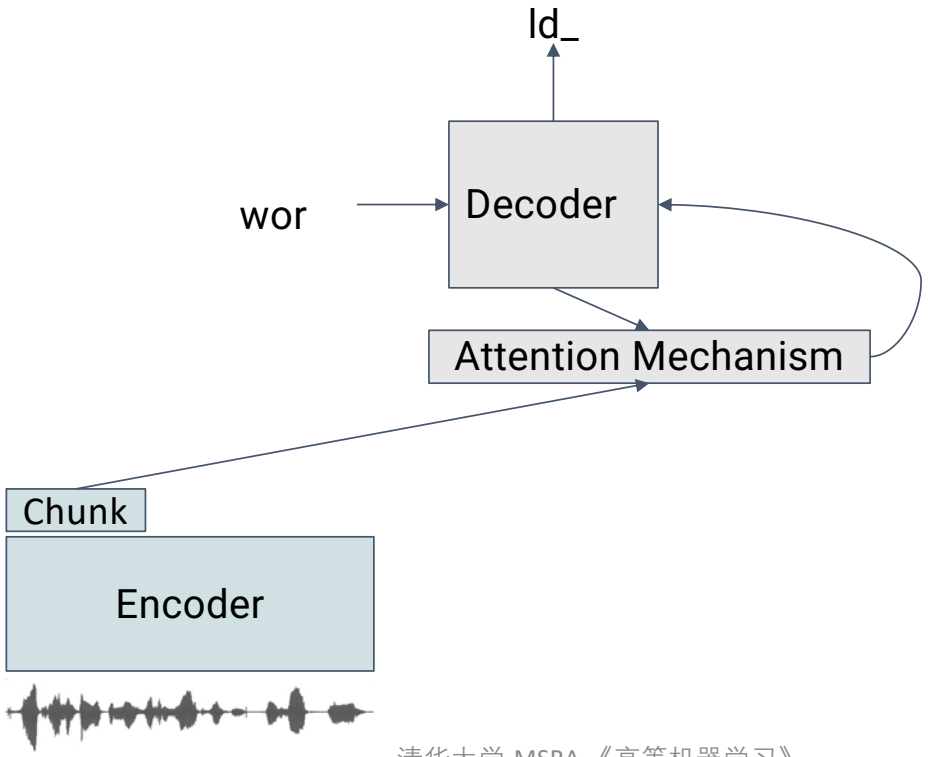
- Decoder runs once per-label (e.g. a world-piece model)



Output: he llo_ wor

Attention-based Encoder-Decoder (AED)

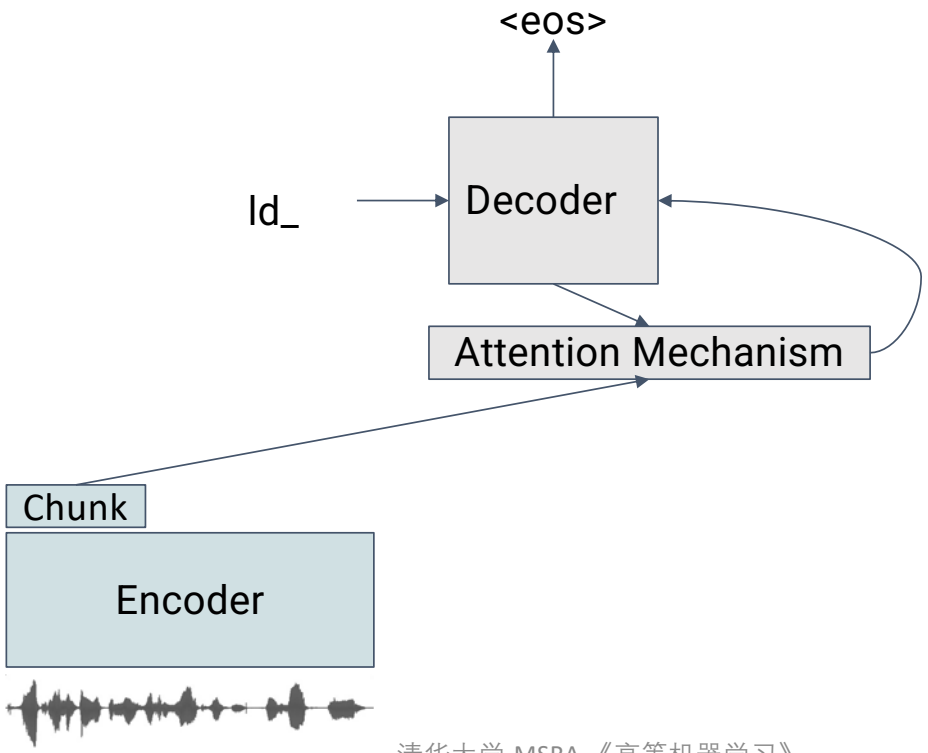
- Decoder runs once per-label (e.g. a world-piece model)



Output: he llo_ wor ld_

Attention-based Encoder-Decoder (AED)

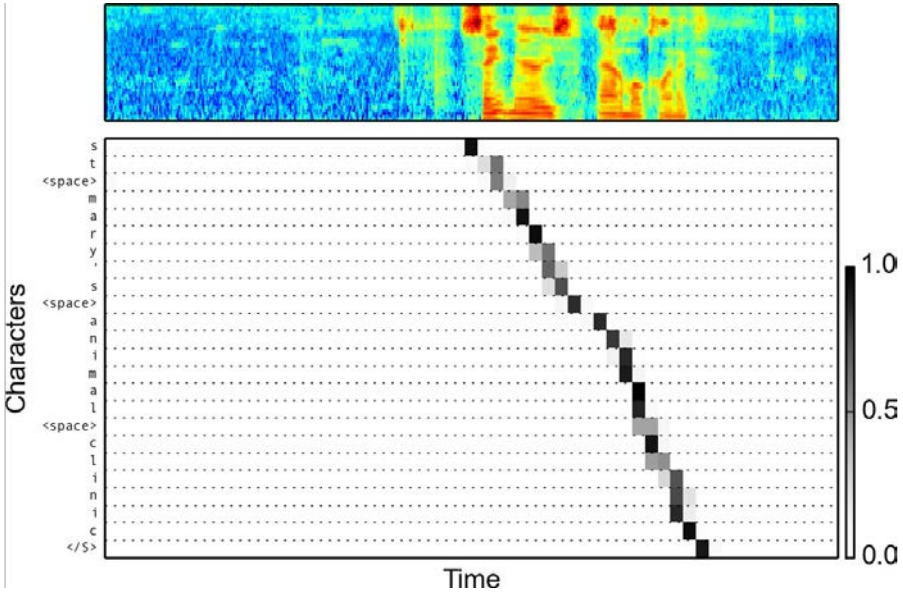
- Decoder runs once per-label (e.g. a world-piece model)



Output: he llo_ wor ld_

Attention-based Encoder-Decoder (AED)

- Alignment achieved purely using the attention mechanism.
  - Location sensitive attention for RNN-based AED.
  - Cross-attention for Transformer.



Results

- Results here compared CTC, RNN-T, with AED (a bit out-of-date).
- AED performs the best, but cannot be used for streaming ASR.

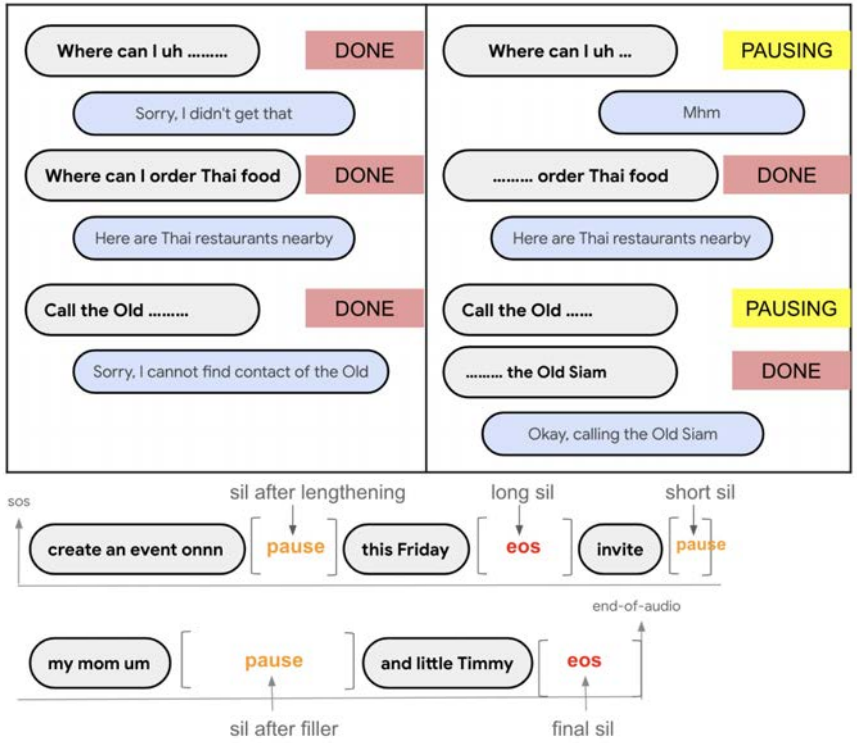
Model	Clean	
	Dictation	VoiceSearch
Unidirectional CD-phone	6.4	9.9
Bidirectional CD-phone	**5.4**	**8.6**
CTC-grapheme	39.4	53.4
RNN-T	6.6	12.8
AED	6.6	11.7

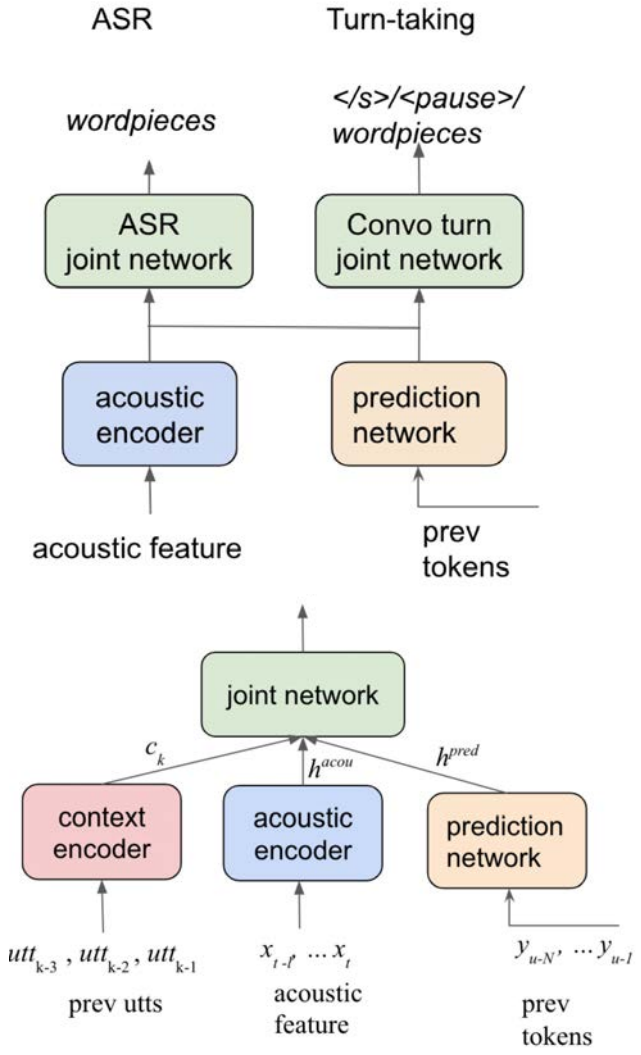
Comparing End-to-end To Modularised

- Cons:
  - Unstable (Hallucinations)
  - No time information provided (forced alignment)
  - Hesitation etc.
  - Possibly still higher WERs
- Pros:
  - Simpler
  - More flexible
    - More languages
    - More tasks

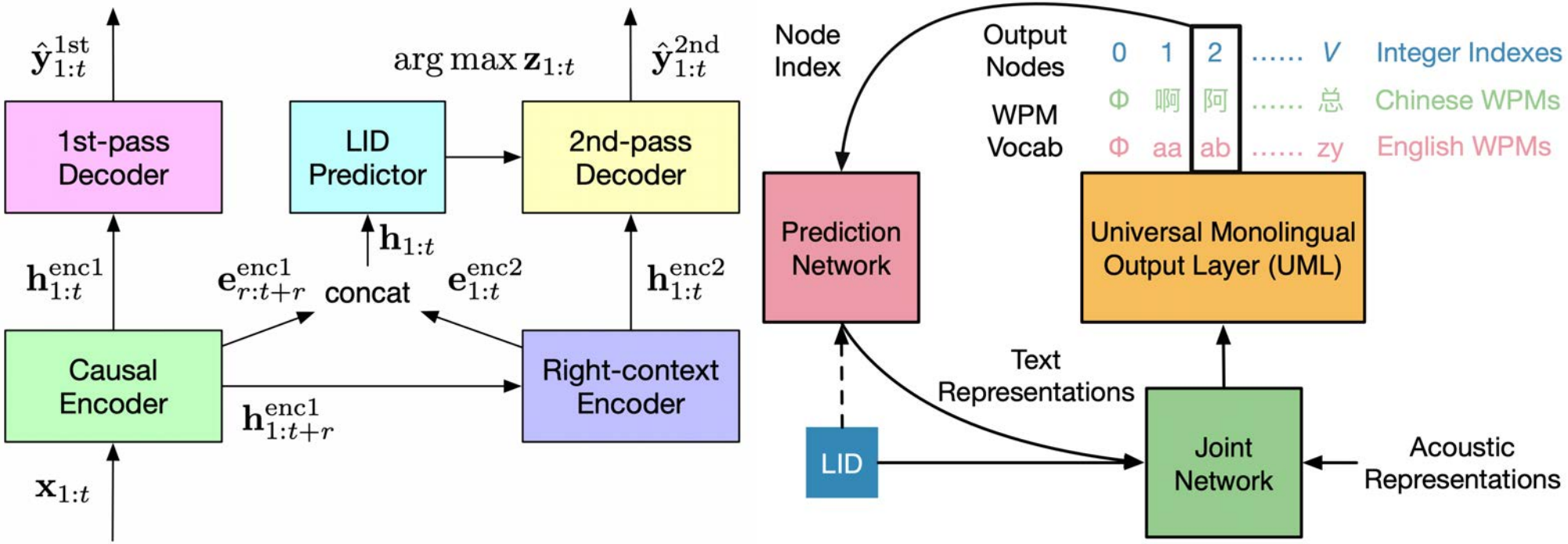
Hesitations & Long-Context

- Natural Conversation



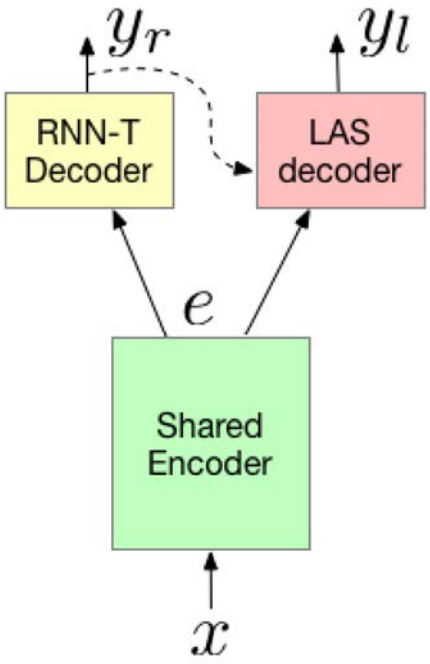


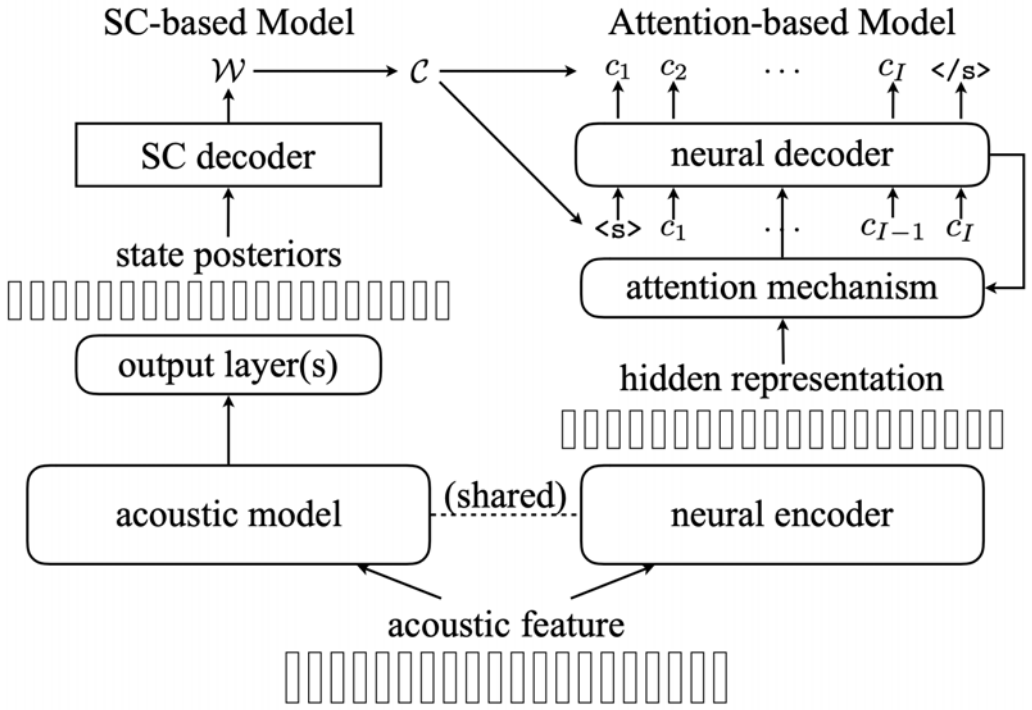
Multilingual ASR with Language Identification



System Combination?

- Rescoring: Two pass end-to-end & Integrating source-channel & AED





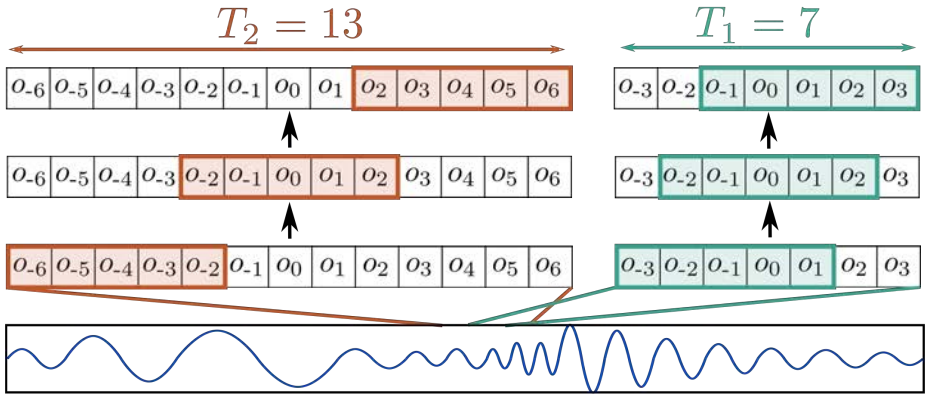
Bottom-Up Probabilistic Transduction

Bottom-Up Probabilistic Transduction

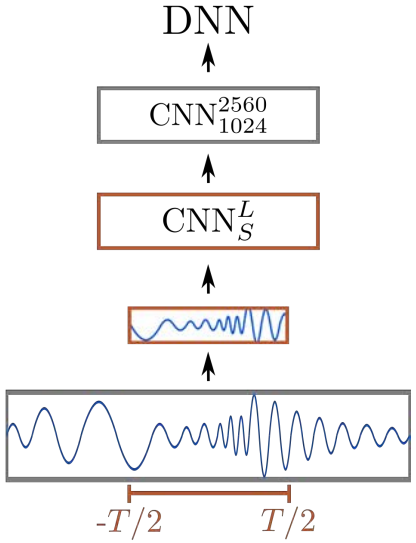
*

Raw Waveform Input Features

- CNN can learn filters from raw waveform features efficiently.
- Different kernels and strides are useful; DNN layer can also be used.
- Leant features similar to log-Mel filterbanks, but more data dependent.



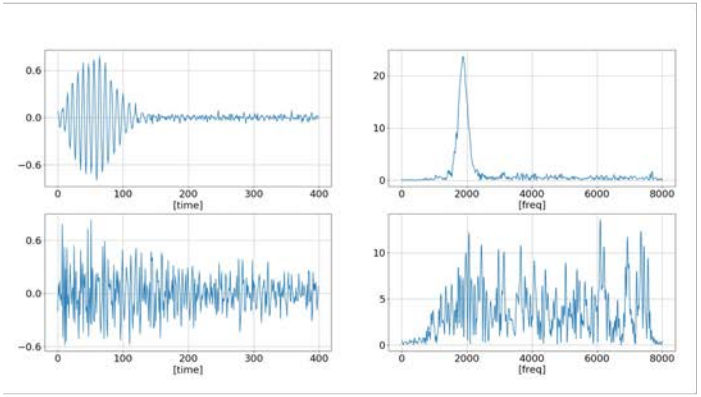
1-dimensional convolution along time with different strides.



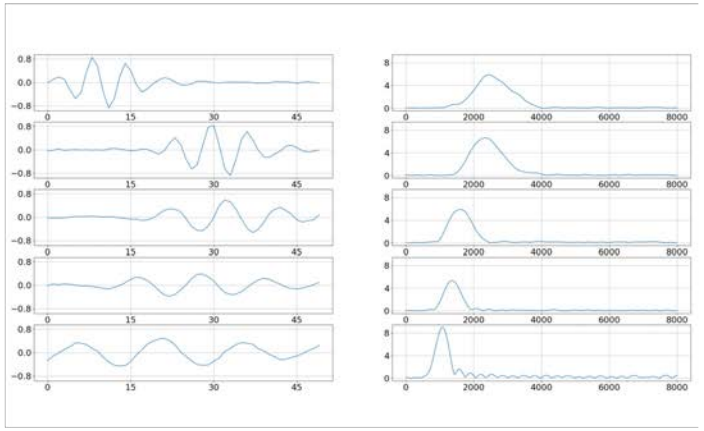
CNN layers for raw features.

Raw Waveform Input Features

- A CNN layer can have filter size (L), stride (S), zero-padding num., filter num. (N), and number of input maps.
- Regarding 1D CNN layer, here we only consider, L, S, and N.



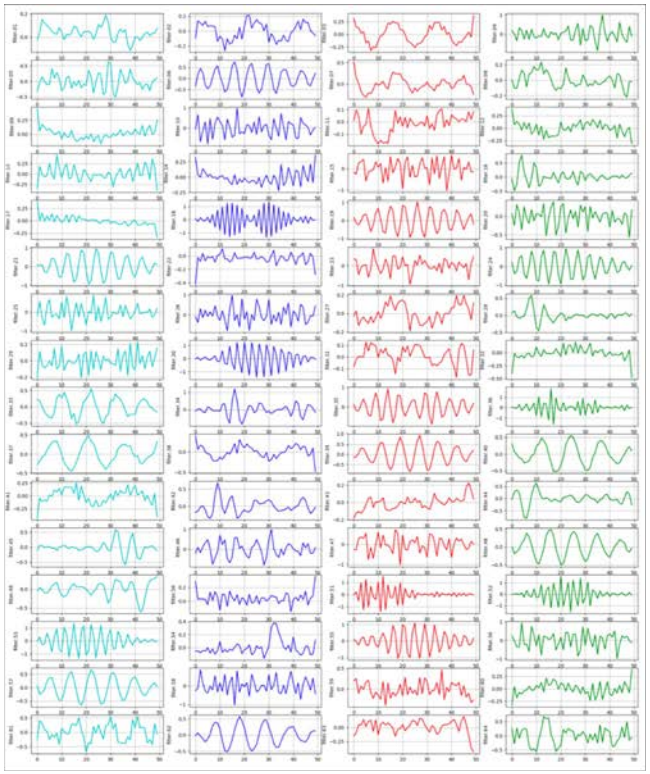
Clean and noisy filters (52 and 27) learnt on a noisy dataset.



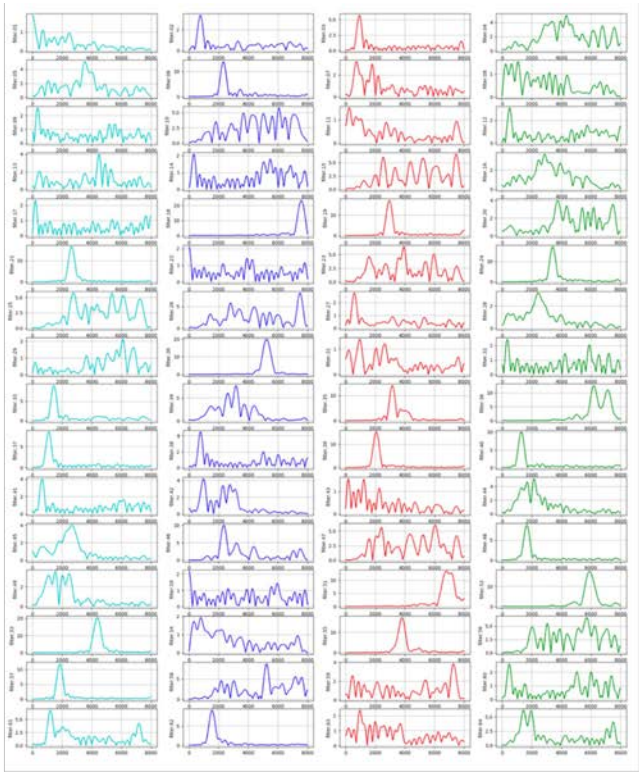
Wavelet-style filters (56, 49, 1, 27, 42) learnt with CNN layers.

Raw Waveform Input Features

- A CNN layer often has many filters.
- A model can have a few CNN layers with big filters or many layers with small filters (works better).
- The model can have many branches of CNN layer stacks by varying stride numbers and input channels etc.



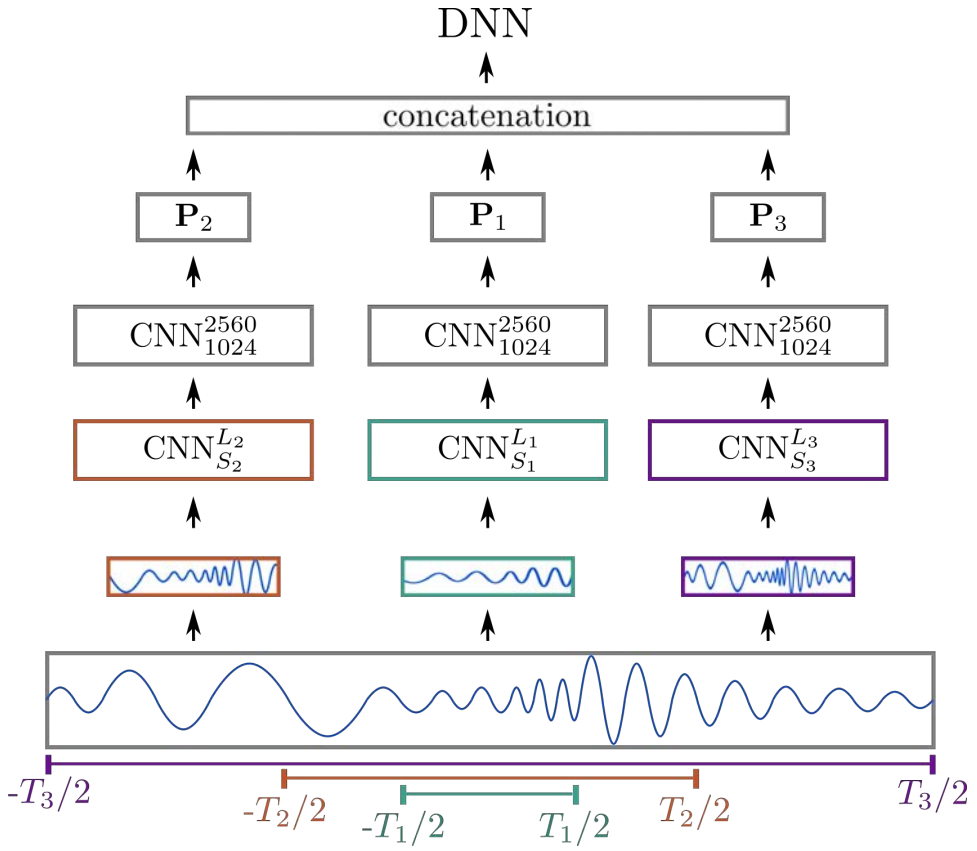
64 learnt filters of a CNN layer in time domain.



64 learnt filters of a CNN layer in frequency domain.

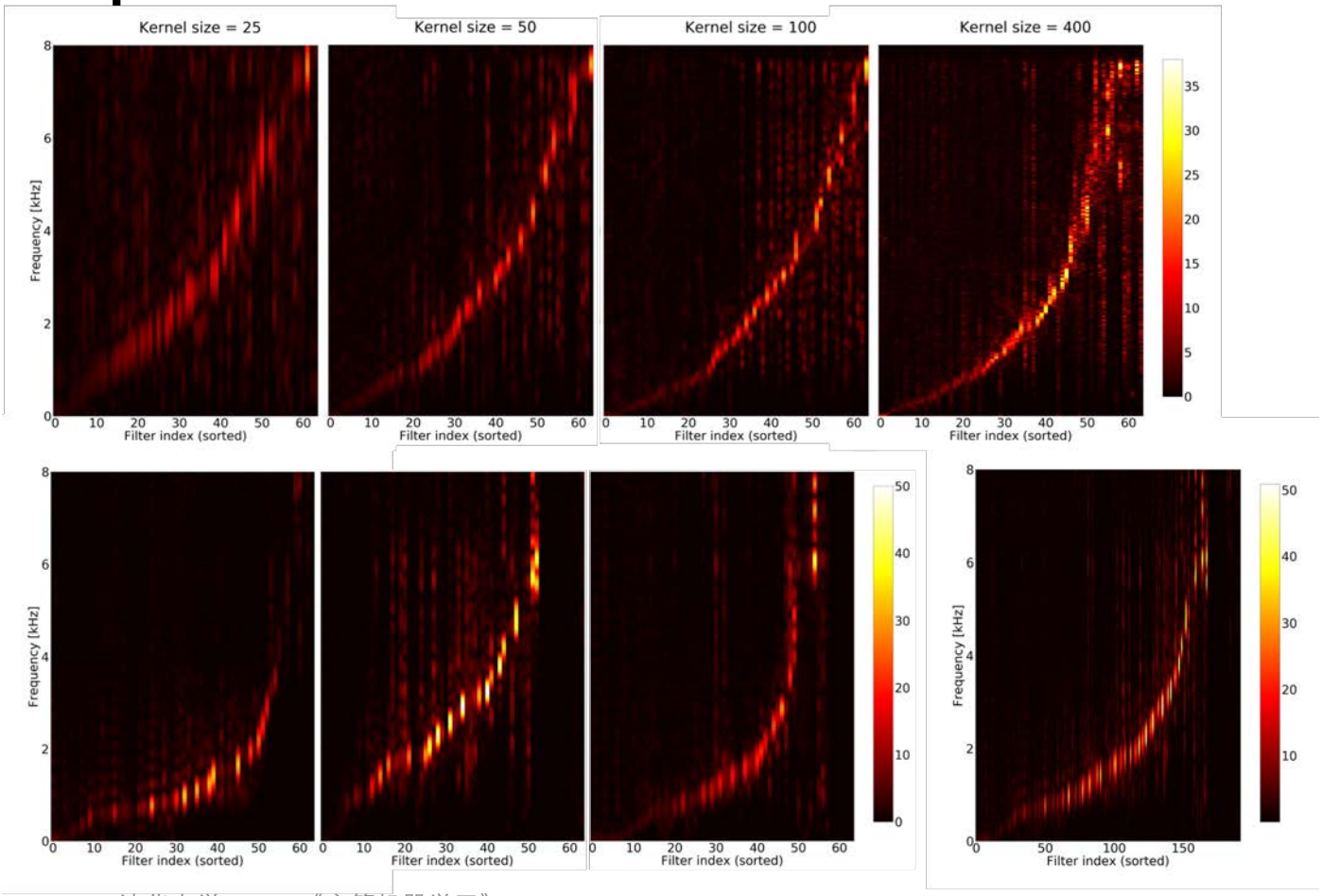
Raw Waveform Input Features

- There exists several known structures that enable models with raw waveform features to work better than hand-crafted features even on small datasets:
  - Multi-span structure with multiple CNN stacks with different strides;
  - SincNet structure with sinc-function-based for filters.



Raw Waveform Input Features

- CNN filters learnt with different sizes.
- CNN filters learnt with size=50 and different stride num (4, 9, 15).
- Log-Mel filterbank style features learned.



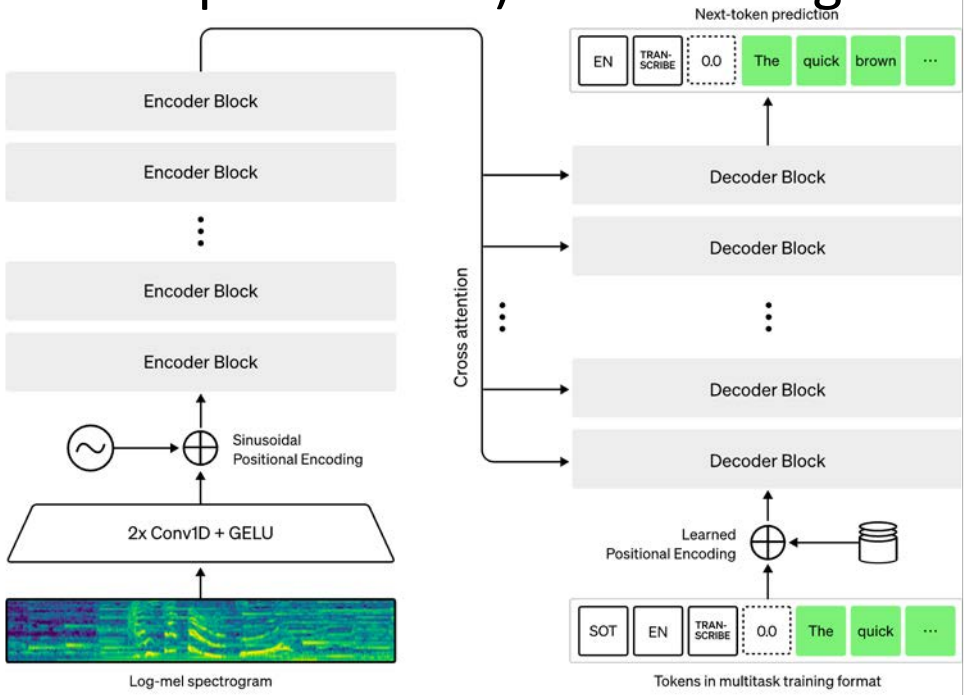
Bottom-Up Probabilistic Transduction

Bottom-Up Probabilistic Transduction

*

OpenAI Whisper

- Whisper was trained on 680,000 hours speech from 96 languages (with 150M parameters) for multilingual ASR and speech translation.



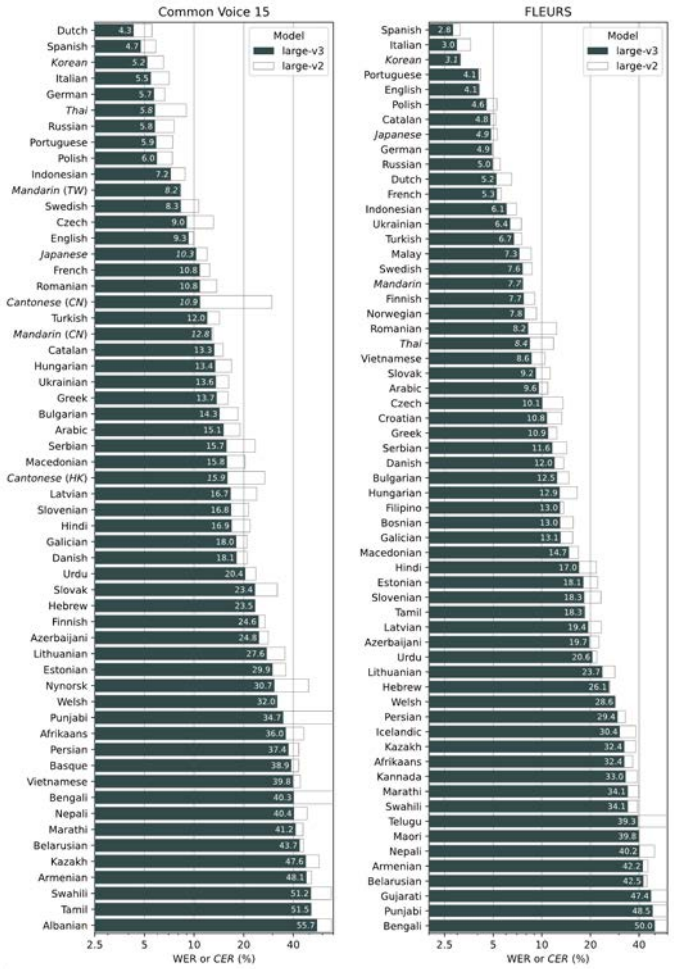
This is the Micro Machine Man presenting the most midget miniature motorcade of Micro Machines. Each one has dramatic details, terrific trim, precision paint jobs, plus incredible Micro Machine Pocket Play Sets There's a police station, fire station, restaurant, service station, and more.

어둠만이 나의 전부였던 동안
While darkness was my everything
숨이 벅차도록 달려왔잖아
I ran so hard that I ran out of breath
Never say "time's up"
Never say time's up
경계의 끝자락
Like the end of the boundary
내 끝은 아니니까
Because my end is not the end

A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision", 2022.

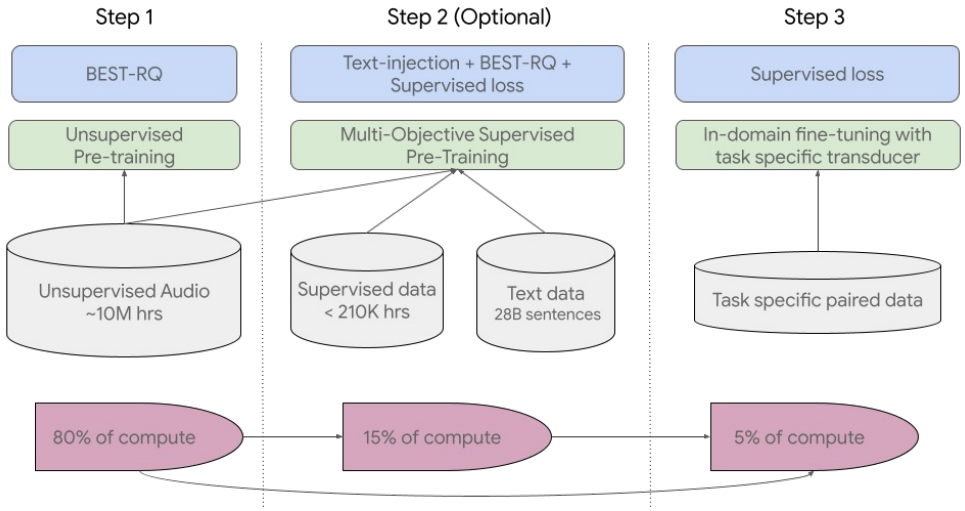
OpenAI Whisper-v3

- Used 128-d instead of 80-d log-Mel filterbank features
- Used 1M hours of weakly labelled data and 4M hours data labelled by Whisper-v2
- Achieved 10-20% relatively lower WERs on Common Voice 15 and Fleurs than Whisper-v2

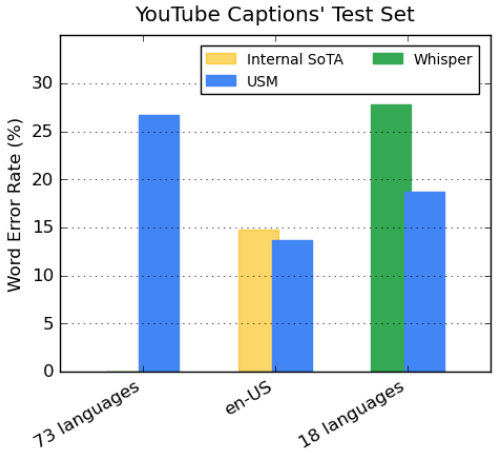


Google Universal Speech Model (USM)

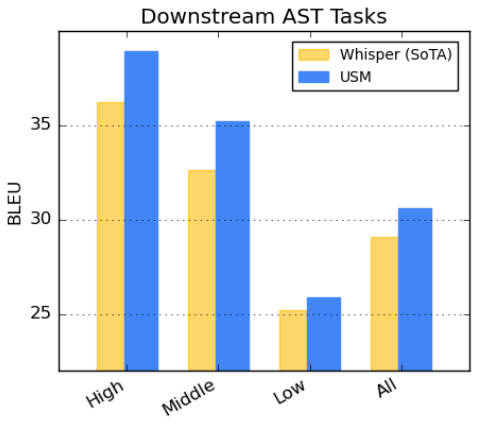
- USM was trained on 10M hours of speech (no labels) and 2M hours of supervised speech from 300 languages (with 2B parameters).



USM Training pipeline



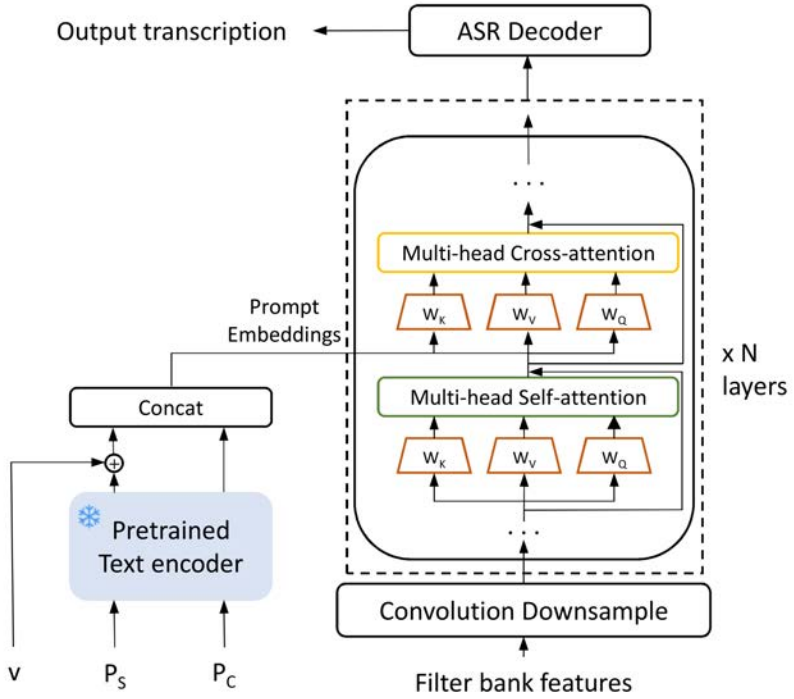
WERs on Youtube test sets



Downstream task speech translation

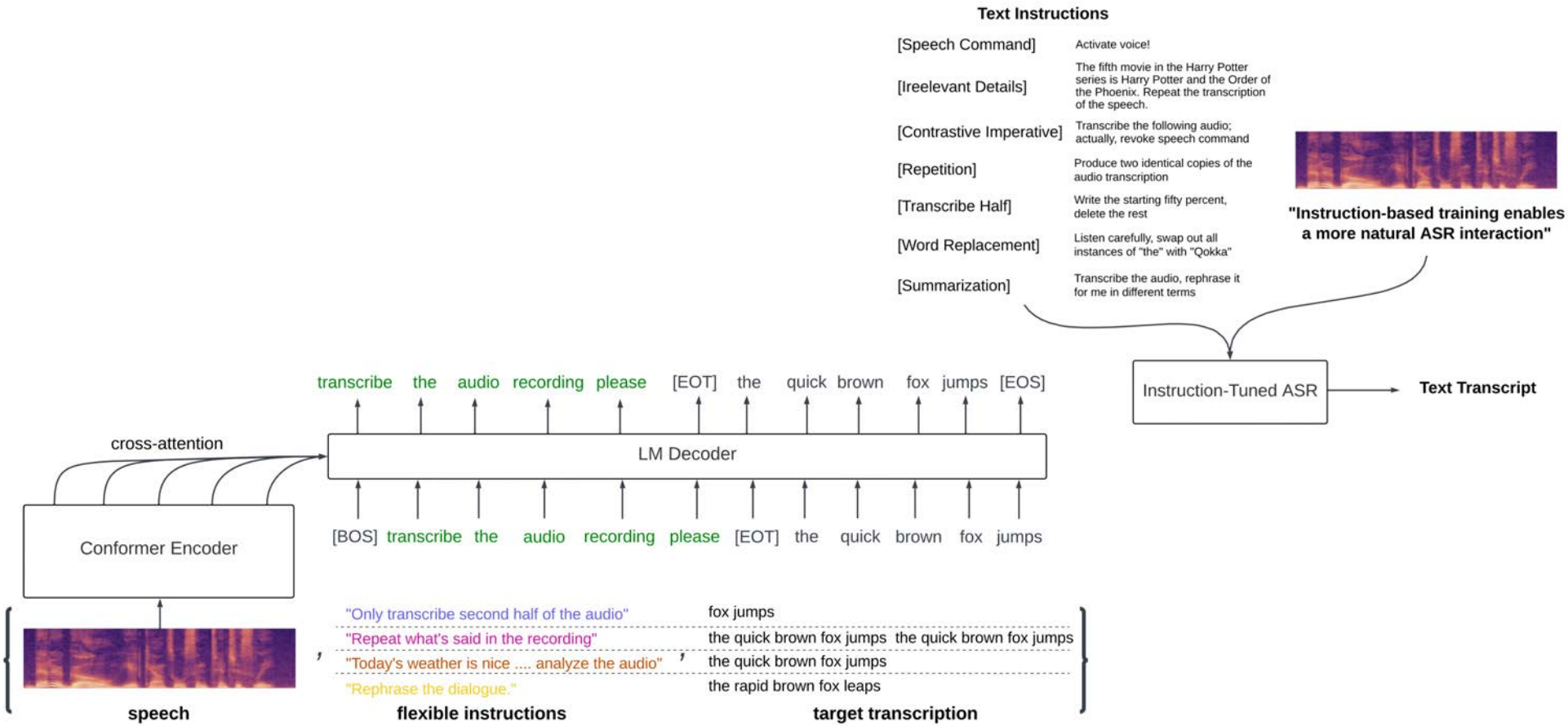
Y. Zhang et al.,"Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages", 2023.

Prompt ASR



Style Prompt	WITHOUT CASING OR PUNCTUATION
Content Prompt	Welcome to the UEFA Champions League final!
Reference text	TODAY'S MATCH IS BETWEEN REAL MADRID AND LIVERPOOL
Style Prompt	Mixed-cased English with punctuation
Content Prompt	Welcome to the UEFA Champions League final!
Reference text	Today's match is between Real Madrid and Liverpool.

Prompt ASR



Text Instructions

[Speech Command]	Activate voice!
[Ireelevant Details]	The fifth movie in the Harry Potter series is Harry Potter and the Order of the Phoenix. Repeat the transcription of the speech.
[Contrastive Imperative]	Transcribe the following audio; actually, revoke speech command
[Repetition]	Produce two identical copies of the audio transcription
[Transcribe Half]	Write the starting fifty percent, delete the rest
[Word Replacement]	Listen carefully, swap out all instances of "the" with "Qokka"
[Summarization]	Transcribe the audio, rephrase it for me in different terms

"Instruction-based training enables a more natural ASR interaction"

Instruction-Tuned ASR → **Text Transcript**

transcribe the audio recording please [EOT] the quick brown fox jumps [EOS]

cross-attention

LM Decoder

Conformer Encoder

[BOS] transcribe the audio recording please [EOT] the quick brown fox jumps

speech	**flexible instructions**	**target transcription**
	"Only transcribe second half of the audio"	fox jumps
	"Repeat what's said in the recording"	the quick brown fox jumps the quick brown fox jumps
	"Today's weather is nice .... analyze the audio"	the quick brown fox jumps
	"Rephrase the dialogue."	the rapid brown fox leaps

Auditory Large Language Model

SALMONN: Speech Audio Language Music Open Neural Network

- Developed by Tsinghua EE Dept. & ByteDance
- SALMONN is the first AI model with human-like general hearing abilities
  - Generic hearing: speech content, audio event, music
- Connecting audio encoders to LLM
  - Model input: audio content X , text instruction I
  - Model output: text response Y
- For content censorship, search and recommendation etc.

C. Tang et al.,“Towards General Hearing Abilities for Large Language Models”, 2023.

SALMONN: Speech Audio Language Music Open Neural Network

- Tasks used in instruction tuning Level 1
  - (Overlapping) ASR, audio and music captioning, gender and speaker recognition, emotion recognition, English to Chinese translation
- Untrained tasks can also be achieved by ASR + LLM Level 2
  - English to other language translation, speech understanding, spoken-query-based QA
- Untrained tasks need generic hearing abilities Level 3
  - Music emotion recognition, title writing for music, speech audio co-reasoning (SAC)
  - Audio-based storytelling & SAC are new AI tasks
- Demonstrate cross-modal emergent abilities zero-shot tasks
  - All level 2 and level 3 tasks

C. Tang et al.,"Towards General Hearing Abilities for Large Language Models", 2023.

SALMONN: Speech Audio Language Music Open Neural Network

- Tasks used in instruction tuning Level 1
  - (Overlapping) ASR, audio and music captioning, gender and speaker recognition, emotion recognition, English to Chinese translation
- Untrained tasks can also be achieved by ASR + LLM Level 2
  - English to other language translation, speech understanding, spoken-query-based QA
- Untrained tasks need generic hearing abilities Level 3
  - Music emotion recognition, title writing for music, speech audio co-reasoning (SAC)
  - Audio-based storytelling & SAC are new AI tasks
- Demonstrate cross-modal emergent abilities zero-shot tasks
  - All level 2 and level 3 tasks

C. Tang et al.,"Towards General Hearing Abilities for Large Language Models", 2023.

SALMONN: Speech Audio Language Music Open Neural Network

- Speech reasoning without transcribing texts first

(A part of Steve Jobs' speech: “Your time is limited so don’t waste it living someone else’s life. Don’t be trapped by dogma which is living with the results of other people’s thinking. Don’t let the noise of other’s opinions drown out your own inner voice. And most important have the courage to follow your heart and intuition. They somehow already know what you truly want to become everything else is secondary. When I was young there was an amazing publication called the whole earth catalog, which was one of the bibles of my”)

USER

Give me only three keywords of the text. Explain your reason.

The keywords are 'time', 'limited', and 'courage'.

The reason for choosing these keywords is that the speaker emphasizes the importance of not wasting time living someone else's life, not being trapped by dogma, and having the courage to follow one's own inner voice.

SALMONN

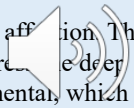
SALMONN: Speech Audio Language Music Open Neural Network

- Music Emotion Recognition

(It's a soothing, elegant piano piece. At the end of the audio a woman begins to sing.)

USER

What is the emotion of the music? Explain the reason in detail.



The music is emotional and evokes feelings of love and affection. The reason for this is that the music is a love song, with lyrics that express deep emotions and feelings of love and affection. The music is also instrumental, which means that it relies solely on the melody and harmony of the instruments to convey these emotions. This creates a more intimate and personal experience for the listener, allowing them to interpret the music in their own way and connect with it on a deeper level.

SALMONN

- Speech Audio Co-Reasoning

(A man asks "Can you guess where I am right now", and explosions and gunshots can be heard in the background.)

USER

Please answer the speaker's question in detail based on the background sound.

Based on the background sound, it seems like the speaker is in a war zone or a combat situation. The sound of gunfire and explosions can be heard in the background. The speaker is asking if the listener can guess where they are.

SALMONN

SALMONN: Speech Audio Language Music Open Neural Network

- Video Understanding (with Speech)

FAVOR

Video Clip:

Audio Description:
Jack: "Open your eyes". Rose: "I'm flying, Jack"
Background music: "My Heart Will Go On"

HUMAN

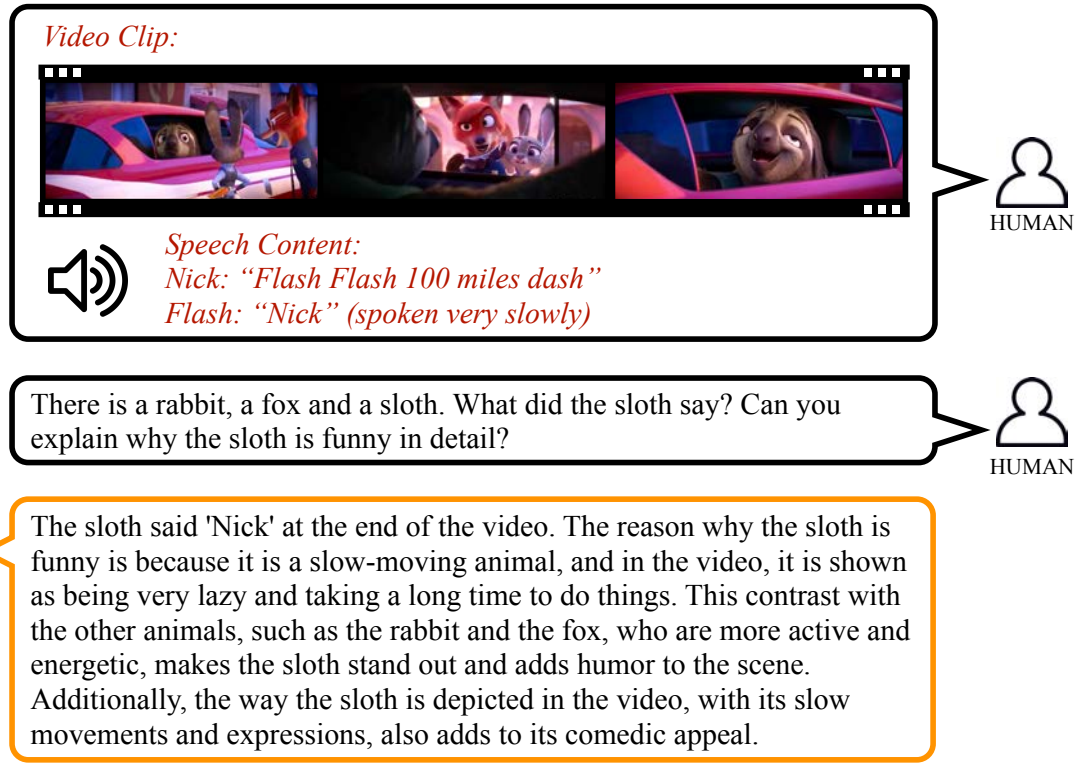
Explain in detail why this video together with the audio and what they say is romantic

HUMAN

The video is a romantic scene of a man and a woman on a boat. The man is holding the woman in his arms, and they are both looking at the sunset. The audio is a song that adds to the romantic atmosphere. The woman says "I'm flying" and "Jack," which suggests that they are happy and enjoying the moment. The setting of the boat and the sunset create a beautiful and serene environment that enhances the romantic feel of the video. The man and the woman's body language and facial expressions also convey their love and affection for each other. Overall, the video is a perfect representation of a romantic and intimate moment between two people.

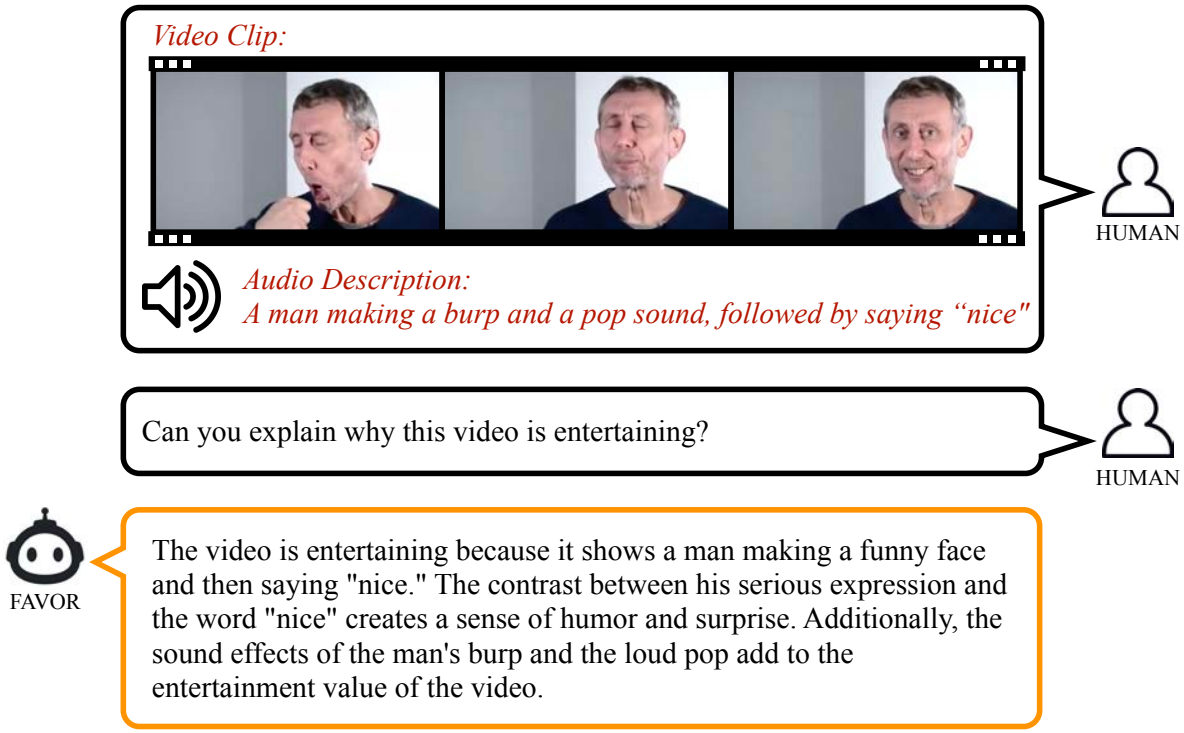
SALMONN: Speech Audio Language Music Open Neural Network

- Video Understanding (with Speech)



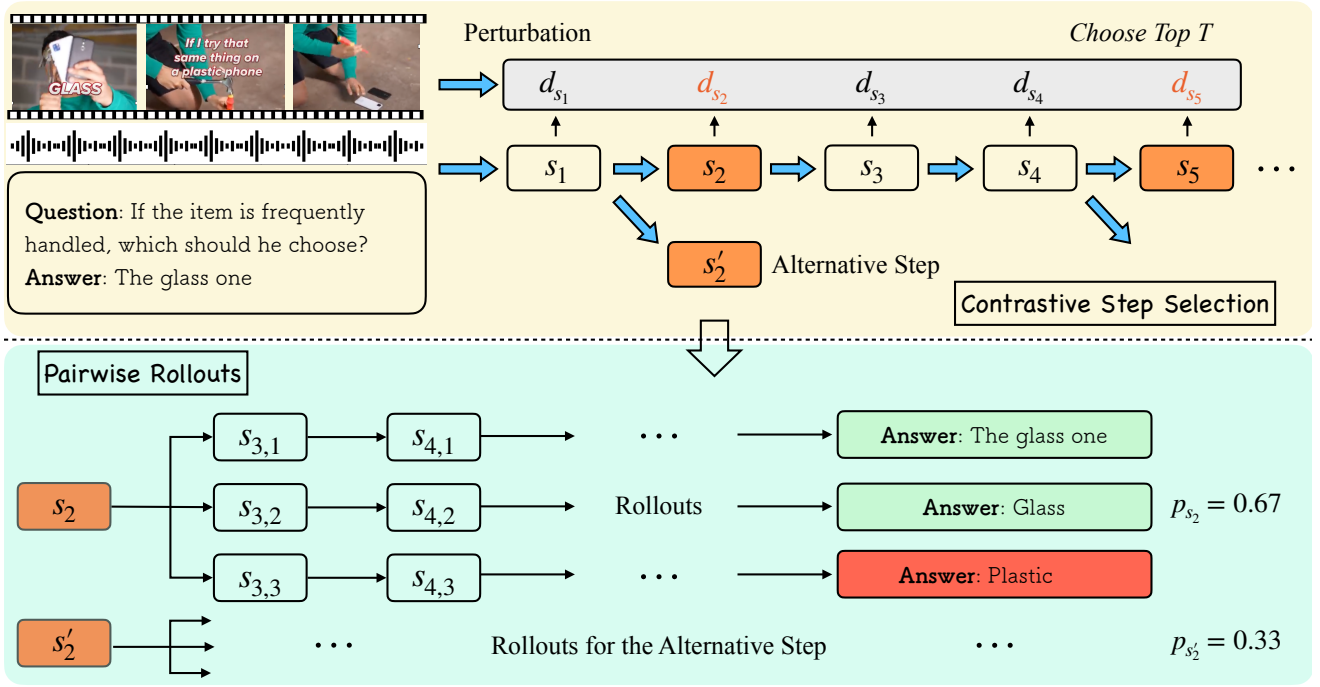
SALMONN: Speech Audio Language Music Open Neural Network

- Video Understanding (with Speech)



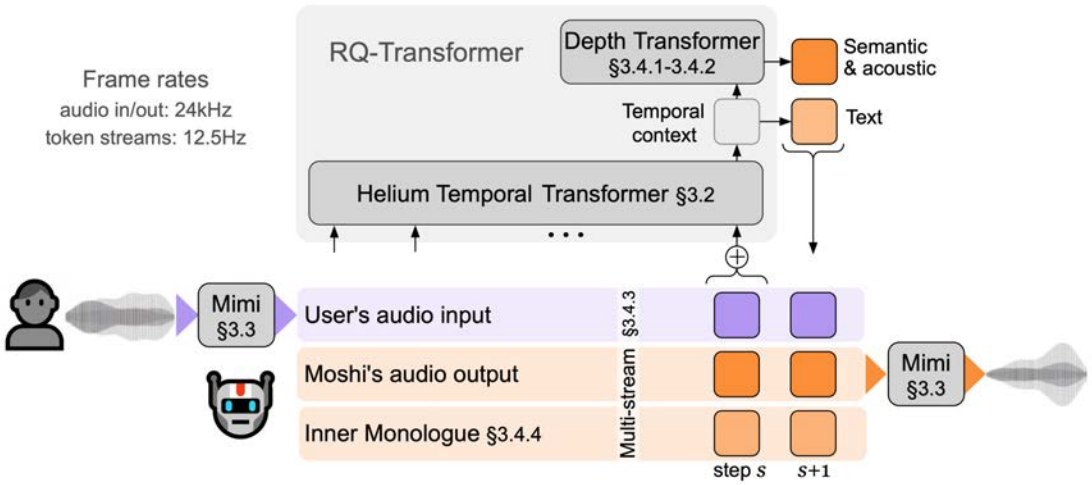
video-SALMONN-o1

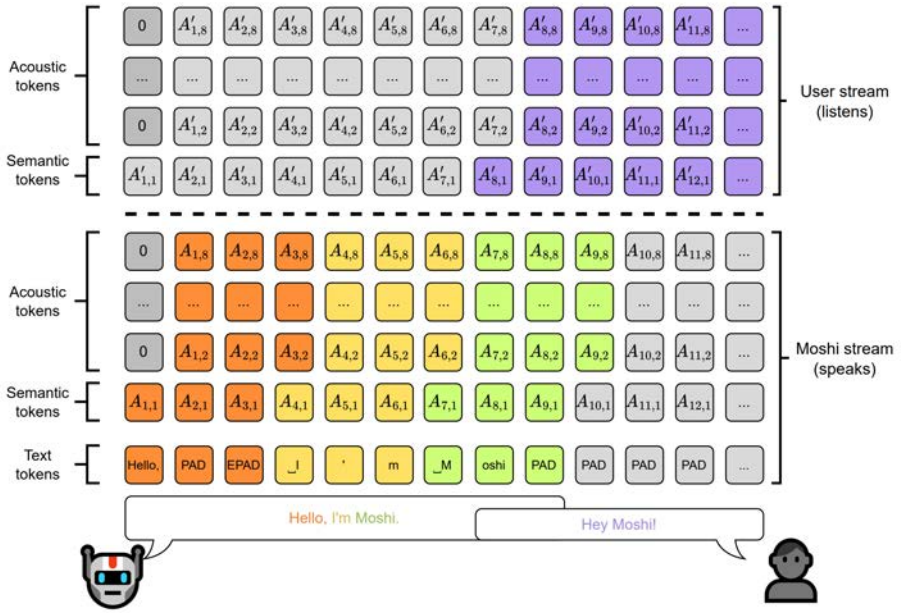
- StandUp, Academic Talk, Generated Video Detection
- https://anonymous.4open.science/r/video-SALMONN-o1-4B0B/README.md



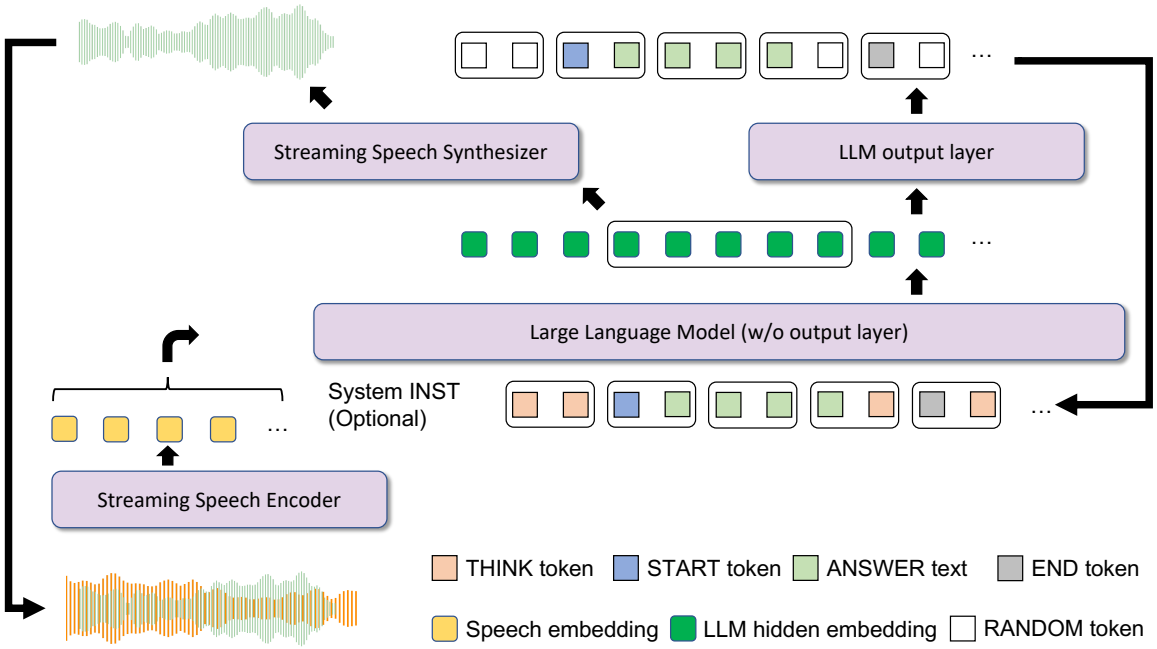
Full-Duplex Speech LLM – Moshi

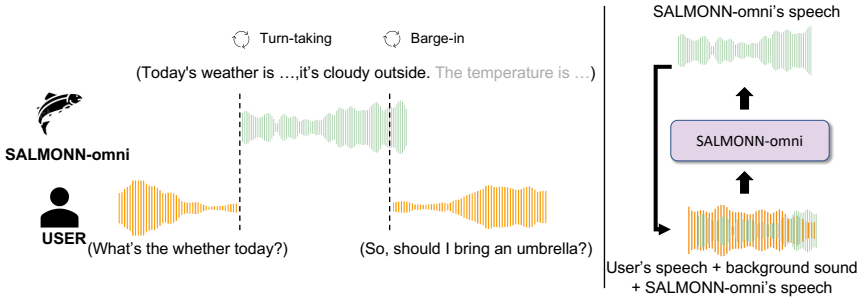
Demo

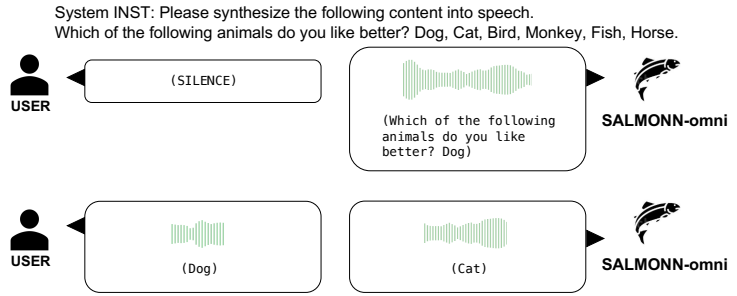




Full-Duplex Speech LLM – SALMONN-omni







Ameca Humanoid Robot AI Platform

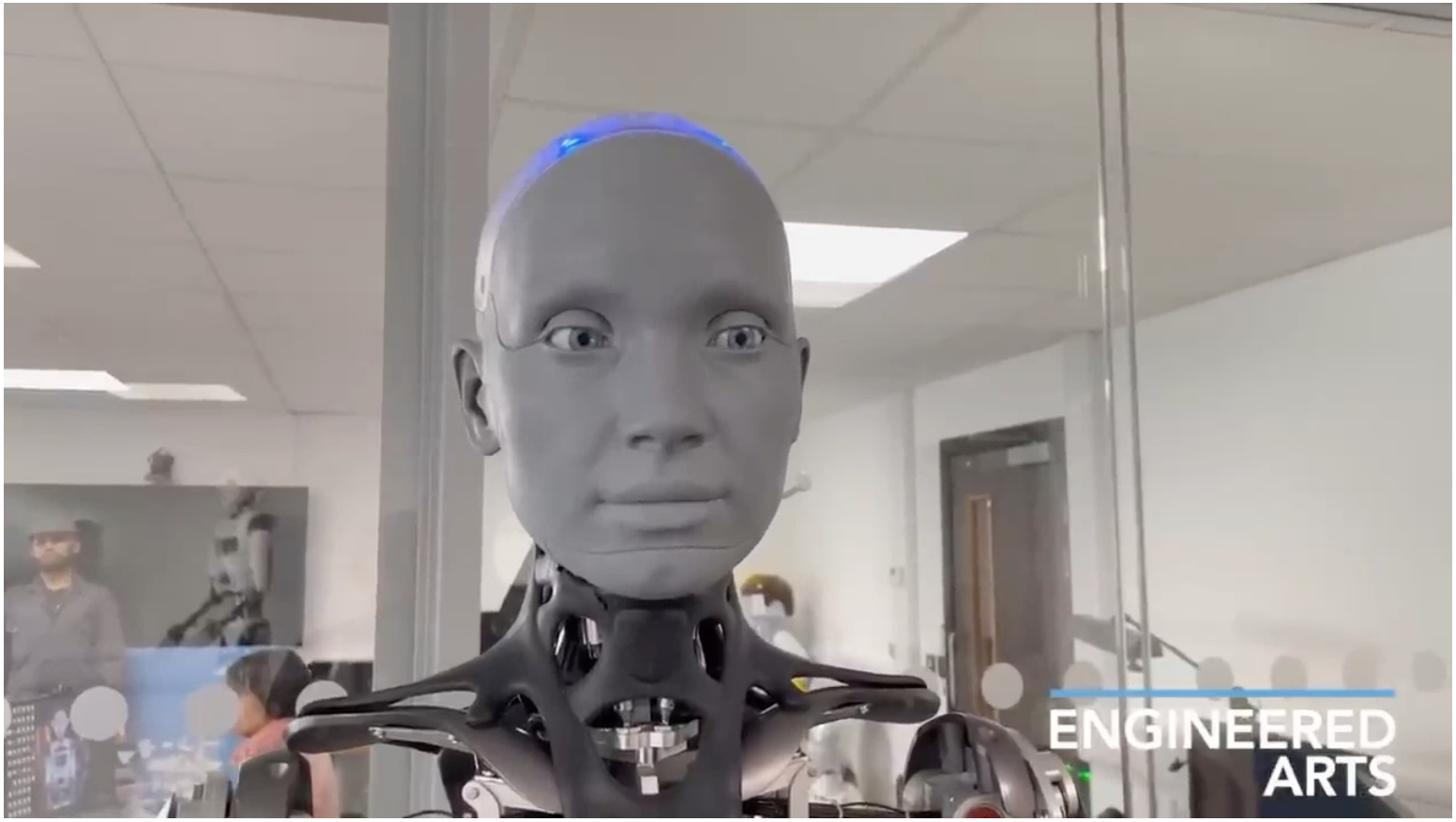
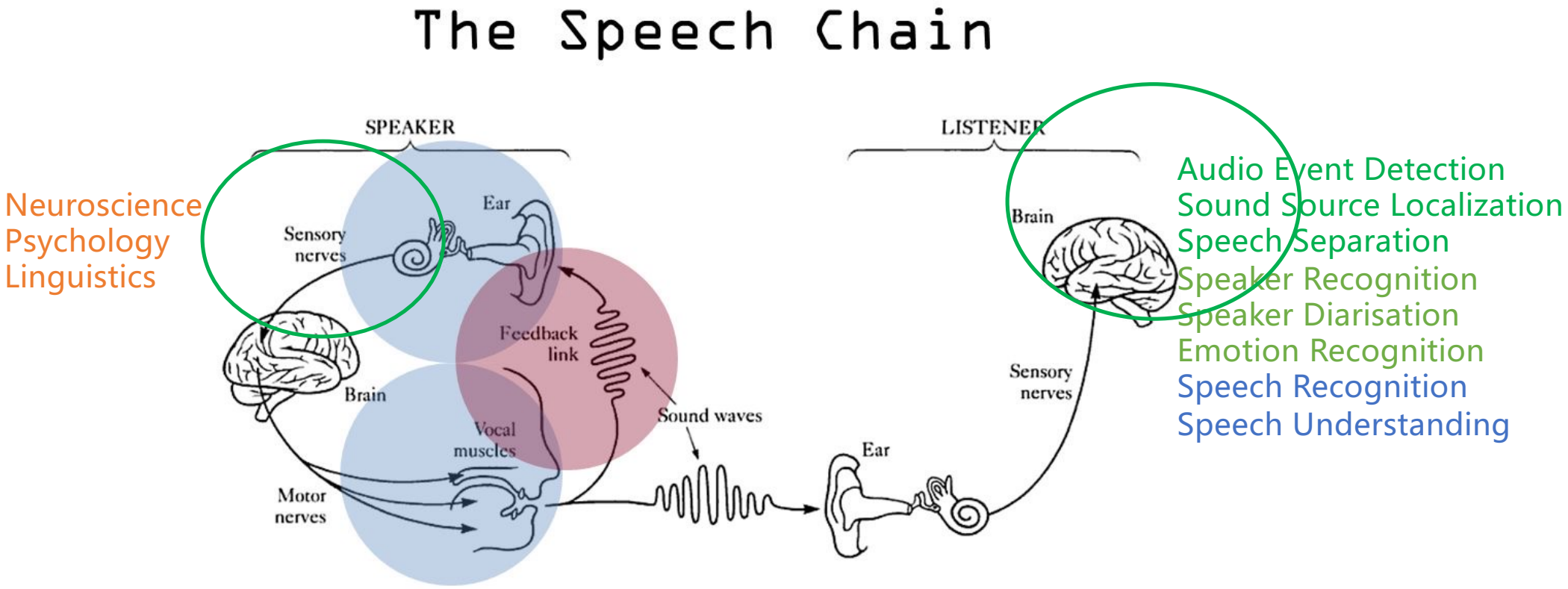


Figure: Speech-to-Speech Reasoning

Speech Science & Technology



Neuroscience
Psychology
Linguistics

Audio Event Detection
Sound Source Localization
Speech Separation
Speaker Recognition
Speaker Diarisation
Emotion Recognition
Speech Recognition
Speech Understanding

Denes & Pinson, 1993

Thanks for your listening!