深蓝学院
shenlanxueyuan.com

Speech Technology: Frontiers and Applications

From GMM-HMM to End-to-End

Xiangang Li, Guoguo Chen

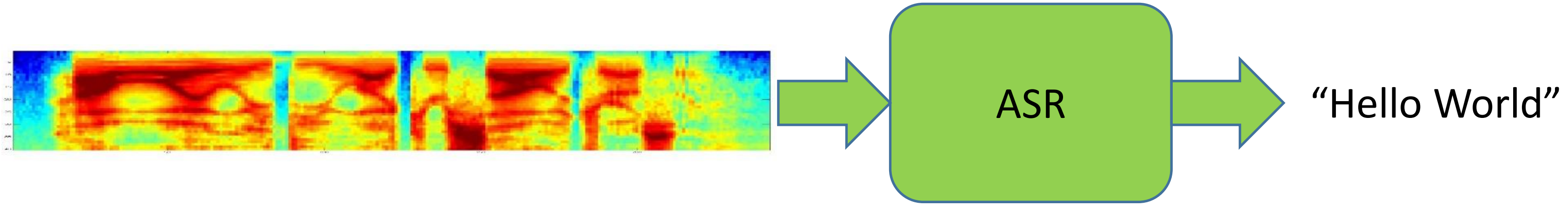
Outline

- Speech recognition: classic methods
- Speech recognition: DNN-HMM approaches
- Speech recognition: end-to-end approaches

Outline

- Speech recognition: classic methods
- Speech recognition: DNN-HMM approaches
- Speech recognition: end-to-end approaches

Speech recognition: basic concepts

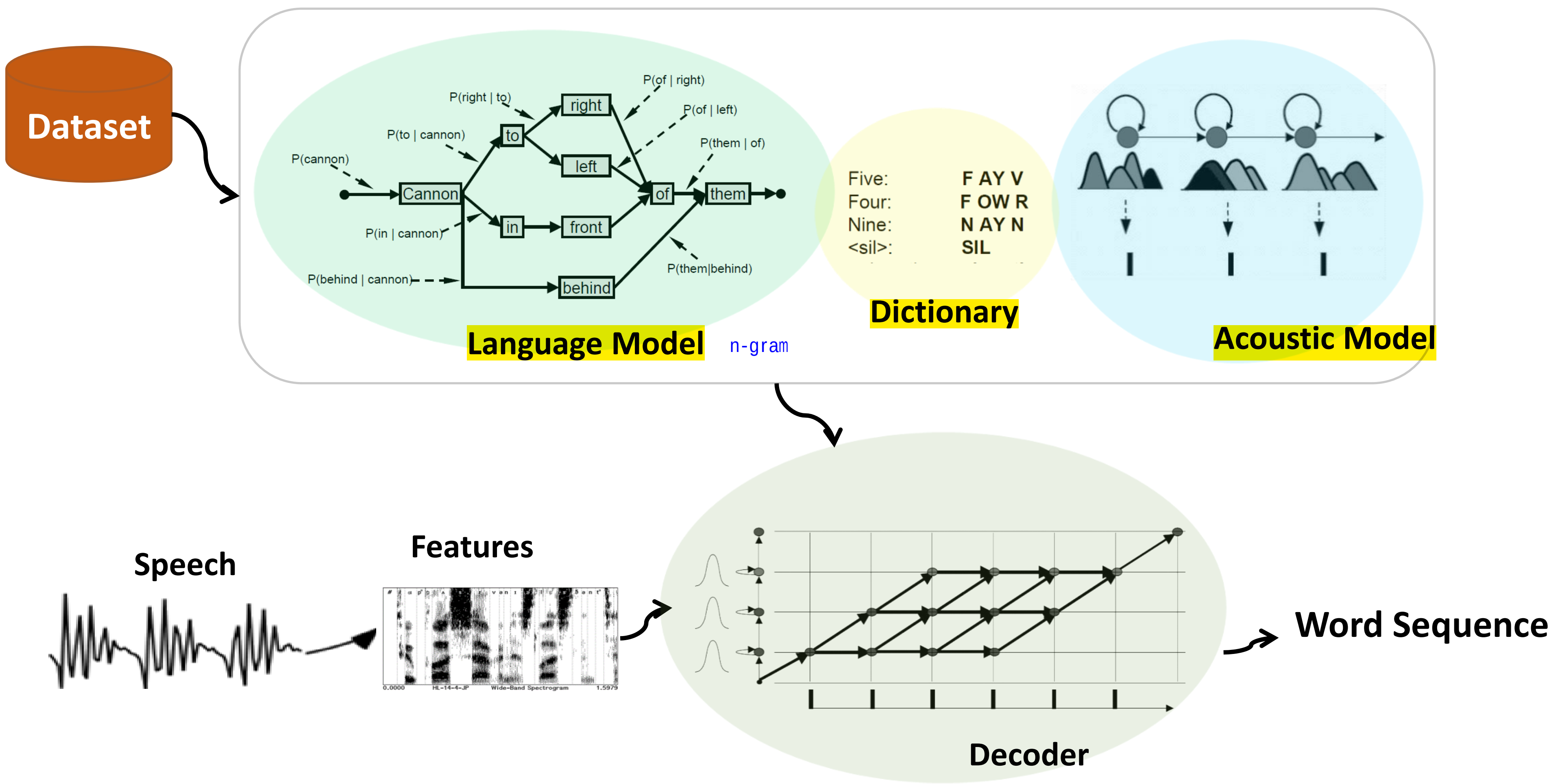


- Speech signal -> Transcripts

$$\hat{W} = \arg\max_{W} p(W|X) = \arg\max_{W} \frac{p(W)p(X|W)}{p(W)}$$
$$= \arg\max_{W} p(W)p(X|W)$$

- Three main parts:
  - Acoustic Model: $p(X|W)$
  - Language Model: $p(W)$
  - Decoder: $\arg\max\,(\cdot)$

Speech recognition: basic concepts



Speech recognition: classic methods

深蓝学院
shenlanxueyuan.com

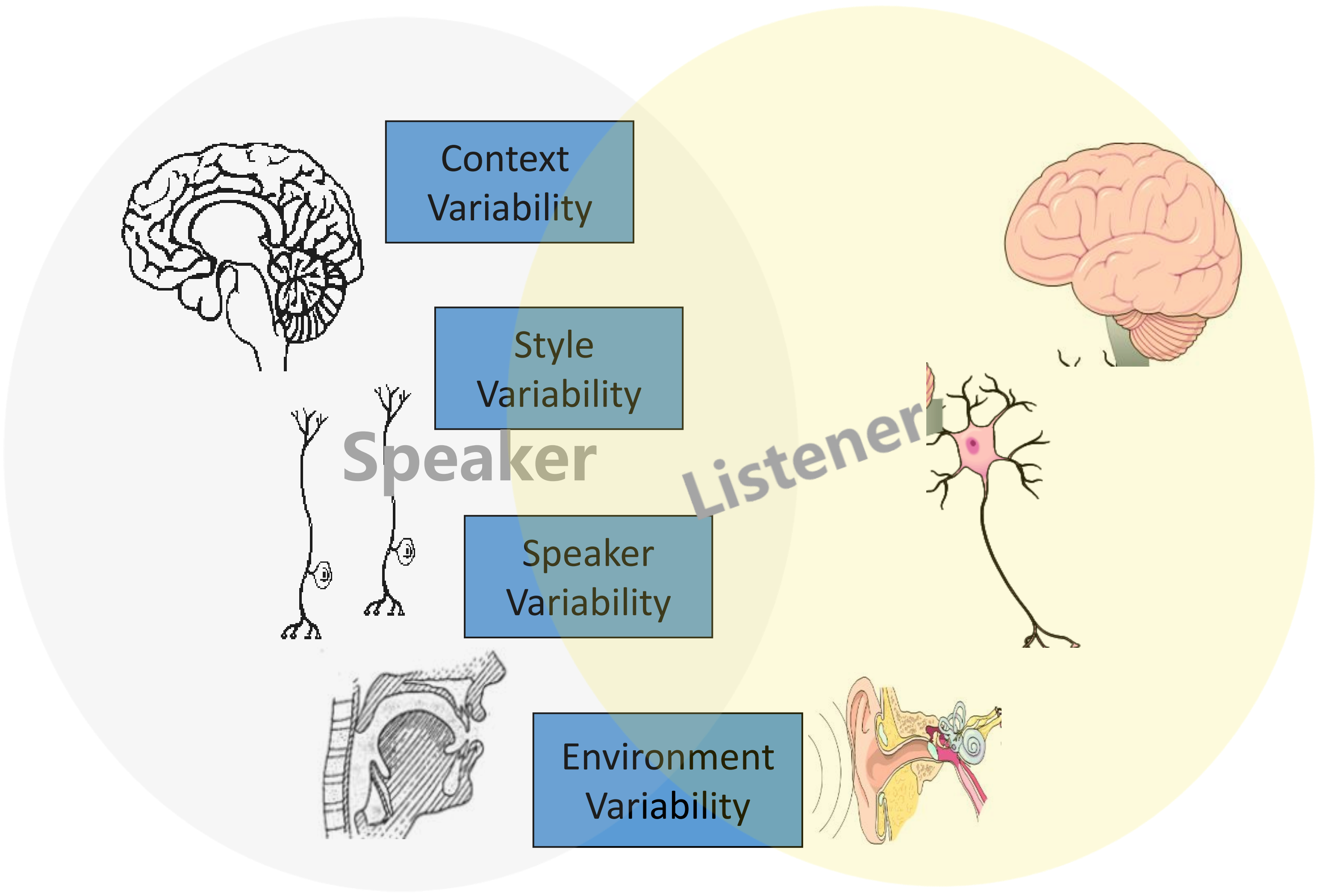
基于字的n-gram
基于词的n-gram:词的粒度选择,六七万词,效果比基于字的更好

- Language Model： 上TB的文本训练语言模型
  - N-gram for computing $p(W)$
  - Markov hypothesis

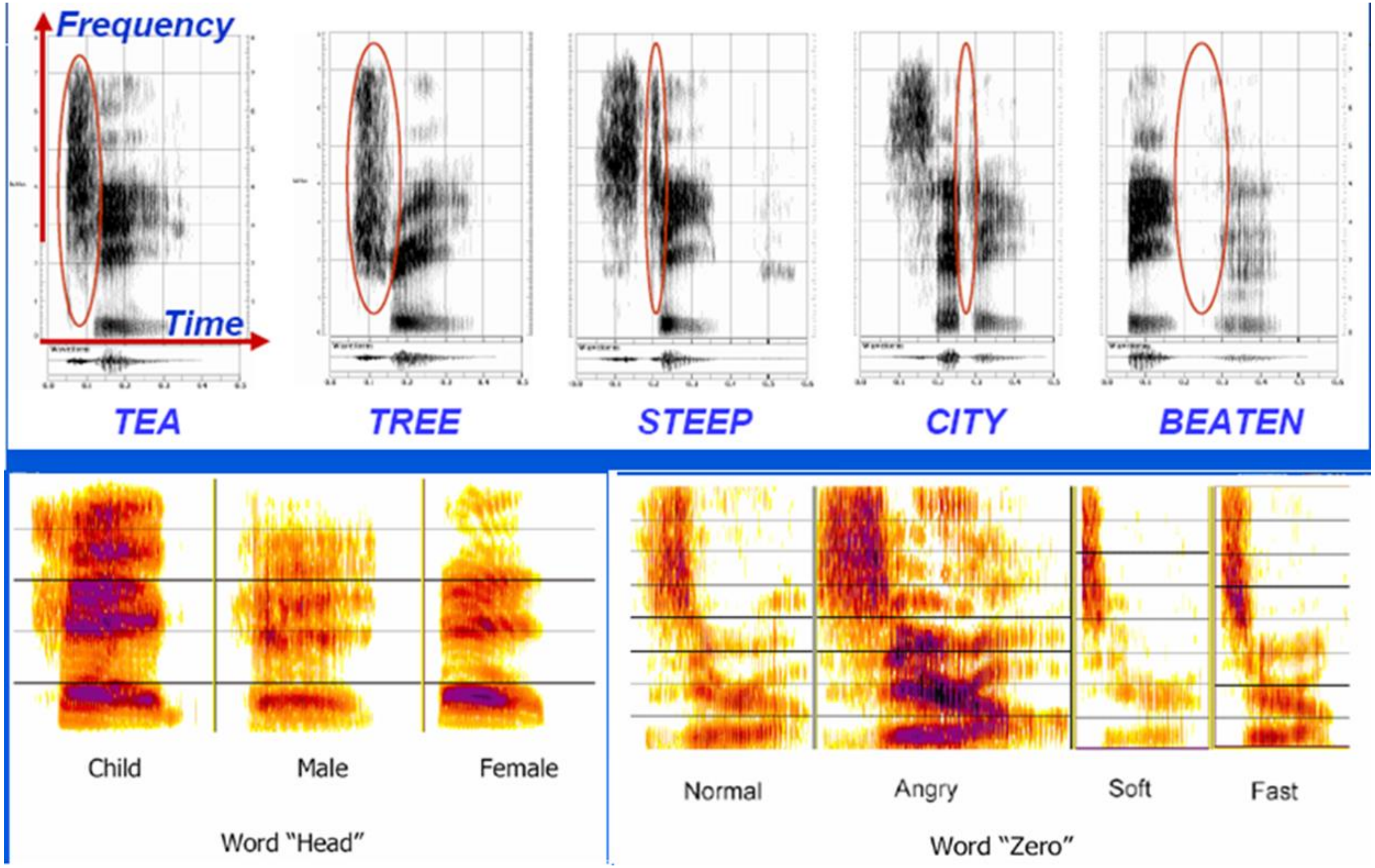
$$p(W) = p(w_1, w_2, \dots, w_m)$$
$$= p(w_1)p(w_2|w_1)p(w_3|w_1, w_2) \dots p(w_n|w_1, w_2, \dots, w_{m-1})$$
$$= \prod_{i=1}^{m} p(w_i|w_1, w_2, \dots, w_{i-1})$$
$$\approx \prod_{i=1}^{m} p\left(w_i \middle| w_{i-(n-1)}, w_{i-(n-2)}, \dots, w_{i-1}\right)$$

- Decoder:
  - Viterbi Algorithm: dynamic programming for combining all models
  - Usually using WFST (Weighted Finite State Transducers) 静态解码网络

Acoustic Models: Variability



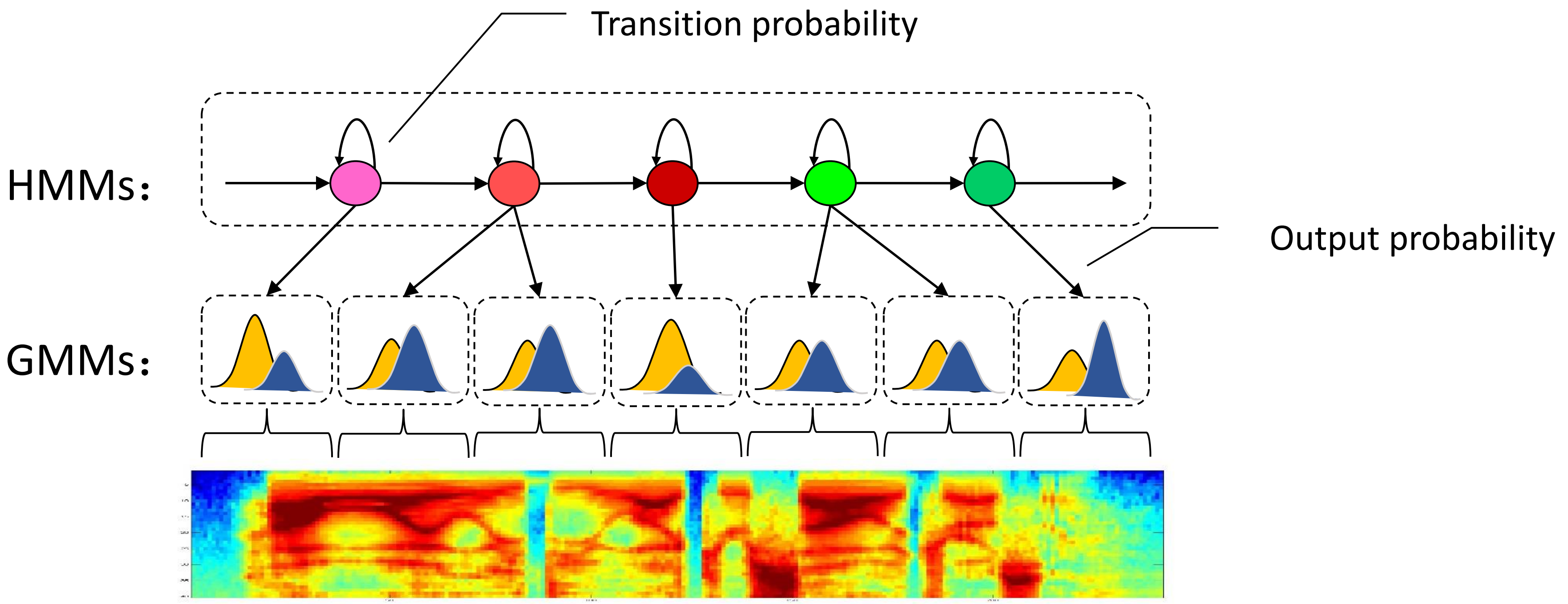
Acoustic Models: Variability



Acoustic Models: GMM-HMM

深蓝学院
shenlanxueyuan.com

- GMM-HMM based Acoustic Model
  - Gaussians for computing p(x|q) as the outputs probability in HMM
  - Markov models the context variability and transitions in acoustic



Acoustic Models: GMM-HMM

- Acoustic model: mapping the speech feature into acoustic unit
- The choice of acoustic modeling units
  - Sentence, phrase, word, character, syllable, initial-final(for Mandarin), phone
  - Selection criteria: the unit should be
    - ***accurate***, to represent the acoustic realization that appears in different contexts
    - ***trainable***. We should have enough data to estimate the parameters of the unit
    - ***generalizable***, so that any new word can be derived from a predefined unit inventory for task-independent speech recognition.

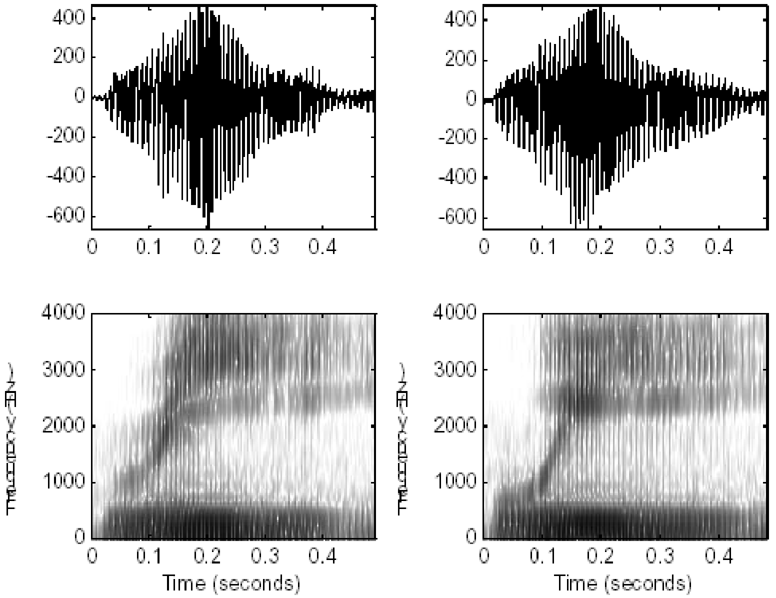
建模单元较小，复用性较高，训练样本会较多，准确性会相对差一点(没有上下文)

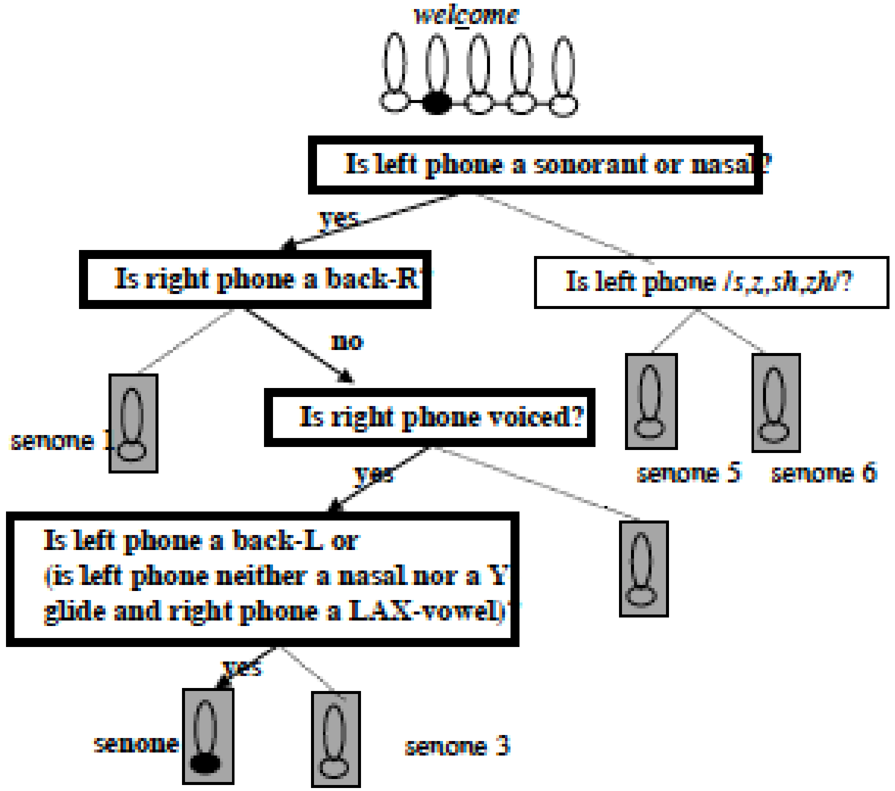
generalizable，trainable ***accurate***

phone syllable character word phase, sentence

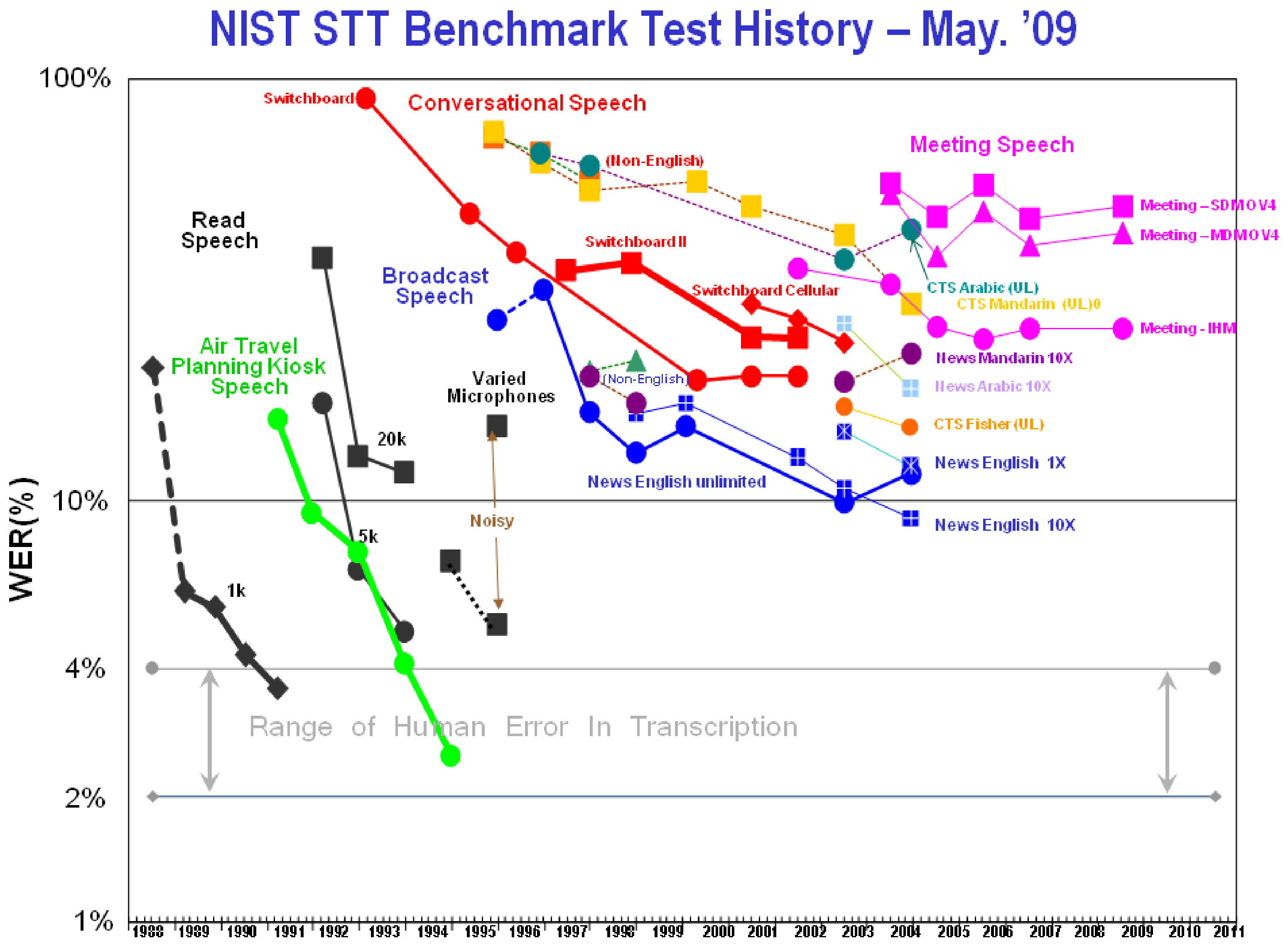
Acoustic Models: GMM-HMM

- Context dependency
  - In GMM-HMM, the triphone is always used:“ae-p+s”
  - Context information in triphone
  - Clustered Acoustic-Phonetic Units: Seno
    - Decision-tree based clustering





The performance benchmark

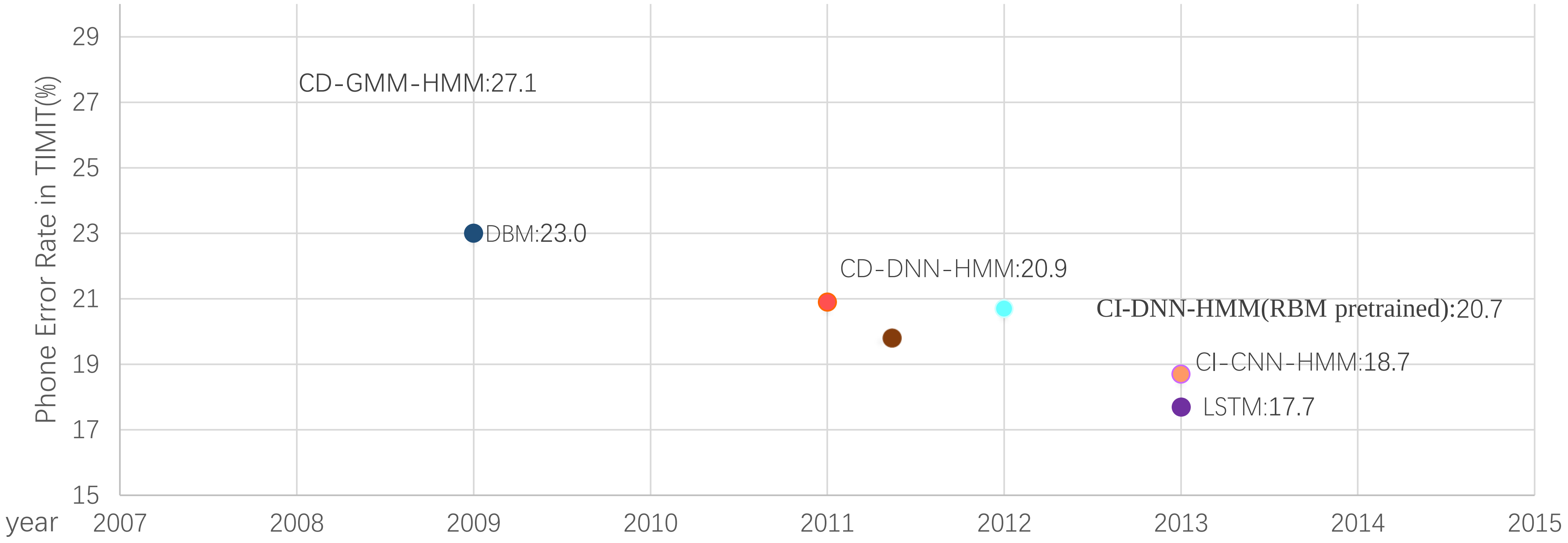


Outline

- Speech recognition: classic methods
- Speech recognition: DNN-HMM approaches
- Speech recognition: end-to-end approaches

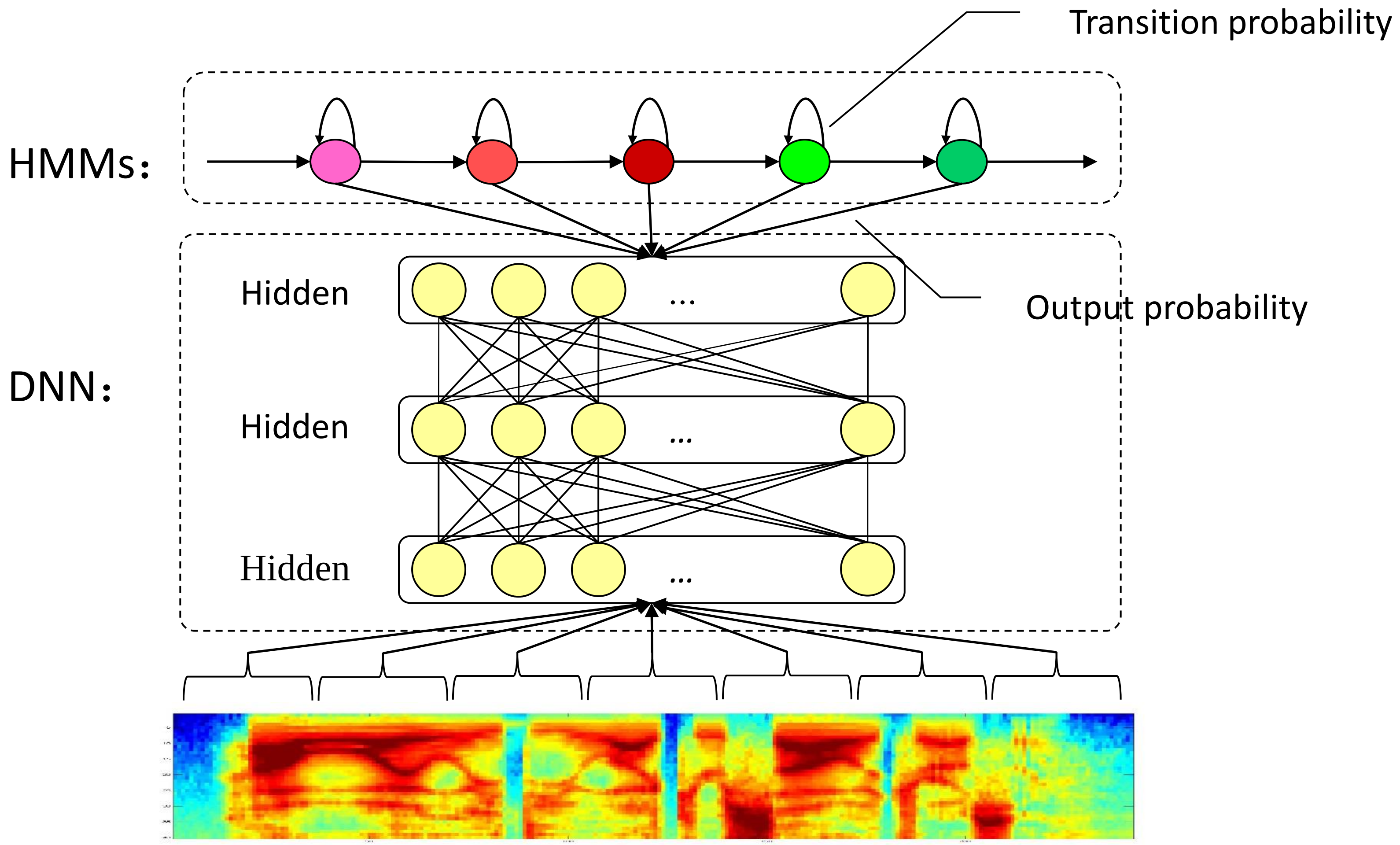
Speech recognition: deep learning approaches

- The introduce of DNN in speech recognition



Acoustic Models: DNN-HMM

- DNN replace GMM: still using HMM



Acoustic Models: DNN-HMM

- DNN replace GMM: still using HMM
  - DNN output the posterior probability

$$y_{s_i}(t) = p(q_t = s_i | x_t)$$

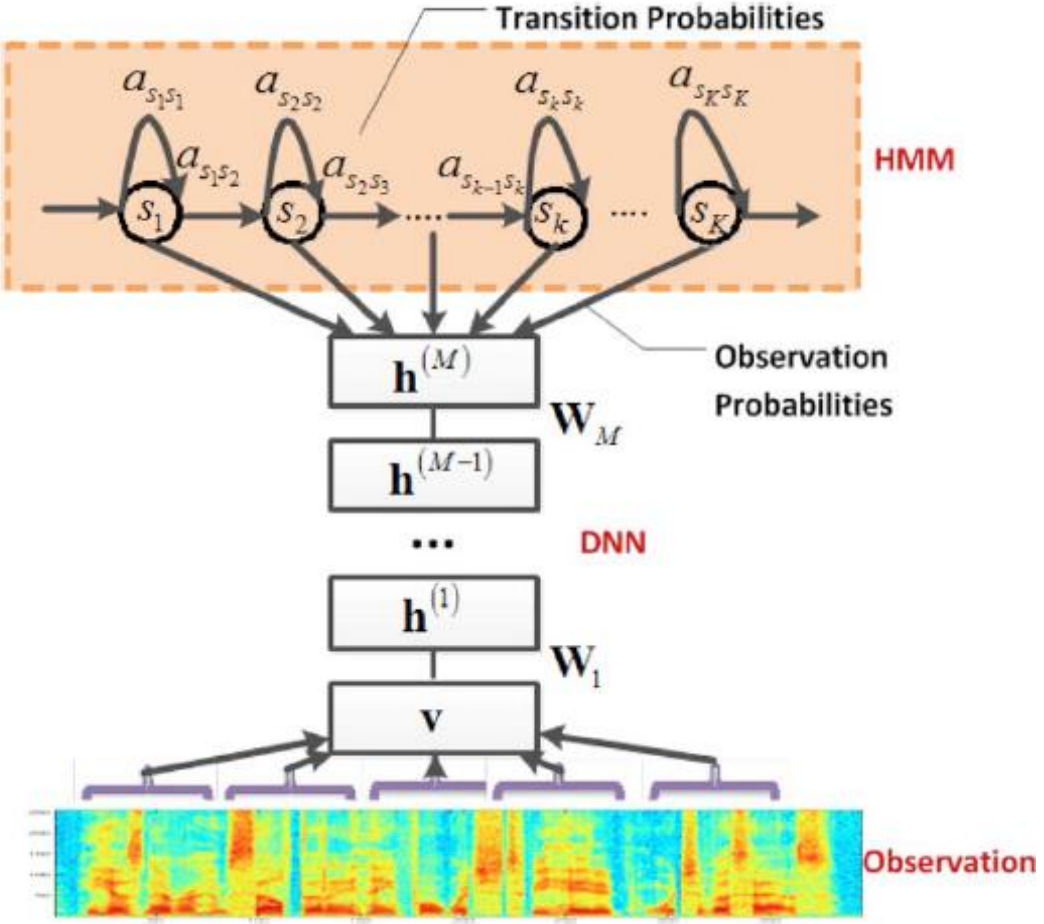
  - Using a pseudo likelihood in the HMM framework

$$p(x_t | q_t = s_i) = \frac{p(q_t = s_i | x_t) p(x_t)}{p(s_i)} \cong \frac{y_{s_i}(t)}{p(s_i)}$$

Acoustic Models: DNN-HMM

- Some references:

[1] G Dahl, D Yu, L Deng, A Acero. Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition. Audio, Speech, and Language Processing, IEEE Transactions on 20 (1), 30 - 42

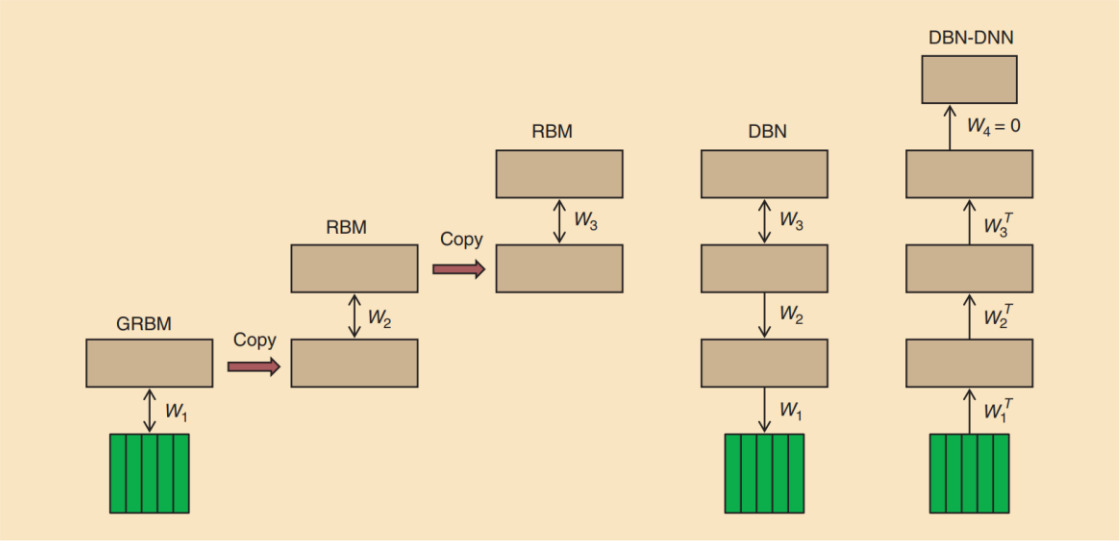


Acoustic Models: DNN-HMM

- Some references:

[1] G Dahl, D Yu, L Deng, A Acero. Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition. Audio, Speech, and Language Processing, IEEE Transactions on 20 (1), 30 – 42

[2] G. Hinton, L. Deng, D. Yu, GE. Dahl, A. Mohamed, and et.al, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine 29 (6), 82-97.



[FIG1] The sequence of operations used to create a DBN with three hidden layers and to convert it to a pretrained DBN-DNN. First, a GRBM is trained to model a window of frames of real-valued acoustic coefficients. Then the states of the binary hidden units of the GRBM are used as data for training an RBM. This is repeated to create as many hidden layers as desired. Then the stack of RBMs is converted to a single generative model, a DBN, by replacing the undirected connections of the lower level RBMs by top-down, directed connections. Finally, a pretrained DBN-DNN is created by adding a "softmax" output layer that contains one unit for each possible state of each HMM. The DBN-DNN is then discriminatively trained to predict the HMM state corresponding to the central frame of the input window in a forced alignment.

DNN-HMM ASR

- The input feature:
  - Trying to remove the hand-crafted features: MFCC -> FBANK
  - Maybe: waveform
- Various neural network structures
  - Feedforward, Convolutions, Recurrent

Acoustic Models: DNN-HMM

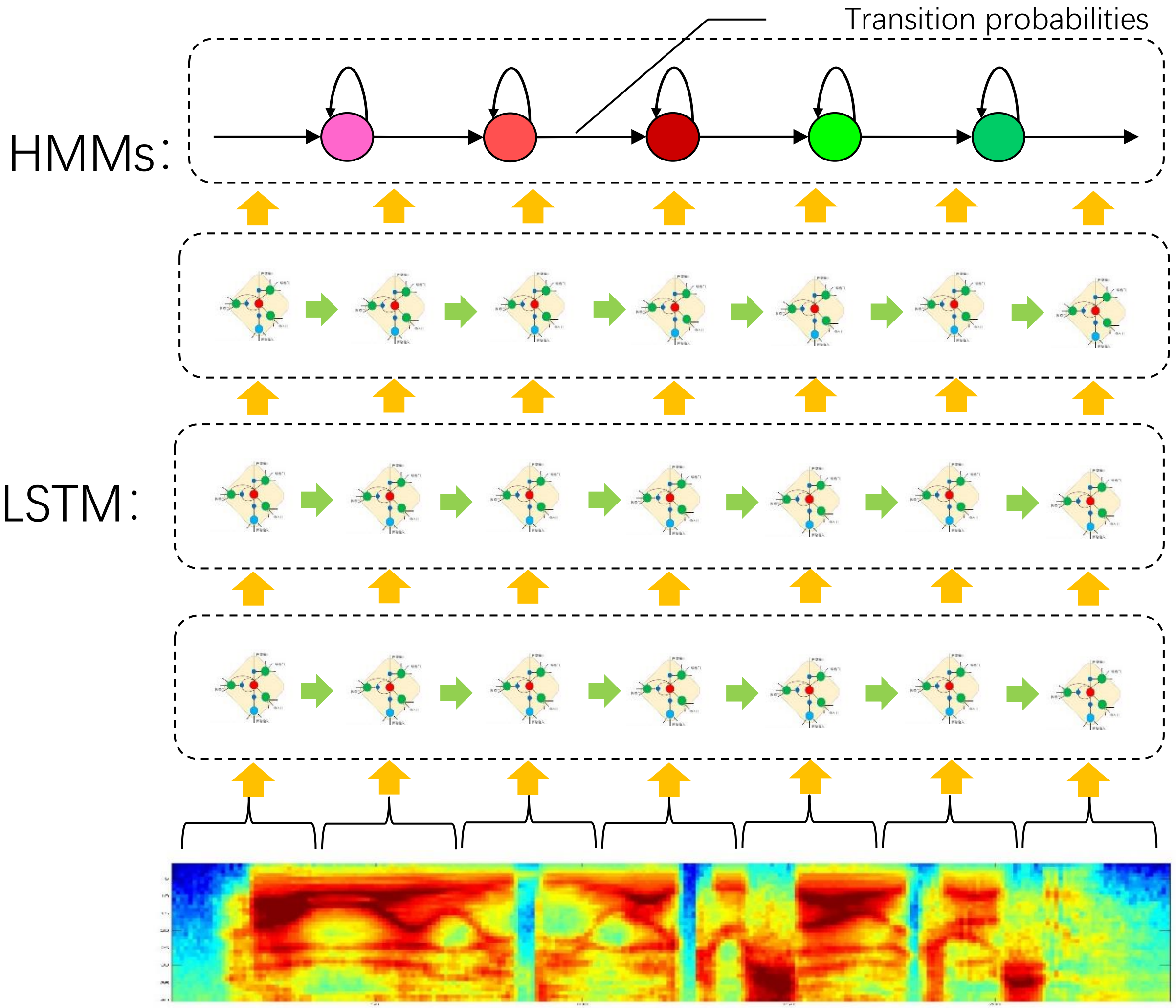
深蓝学院
shenlanxueyuan.com

- Some references:

[1] G Dahl, D Yu, L Deng, A Acero. Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition. Audio, Speech, and Language Processing, IEEE Transactions on 20 (1), 30 – 42

[2] G. Hinton, L. Deng, D. Yu, GE. Dahl, A. Mohamed, and et.al, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine 29 (6), 82-97.

[3] A. Graves, A. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks. ICASSP 2013.



Acoustic Models: DNN-HMM

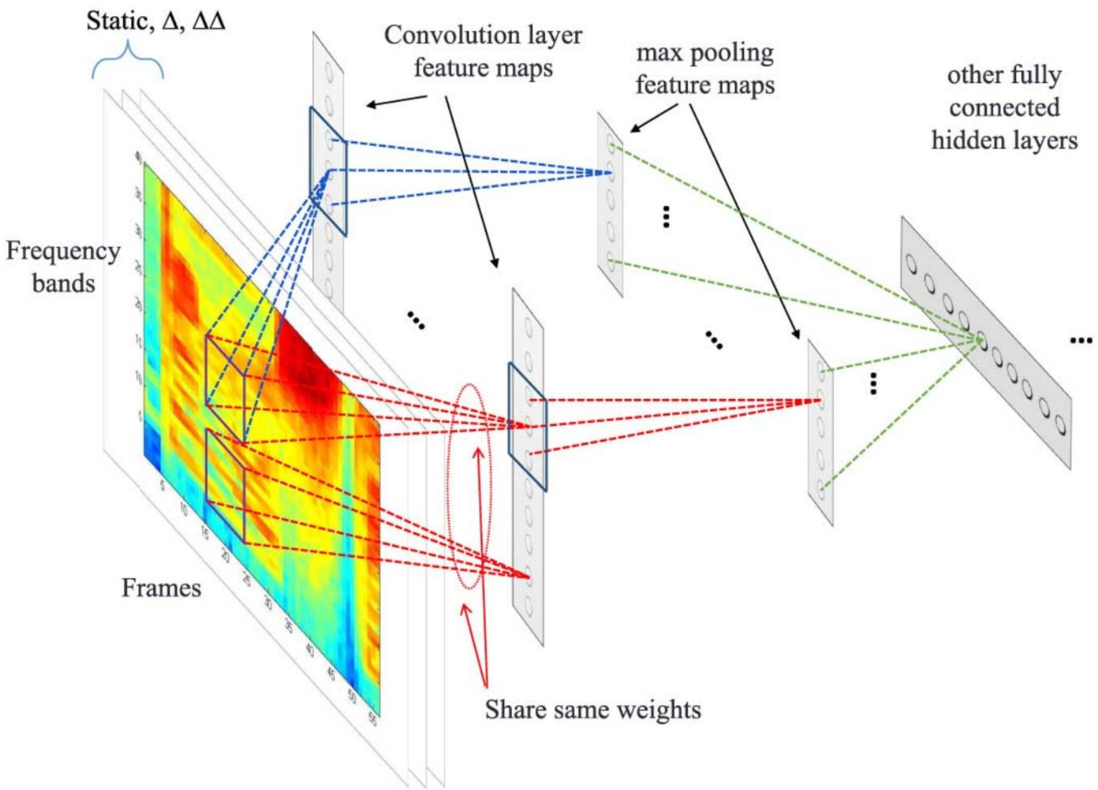
- Some references:

[1] G Dahl, D Yu, L Deng, A Acero. Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition. Audio, Speech, and Language Processing, IEEE Transactions on 20 (1), 30 – 42

[2] G. Hinton, L. Deng, D. Yu, GE. Dahl, A. Mohamed, and et.al, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine 29 (6), 82-97.

[3] A. Graves, A. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks. ICASSP 2013.

[4] O Abdel-Hamid, A Mohamed, H Jiang, L Deng, G Penn, D Yu. Convolutional neural networks for speech recognition. IEEE/ACM Transactions on Audio, Speech and Language Processing, 22(10), 1533-1545.



Outline

- Speech recognition: classic methods
- Speech recognition: DNN-HMM approaches
- Speech recognition: end-to-end approaches

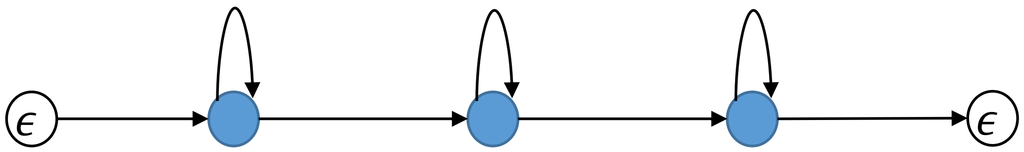
The rise of end-to-end learning

- The rise of end-to-end learning

- Replacing pipeline systems with a single learning algorithm
  - Go directly from the input to the desired output

CTC based speech recognition

- Hybrid: LSTM-HMM



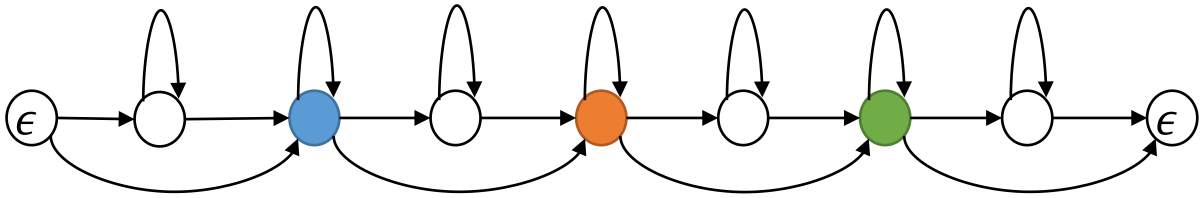
- Connectionist Temporal Classification (CTC)
  - Introduce the blank label

a b c = blank a a b blank c c c blank

= blank a blank b b blank c blank

= blank a a blank b b c c blank

= …



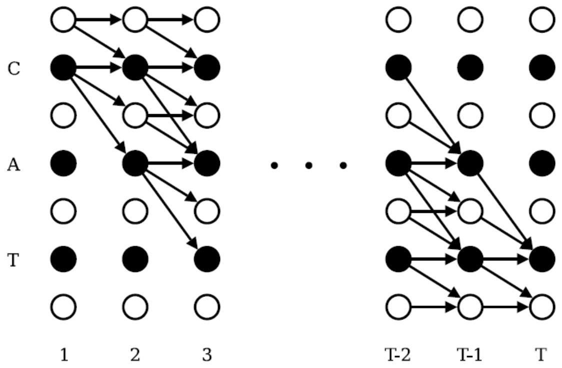
CTC

深蓝学院
shenlanxueyuan.com

- Objective function of CTC is defined as the negative log probability of correctly labelling the entire training set:

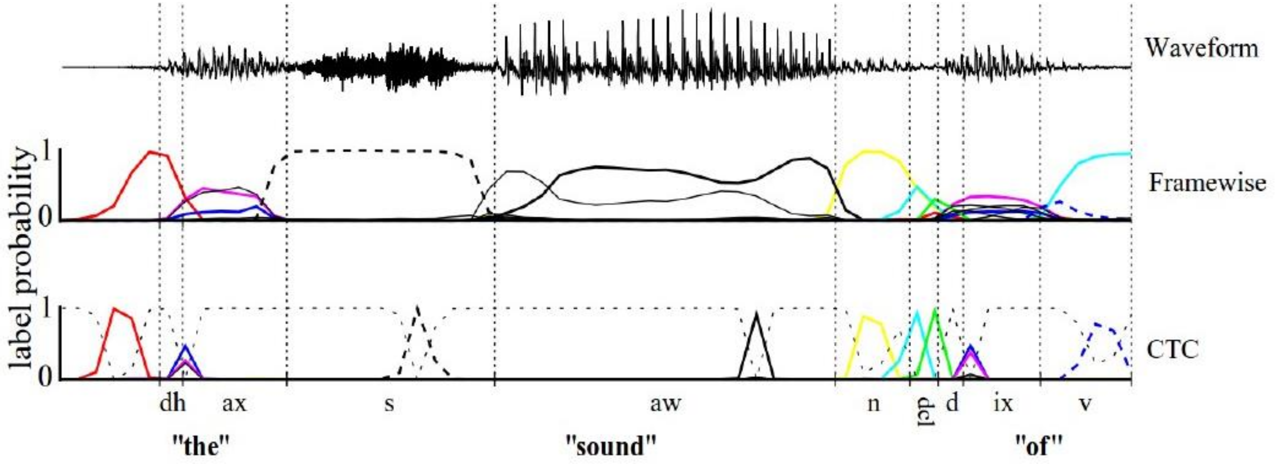
$$O_{ctc} = -\ln\left(\prod_{(\mathbf{x},\mathbf{z})\in S} p(\mathbf{z}|\mathbf{x})\right) = -\sum_{(\mathbf{x},\mathbf{z})\in S} \ln\big(p(\mathbf{z}|\mathbf{x})\big)$$

- Forward and backward variables used for accelerated the calculating the objective function
  - Similar to the forward-backward algorithm of DNN-HMM, but using different topology



CTC vs. HMM

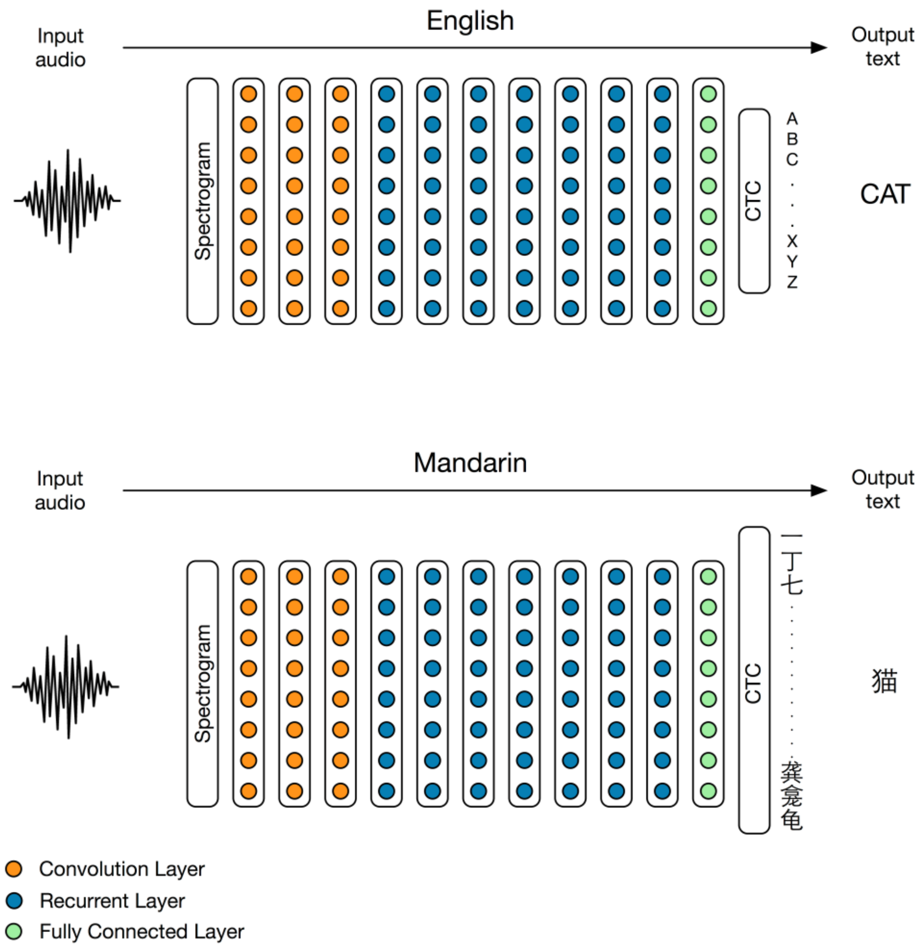
- Map input feat to output symbol（maybe blank）
  - Do not need pre-alignment
  - Conditional independent assumption
  - Possible output peak delay
- Main difference
  - Topology



CTC vs. HMM

- Modeling units in CTC ASR:
  - Some systems use One-state tied tri-phone
  - Trying to perform end-to-end
    - For English: using Grapheme,
    - For Mandarin: Characters or Syllables
- Input features in CTC ASR:
  - Still using FBank
  - But usually 3-fold down-sampling, so 30 ms each frame

DEEPSPEECH system



MIT Technology Review

Log in / Register Search

Topics+ Top Stories Magazine Business Reports More+

10 Breakthrough Technologies The List+ Years+

Conversational Interfaces

Powerful speech technology from China's leading Internet company makes it much easier to use a smartphone.

Availability: now

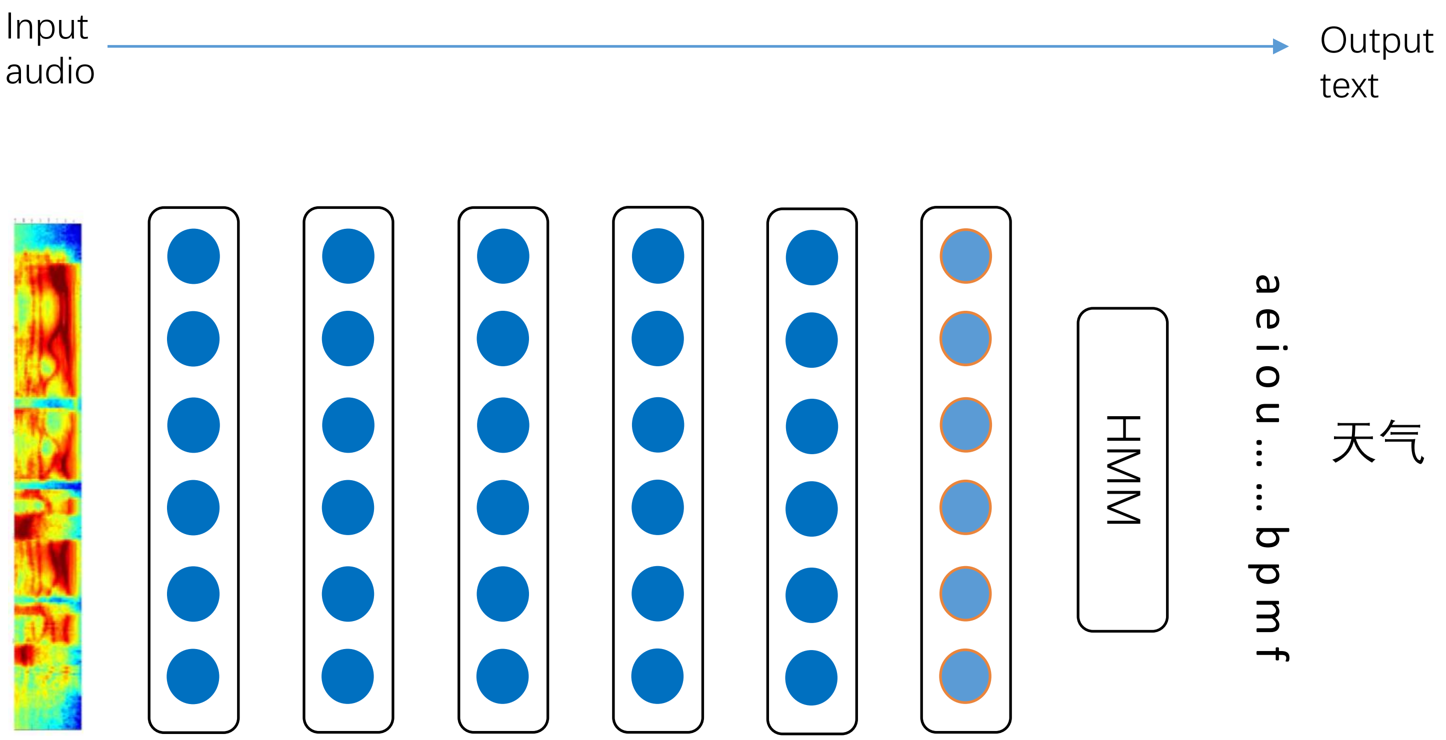
by Will Knight

Speech recognition: from GMM to end-to-end

- GMM-HMM ASR
  - Acoustic Model: GMM-HMM
    - Trained using acoustic training set
    - Using GMM to model the distributions of HMM states
  - Dictionary: Word->Phoneme
  - Language Model: N-gram
    - Trained using text corpus

L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 1989

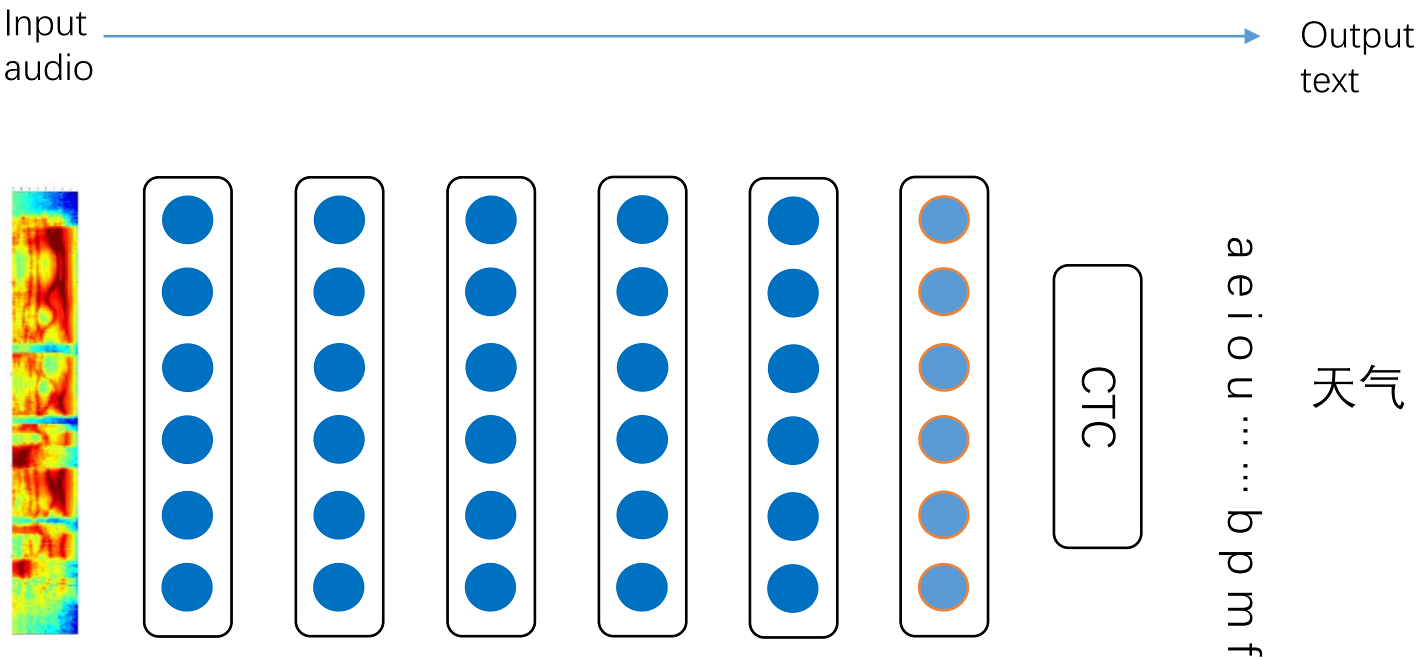
Speech recognition: from GMM to end-to-end



- DNN-HMM ASR
  - Acoustic Model: DNN-HMM
    - Trained using acoustic training set
    - Using deep learning
    - Output HMM states
  - Dictionary: Word->Phoneme
  - Language Model: N-gram
    - Trained using text corpus

George Dahl, Dong Yu, Li Deng, Alex Acero, Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing. 2012

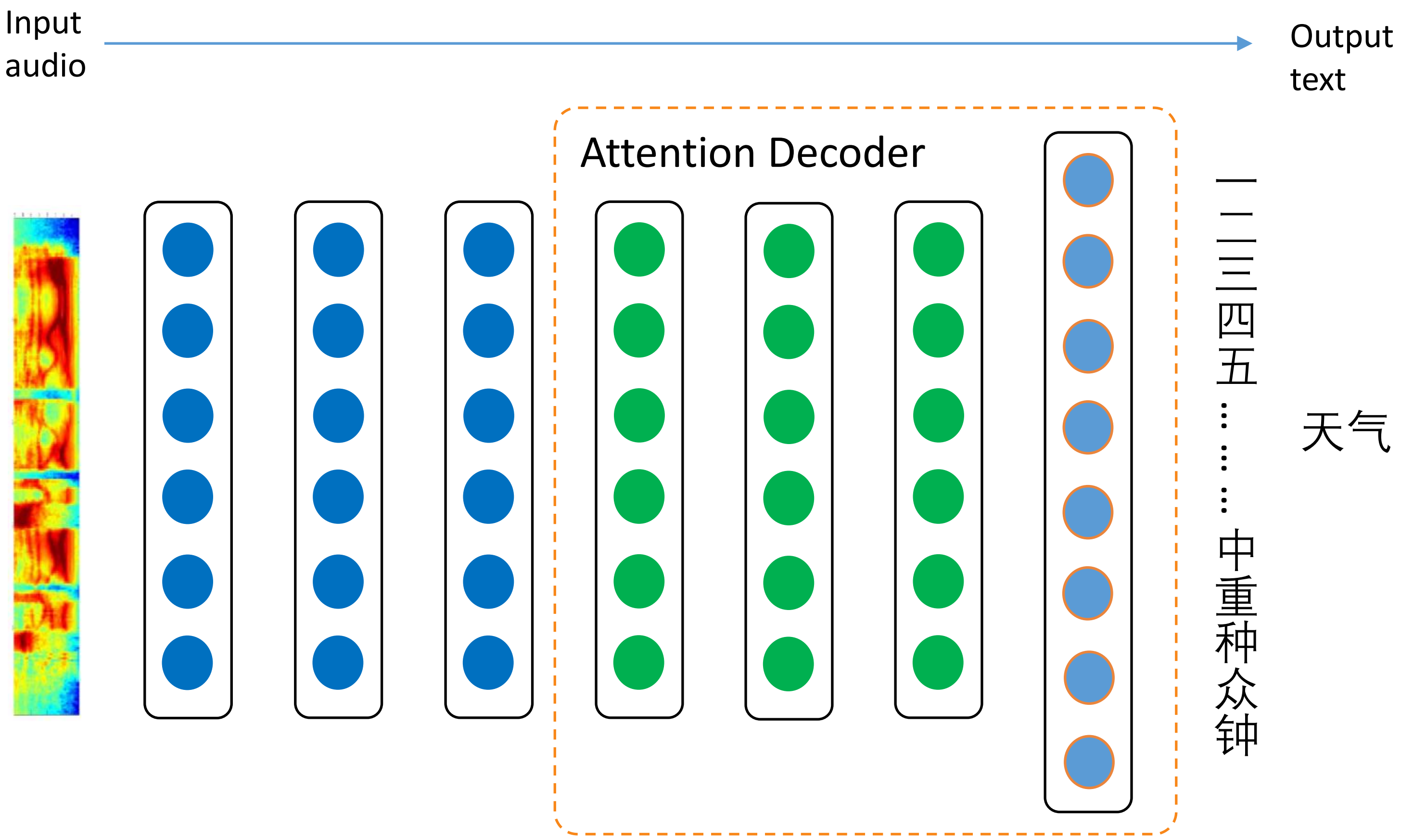
Speech recognition: from GMM to end-to-end



- CTC ASR
  - Acoustic Model: RNN-CTC
    - Trained using acoustic training set
    - Using deep sequence learning
    - Output grapheme for English, syllables for Mandarin
  - Dictionary: Word->Phoneme
  - Language Model: N-gram
    - Trained using text corpus

H. Sak, A. Senior, K. Rao, F. Beaufays, Fast and accurate recurrent neural network acoustic models for speech recognition. arXiv:1507.06947, 2015

Speech recognition: from GMM to end-to-end



- Attentional ASR
  - Acoustic Model: RNN-Attention
    - Trained using acoustic training set
    - Using deep sequence learning
    - Output characters / phonemes
  - ~~Dictionary~~
  - ~~Language Model~~

Attentional ASR

- ~~Dictionary~~
  - The modeling units for Mandarin Chinese ASR

Word	Character	Syllable	Initial-final/phones
北京	北 京	bei jing	b ei j ing

  - Characters are usually selected as the basic modeling units
- ~~Language Model~~
  - How to benefit from the large text corpus without N-gram ?
  - We pre-train RNN-LM and then merged into acoustic neural network

End-to-end speech recognition

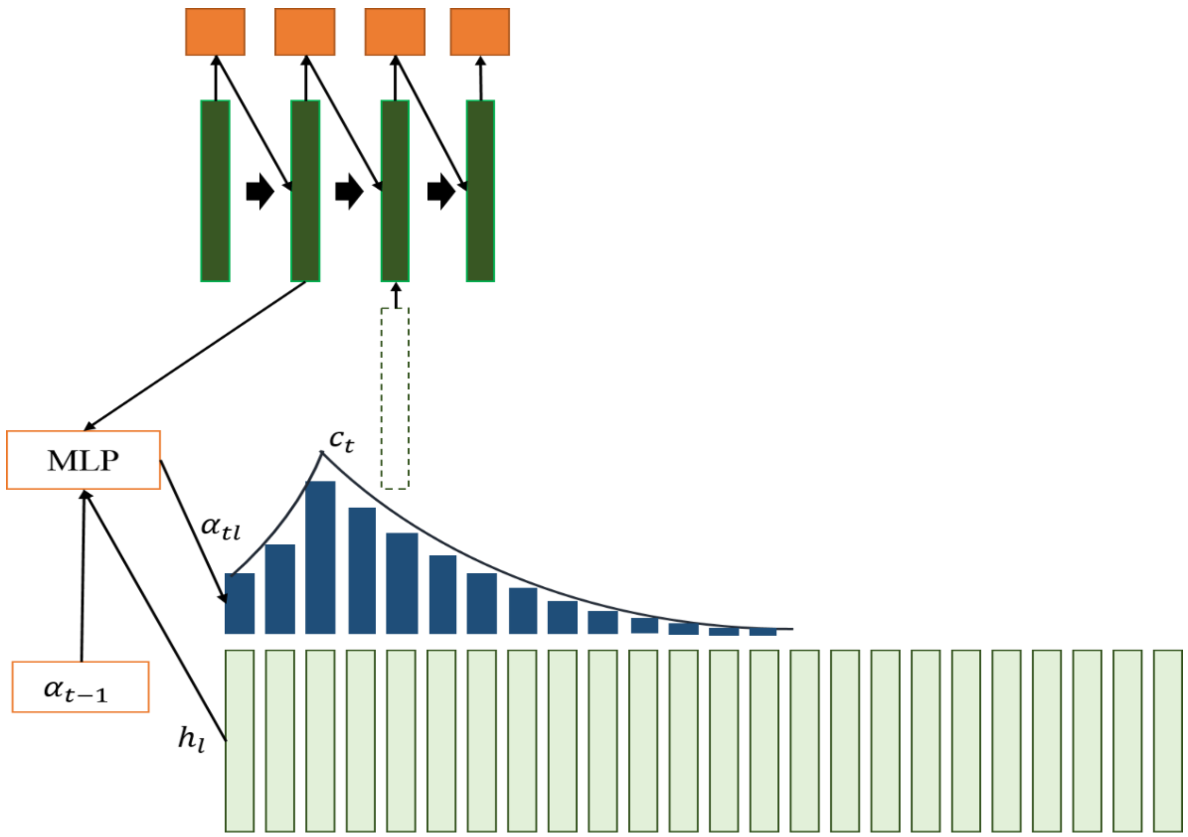
深蓝学院 shenlanxueyuan.com

- End-to-end is a relative concept

	phoneme	syllable/character
DNN-HMM	We need decision-tree based state clustering, dictionary, language model	
RNN-CTC	We need dictionary, language model, (If we use the cd-phone as modeling units, we still need decision-tree based state clustering)	The N-gram based language models would improve the performance
RNN-Attention		We do not need extra models

Attentional ASR

- Sequence-to-sequence model from translation



First Attention in Speech

深蓝学院
shenlanxueyuan.com

- Same structure with Bahdanau's neural translation model

End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results

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

Dzmitry Bahdanau
Jacobs University Bremen, Germany

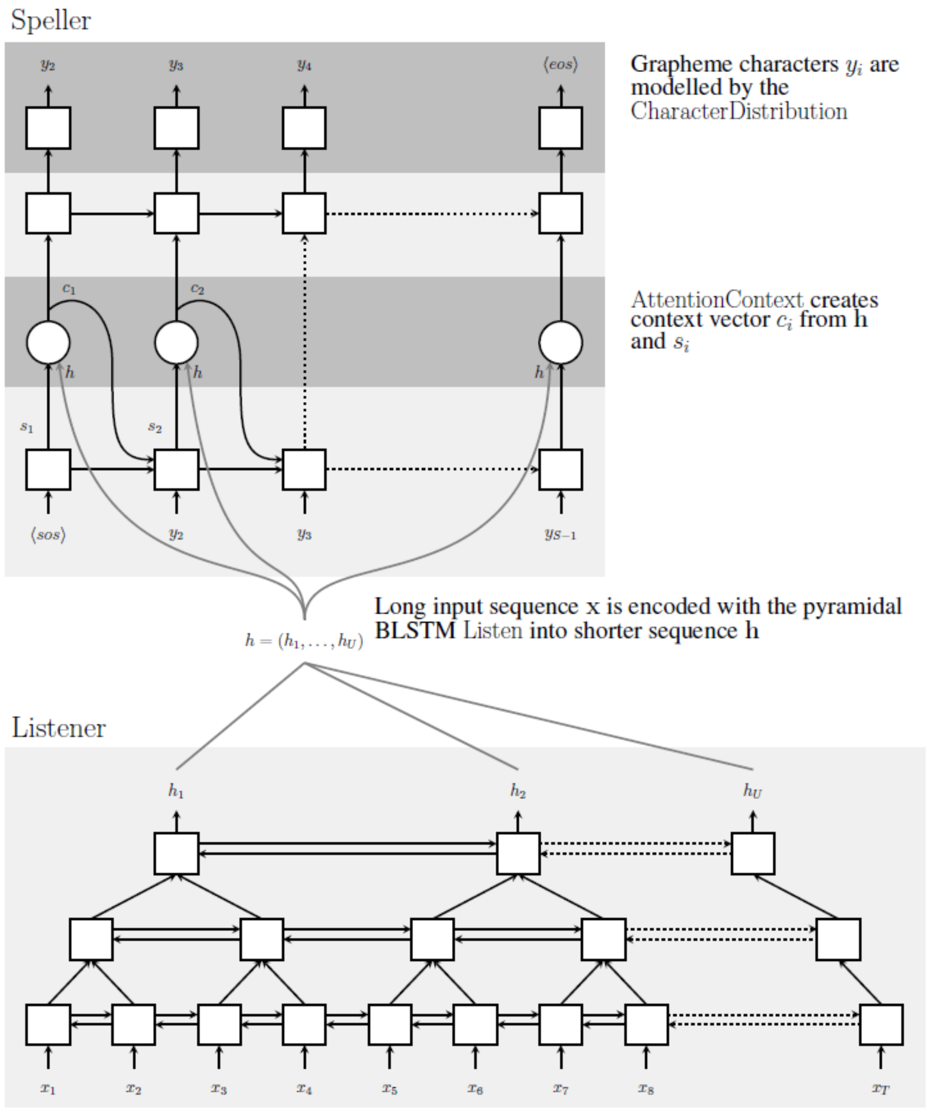
Kyunghyun Cho
Université de Montréal

Yoshua Bengio
Université de Montréal
CIFAR Senior Fellow

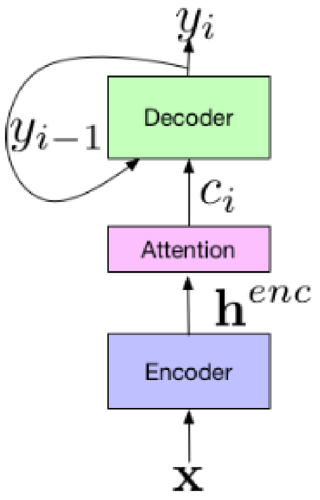
Abstract

We replace the Hidden Markov Model (HMM) which is traditionally used in in continuous speech recognition with a bi-directional recurrent neural network encoder coupled to a recurrent neural network decoder that directly emits a stream of phonemes. The alignment between the input and output sequences is established using an attention mechanism: the decoder emits each symbol based on a context created with a subset of input symbols selected by the attention mechanism. We report initial results demonstrating that this new approach achieves phoneme error rates that are comparable to the state-of-the-art HMM-based decoders, on the TIMIT dataset.

Listen-Attend-Spell



- Encoder
  - Listen, map the input feature sequence to embedding
- Decoder
  - Spell, map the embedding based on the attention information to the output symbols



Attention vs. CTC

- Advantages
  - There is no conditional independence assumptions
  - Joint learning of acoustic information and language information
  - Speech recognition system is more simple

- Disadvantages
  - Not easy to converge, We need more tricks to train attention model
  - Cannot be used for “streaming” speech recognition, during inference, the model can produce the first output token only after all input speech frames have been consumed.

Listen-Attend-Spell

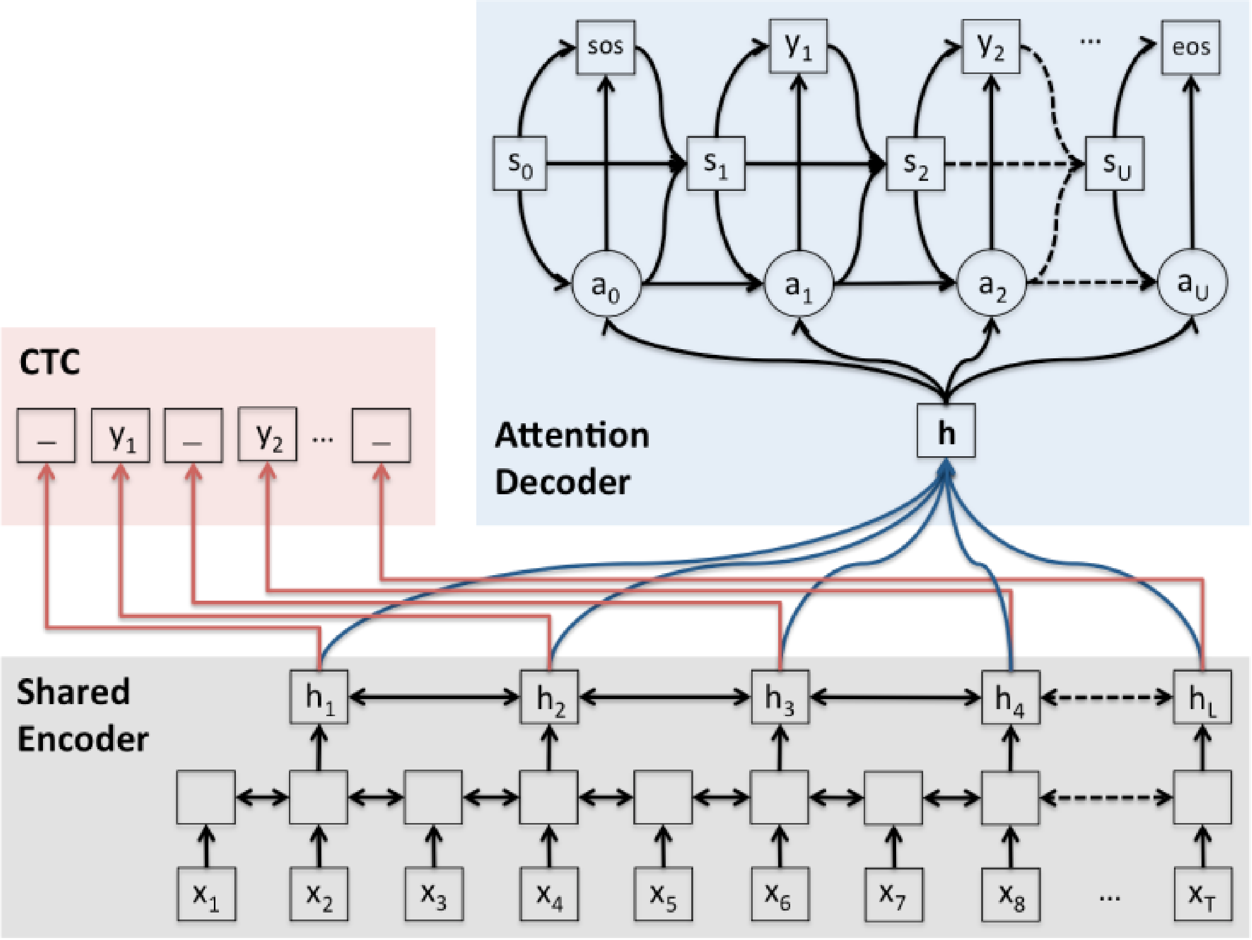
- Hard to train – many "tricks"
  - Schedule sampling
  - Label smoothing (2016)

$$q'(k|x) = (1 - \epsilon)\delta_{k,y} + \epsilon u(k)$$

Listen-Attend-Spell

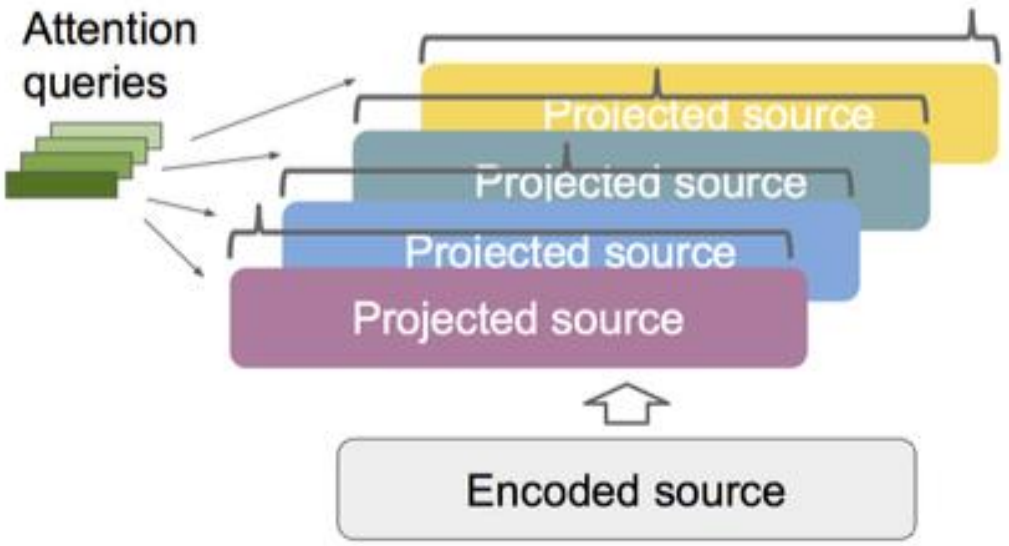
深蓝学院 shenlanxueyuan.com

- Hard to train – many “tricks”
  - Schedule sampling
  - Label smoothing (2016)
  - Multi-Task Learning (2017)
    - Joint CTC-attention based end-to-end framework
    - The shared encoder is trained by both CTC and attention model objectives simultaneously.



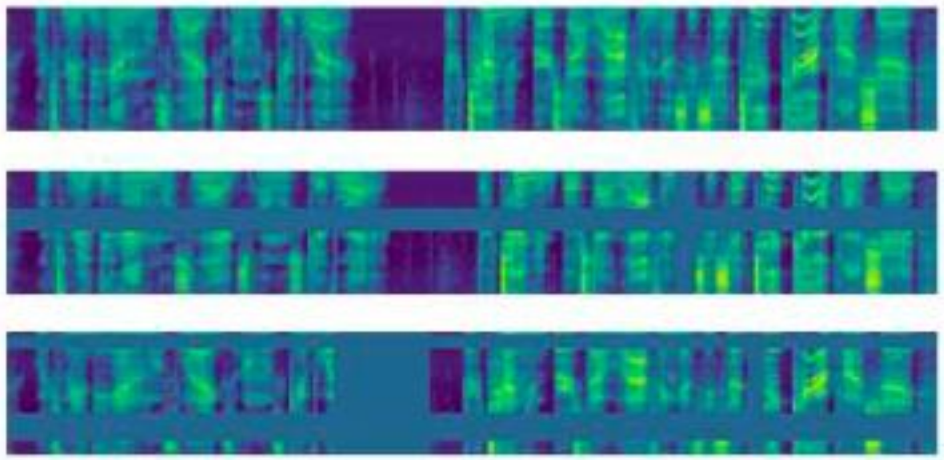
Listen-Attend-Spell

- Hard to train – many “tricks”
  - Schedule sampling
  - Label smoothing (2016)
  - Multi-Task Learning (2017)
  - Multi-headed Attention (2018)
    - Inspired by transformer
    - Replacing single head attention



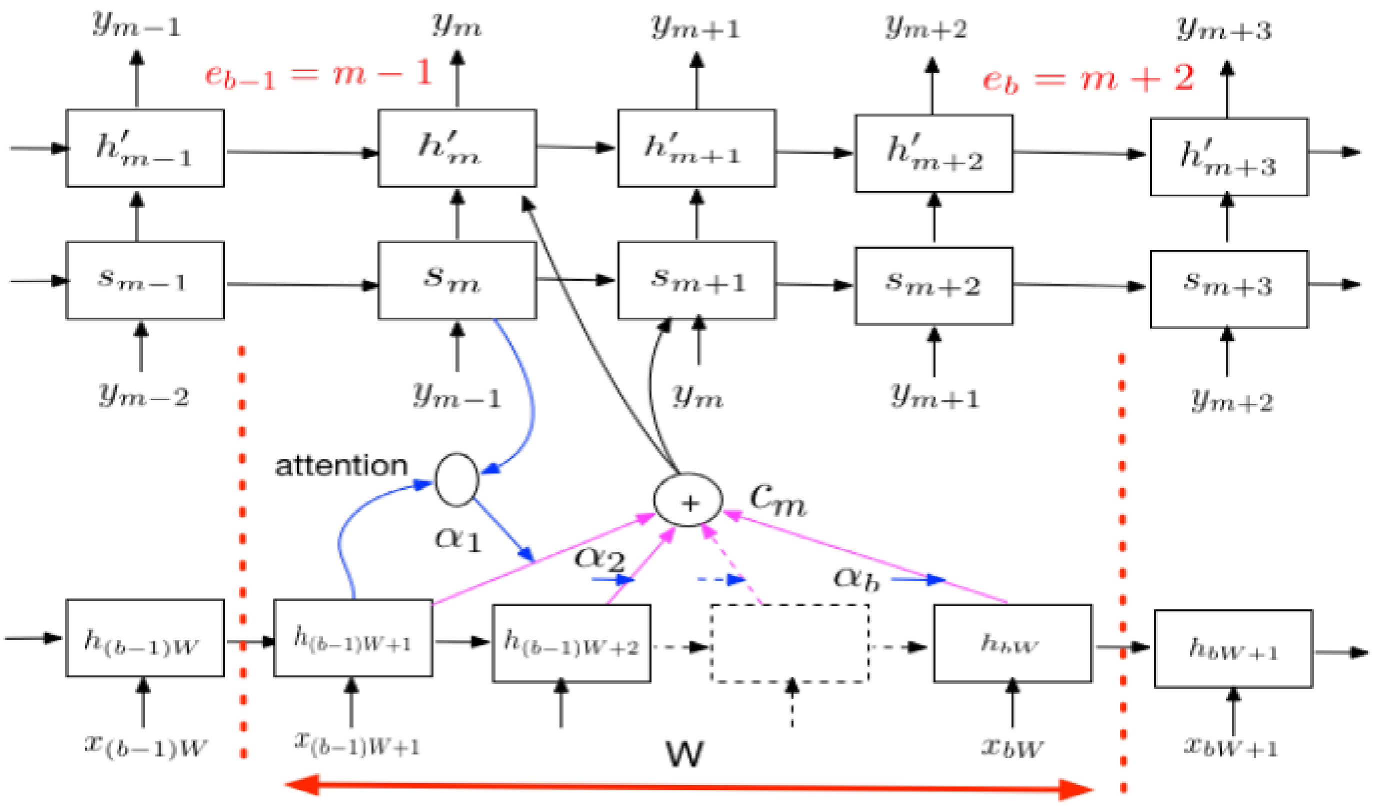
Listen-Attend-Spell

- Hard to train – many “tricks”
  - Schedule sampling
  - Label smoothing (2016)
  - Multi-Task Learning (2017)
  - Multi-headed Attention (2018)
  - SpecAugment (2019)
    - Data augmentation to LAS
    - Achieved sota results on Librispeech a SWBD



Online neural transducer

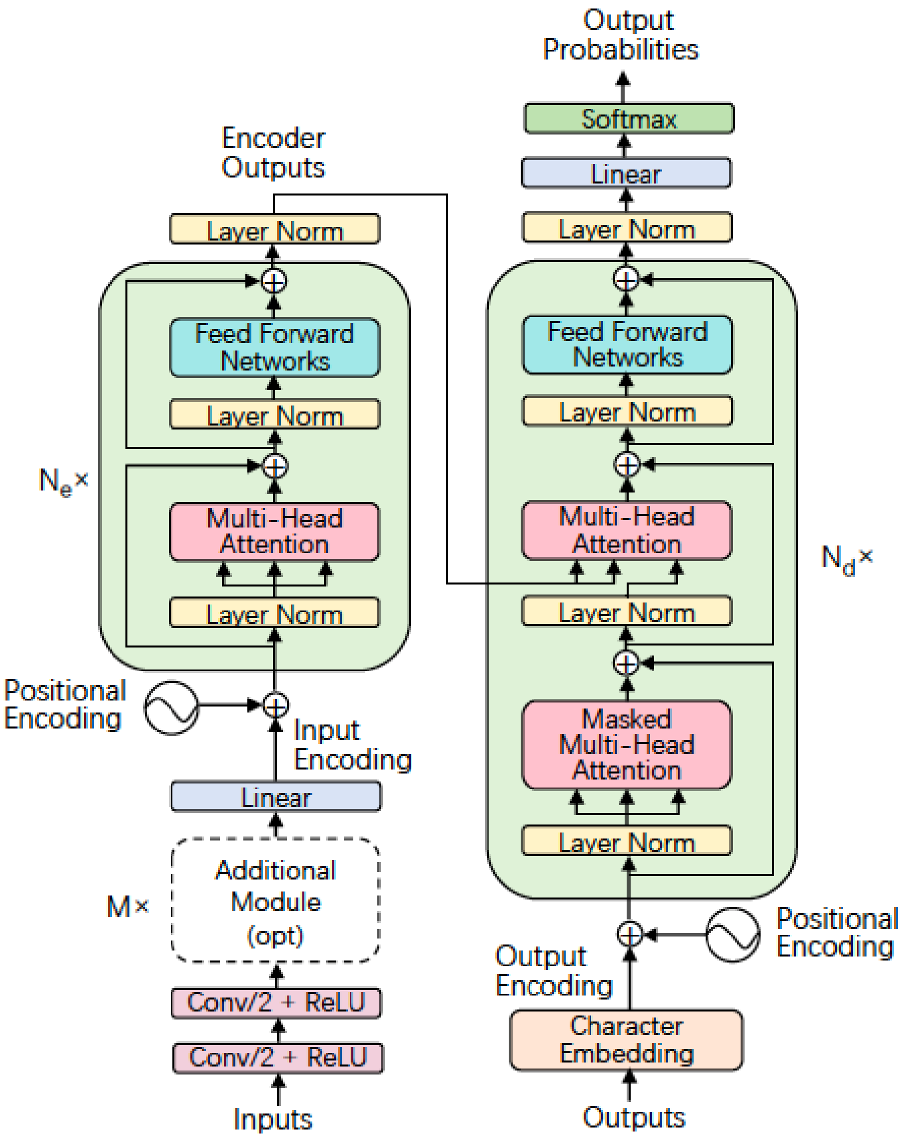
深蓝学院
shenlanxueyuan.com



- A limited sequence streaming attention-based model
- Consumes a fixed number of input frames (a chunk)
- Outputs a variable number of labels before it consumes the next chunk

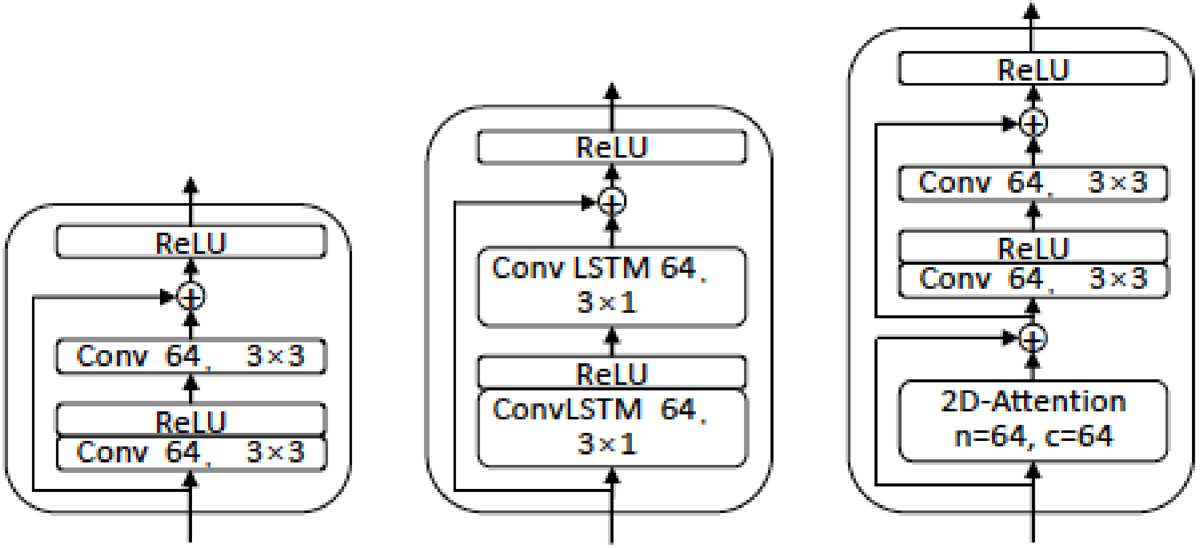
Speech-Transformer

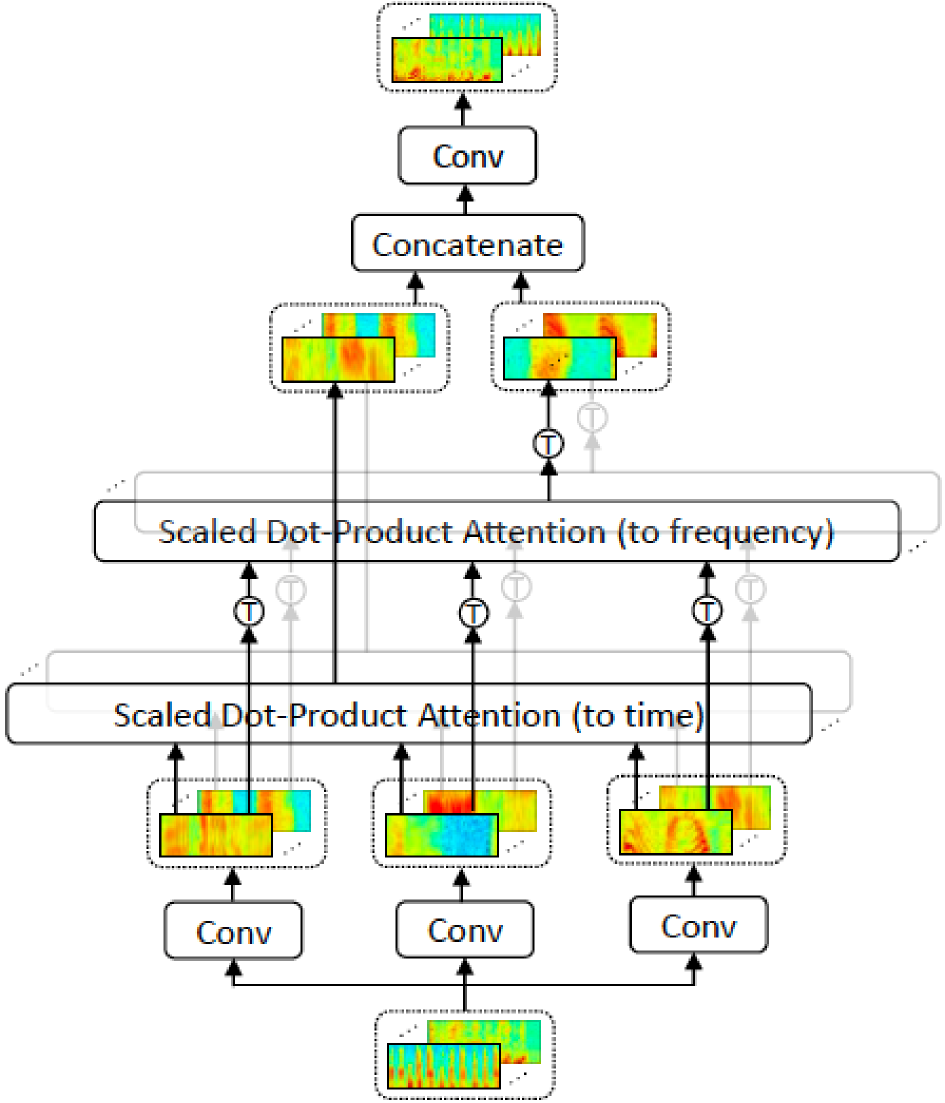
- Speech Transformer
  - Transformer applied to ASR
  - With Conv layers as inputs



Speech-Transformer

- Speech Transformer
  - Transformer applied to ASR
  - With Conv layers as inputs





Speech-Transformer

- Speech Transformer
  - Transformer applied to ASR
  - With Conv layers as inputs
- Time-restricted self-attention
  - Left & Right Contexts restricting the attention mechanism

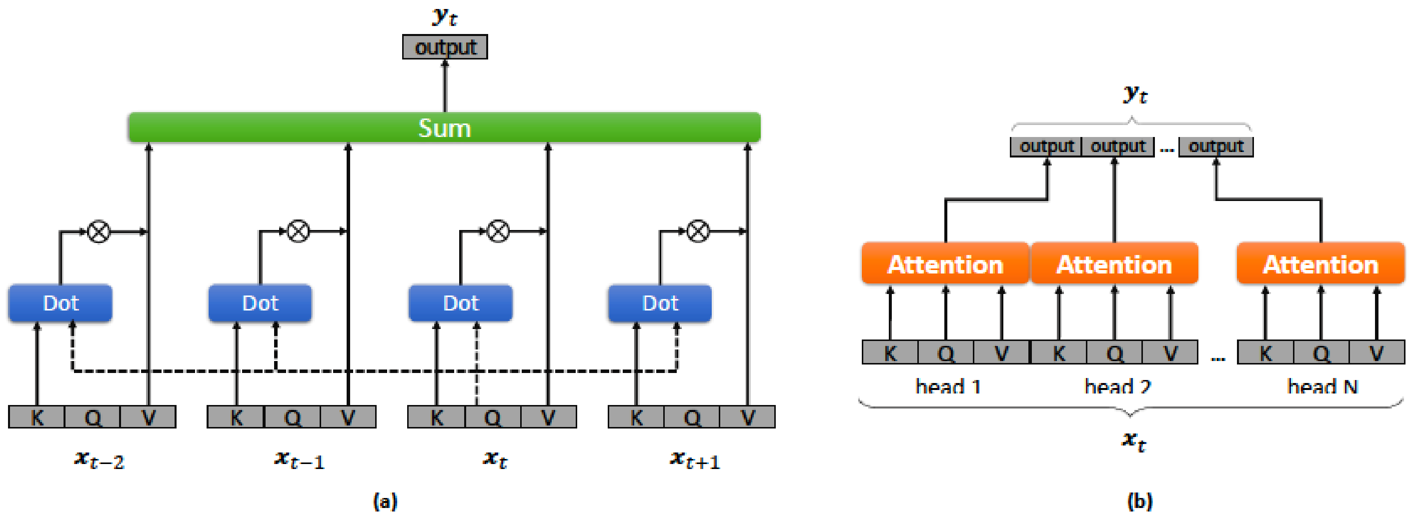


Fig. 2. (a) A single-head attention component. Left and right context sizes are 2 and 1 respectfully. For clarity, positional-encoding and the softmax (which is applied to the dot-products) are not shown. (b) A multi-head attention component using single-head attention blocks. K, Q, and V respectively mean key, query, and value.

Thanks!