# Summary of End-to-end Speech Recognition Researches


Zhehuai Chen

chenzhehuai@sjtu.edu.cn

SJTU Speech Lab

上海交通大學智能語音實驗室

# Outline

- End-to-end Modeling
  - CTC
  - LF-MMI
  - RNN-transducer
  - Sequence-to-sequence

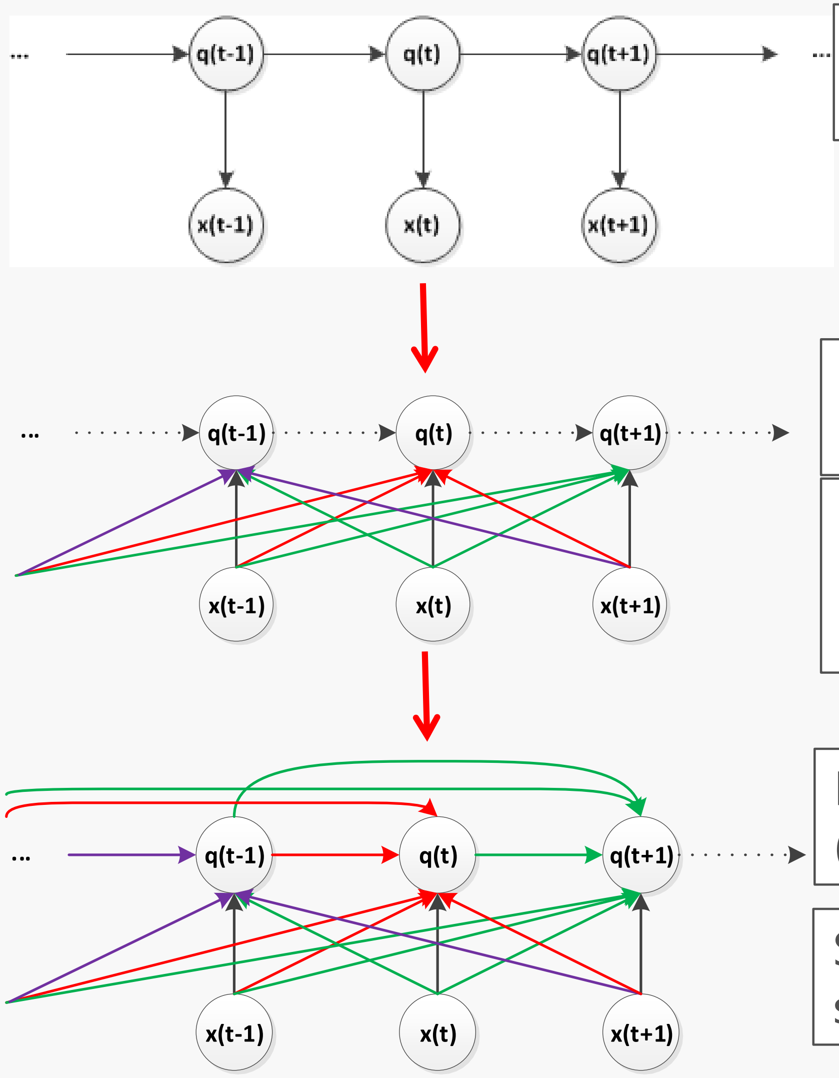
- End-to-end Inference
  - Phone level PSD
  - Word level PSD
  - Reducing WFST sizes

# Outline

- End-to-end Modeling
  - **CTC**
  - **LF-MMI**
  - **RNN-transducer**
  - **Sequence-to-sequence**

- End-to-end Inference
  - Phone level PSD
  - Word level PSD
  - Reducing WFST sizes

# A Brief Comparison




Hidden Markov Model (HMM)

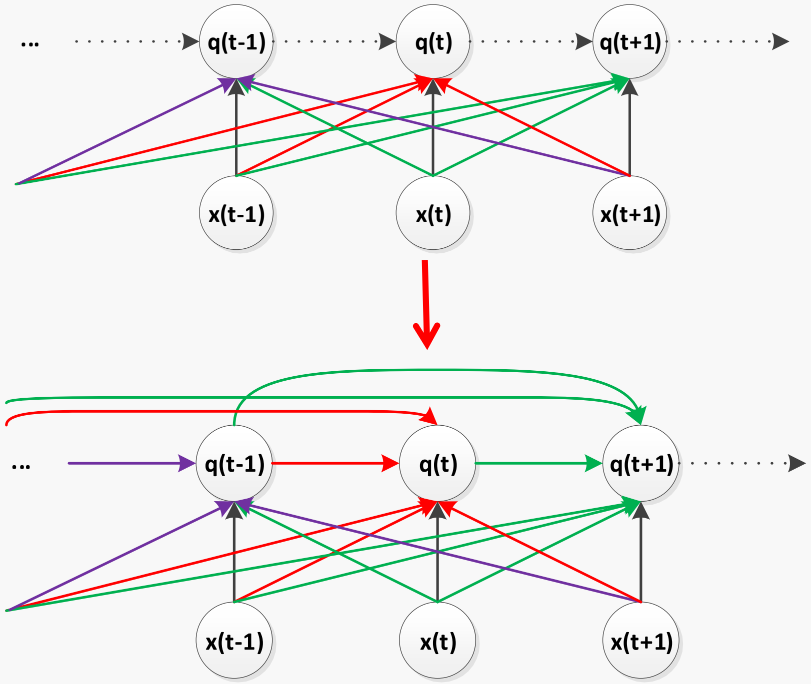
Connectionist Temporal Classification (CTC)

Lattice-free maximum mutual information (LF-MMI)

RNN transducer (RNN-T)

Sequence-to-sequence (S2S)

# A Brief Comparison

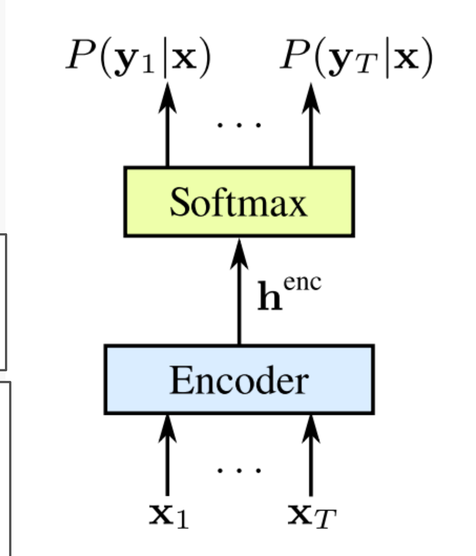


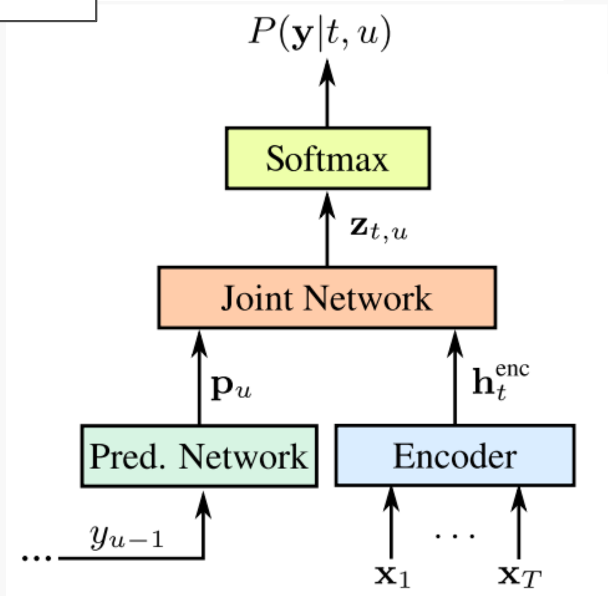

Connectionist Temporal Classification (CTC)

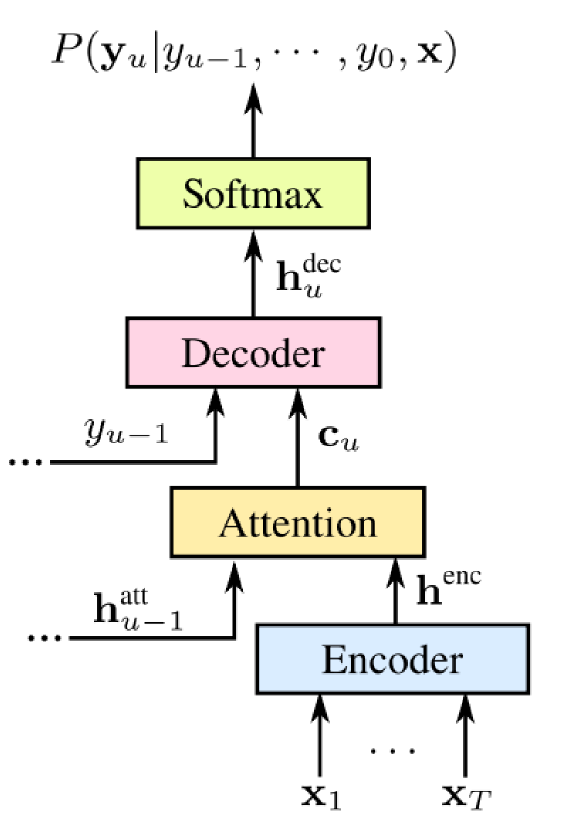
Lattice-free maximum mutual information (LF-MMI)

RNN transducer (RNN-T)

Sequence-to-sequence (S2S)

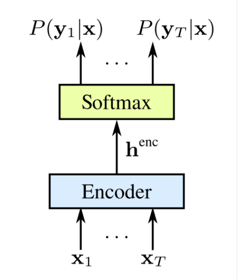


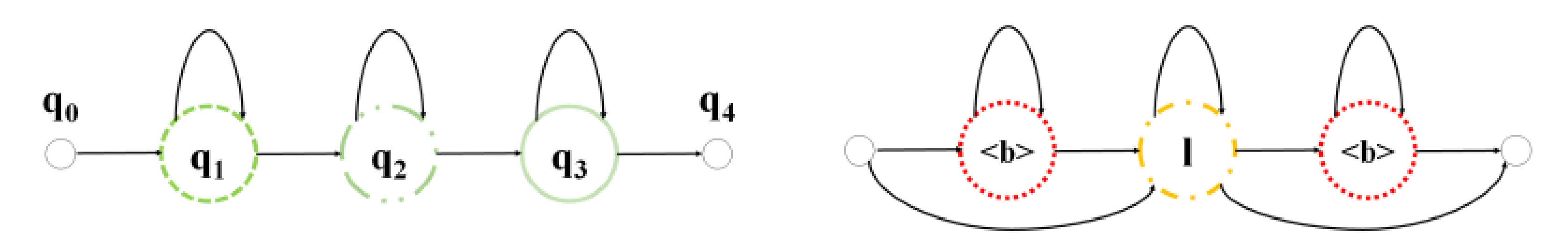



# CTC

- Formula




$$\mathcal{F}_{\text{CTC}} = \sum_u \log P(\mathbf{W}_u|\mathbf{O}_u)$$
$$= \sum_u \log \sum_{\mathbf{L}\in\mathcal{L}(\mathbf{W}_u)} P(\mathbf{L}|\mathbf{O}_u)P(\mathbf{W}_u|\mathbf{L})$$



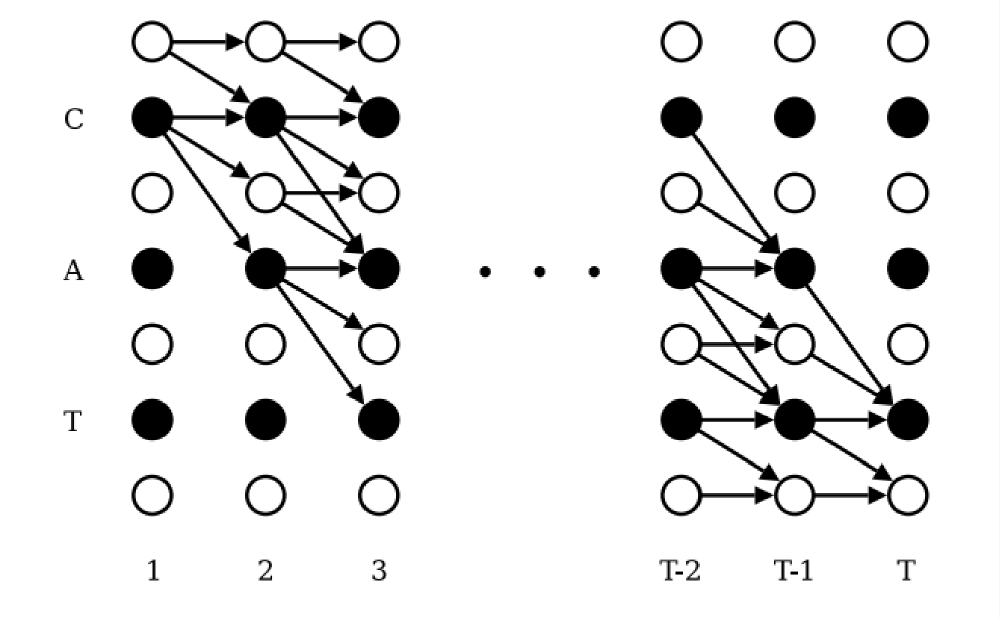

**(a) Traditional HMM** **(b) CTC**

- Training

Probability up to frame t and path length s

Blank or self-loop

$$\alpha_t(s) = y^t_{l'_s} \begin{cases} \sum_{i=s-1}^{s} \alpha_{t-1}(i) & \text{if } l'_s = b \text{ or } l'_{s-2} = l'_s \\ \sum_{i=s-2}^{s} \alpha_{t-1}(i) & \text{otherwise,} \end{cases}$$



# LF-MMI

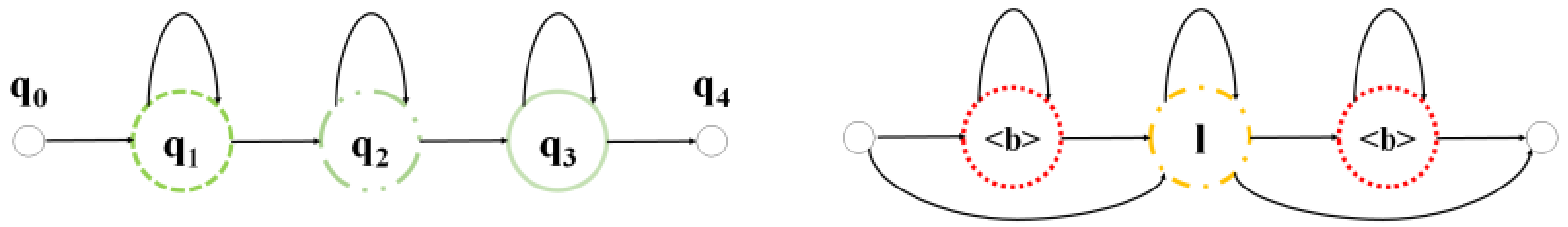
- Formula

$$\mathcal{F}_{\text{CTC}} = \sum_u \log P(\mathbf{W}_u|\mathbf{O}_u)$$
$$= \sum_u \log \sum_{\mathbf{L}\in\mathcal{L}(\mathbf{W}_u)} P(\mathbf{L}|\mathbf{O}_u)P(\mathbf{W}_u|\mathbf{L})$$

$$P(\mathbf{W}_u|\mathbf{O}_u) = \frac{p(\mathbf{O}_u|\mathbf{W}_u)P(\mathbf{W}_u)}{p(\mathbf{O}_u)}$$

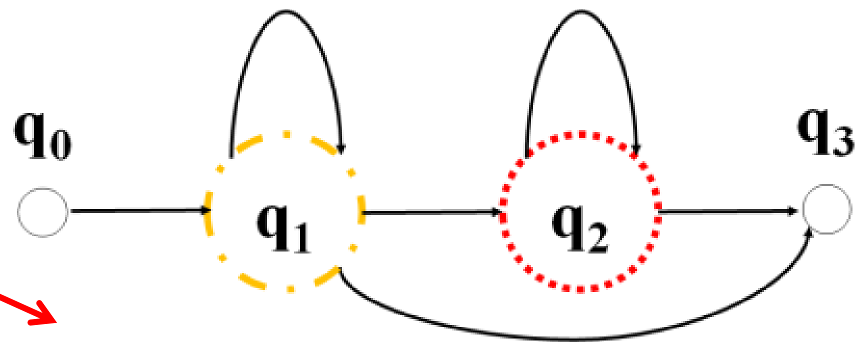
$$p(\mathbf{O}|\mathbf{W}) = \sum_{\mathbf{L}\in\mathcal{L}(\mathbf{W})} p(\mathbf{O}|\mathbf{L})P(\mathbf{L}|\mathbf{W})$$

- Training

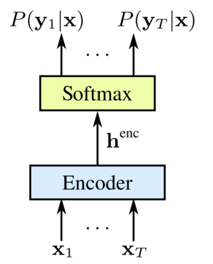



(a) Traditional HMM

(b) CTC

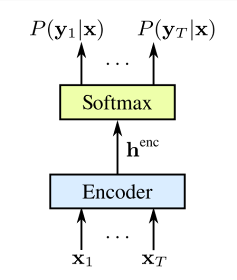



(c) HMM-PB (Povey et al. 2016)



# CTC v.s. LF-MMI

- Normalization
  - Any decoding results in block v.s. softmax of all alignments
- Alignment
  - w/ and w/o Movable window: flexibility + supervision
- Phone-wise Blank: better generalization
- Lower framerate
- Joint training with language model

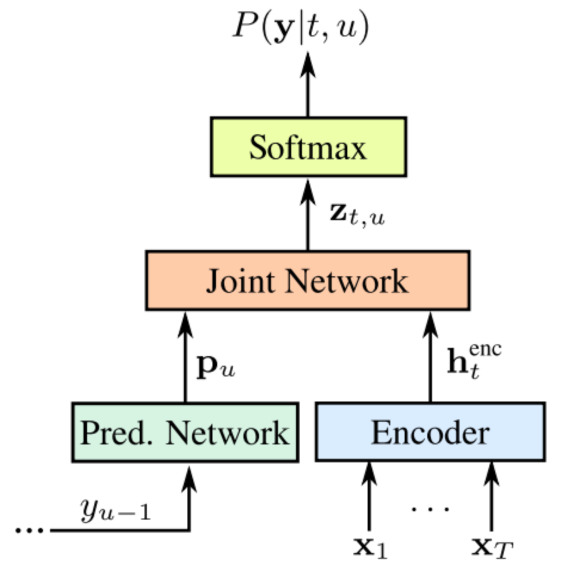


# RNN-Transducer

- Formula

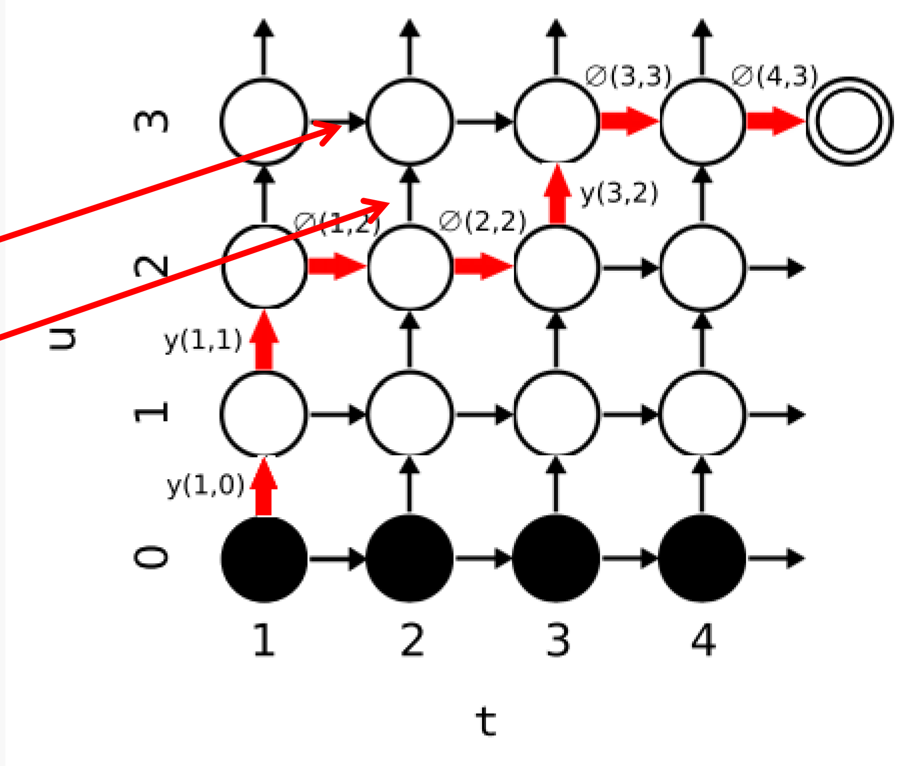
$$h(k,t,u) = \exp\left(f_t^k + g_u^k\right)$$

$$\Pr(k \in \bar{\mathcal{Y}} | t,u) = \frac{h(k,t,u)}{\sum_{k' \in \bar{\mathcal{Y}}} h(k',t,u)}$$




- Training

$$\alpha(t,u) = \alpha(t-1,u)\varnothing(t-1,u) + \alpha(t,u-1)y(t,u-1)$$



# RNN-Transducer

- Formula

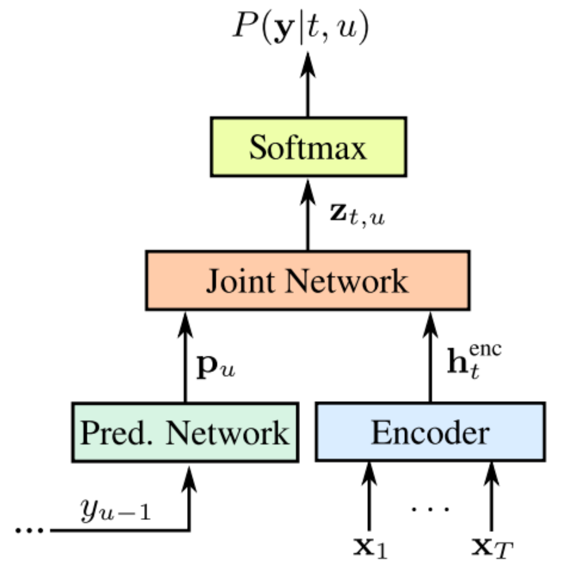
$$h(k,t,u) = \exp\left(f_t^k + g_u^k\right)$$

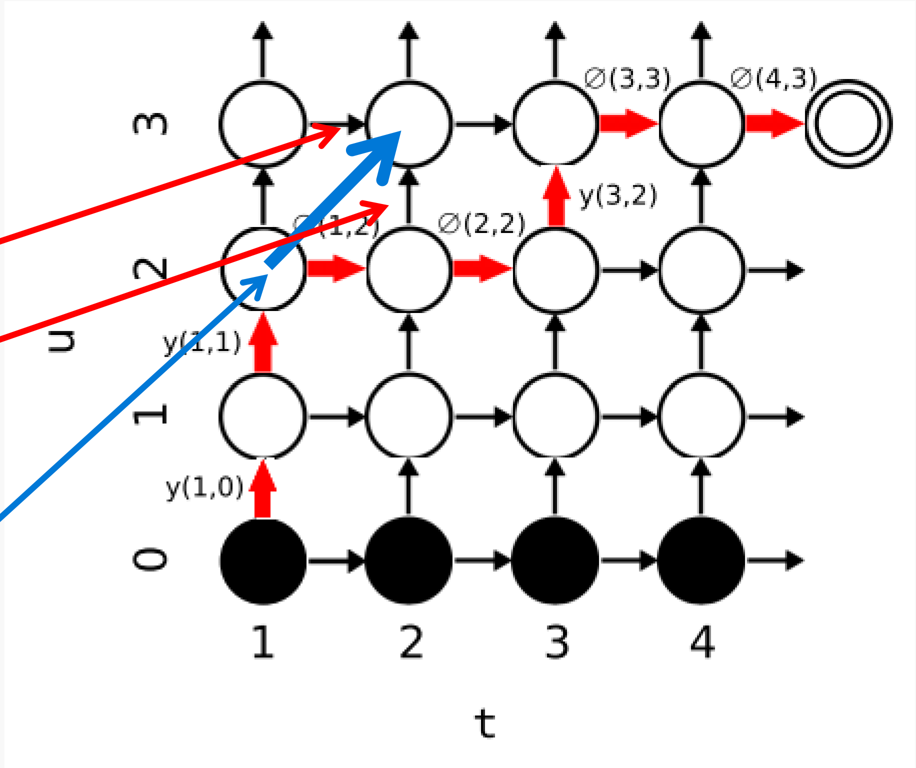
$$\Pr(k \in \bar{\mathcal{Y}} \mid t,u) = \frac{h(k,t,u)}{\sum_{k' \in \bar{\mathcal{Y}}} h(k',t,u)}$$

- Training

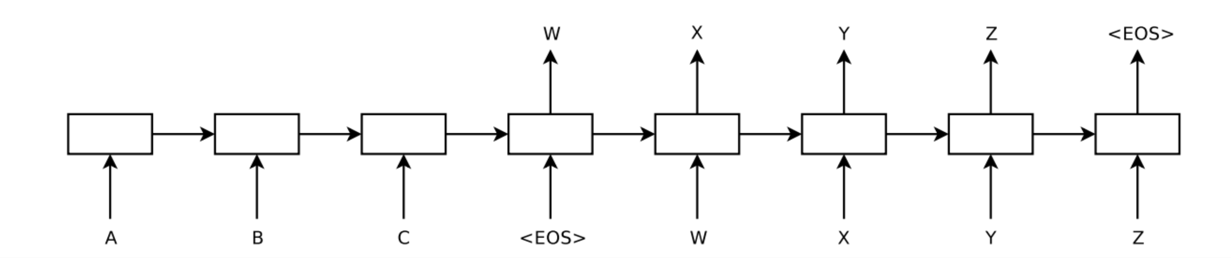
$$\alpha(t,u) = \alpha(t-1,u)\varnothing(t-1,u) + \alpha(t,u-1)y(t,u-1)$$

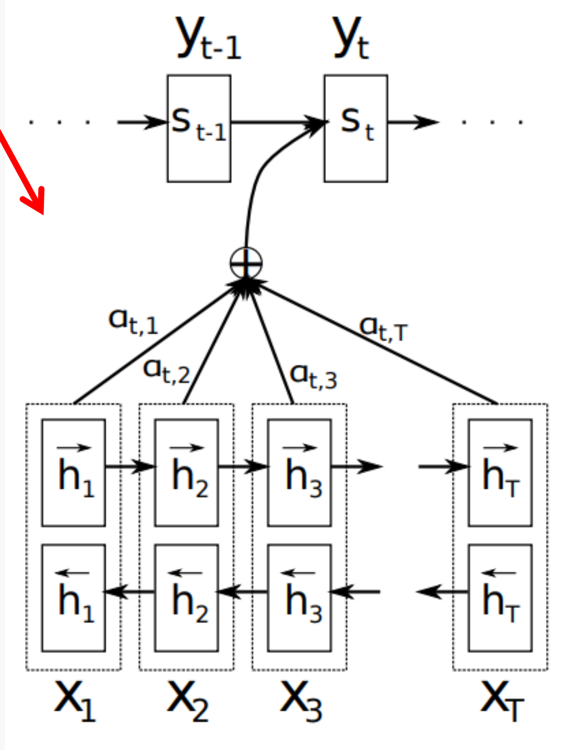
$$\boxed{+\,\alpha(t-1,u-1)y'(t-1,u-1)}$$

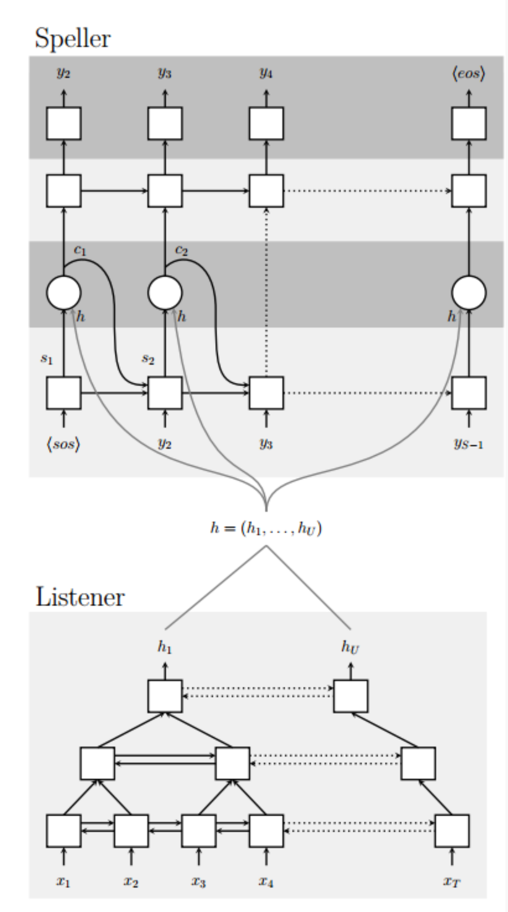


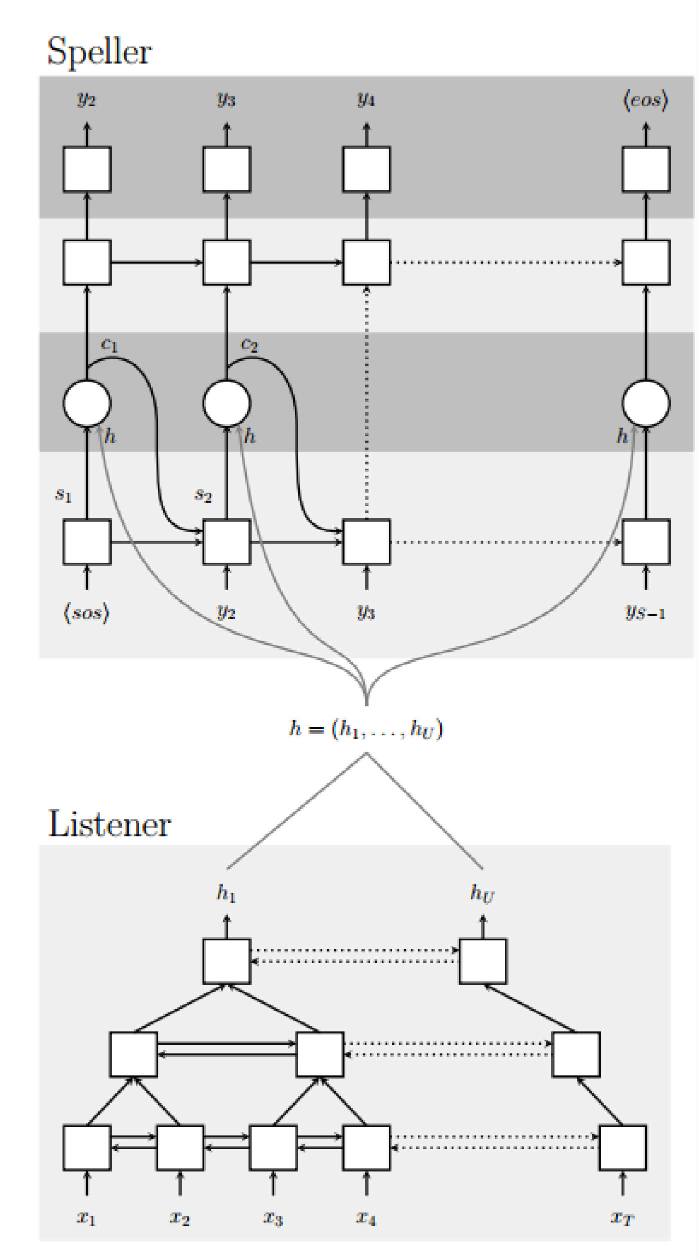
# S2S


W
X
Y
Z
<EOS>
A
B
C
<EOS>
W
X
Y
Z

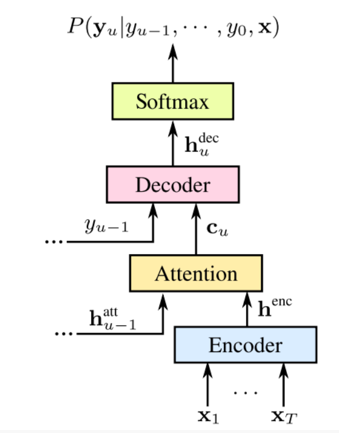

$y_{t-1}$
$y_t$
$s_{t-1}$
$s_t$
$a_{t,1}$
$a_{t,2}$
$a_{t,3}$
$a_{t,T}$
$\overrightarrow{h}_1$ $\overrightarrow{h}_2$ $\overrightarrow{h}_3$ $\overrightarrow{h}_T$
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ $\overleftarrow{h}_T$
$X_1$ $X_2$ $X_3$ $X_T$


Speller
$h = (h_1, \dots, h_U)$
Listener

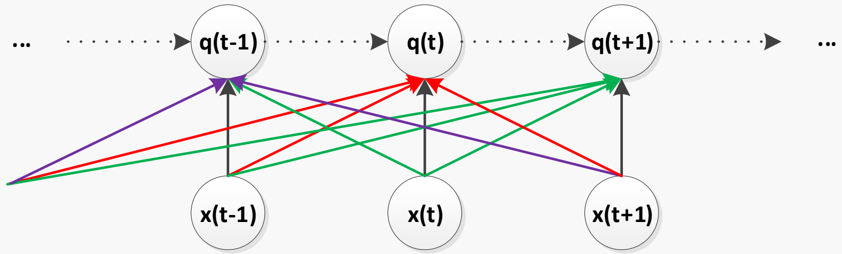
# S2S



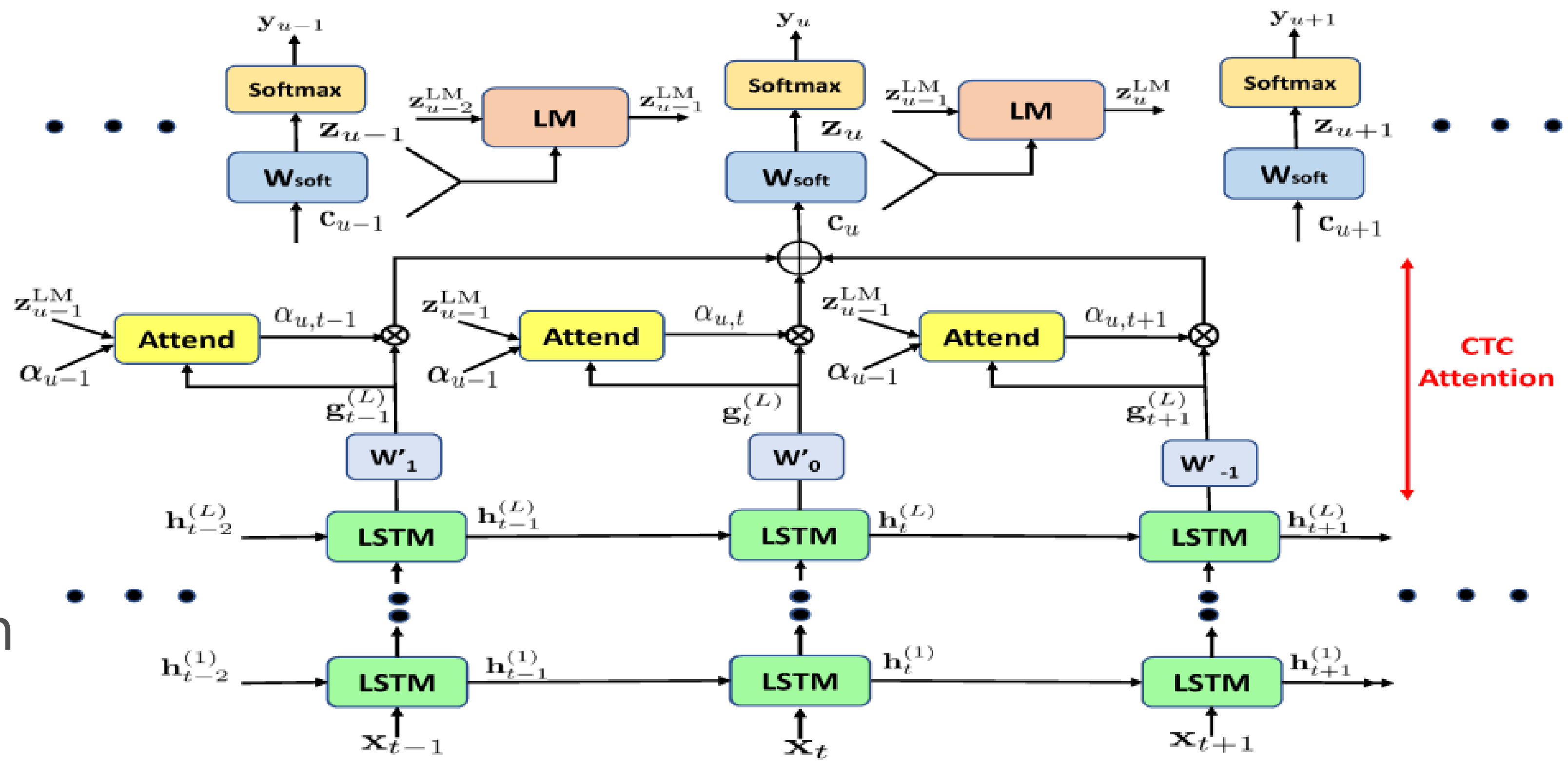

| stage | the last output of the decoder |
|---|---|
| Training | Ground truth |
| Inference | Last inference with highest probability |



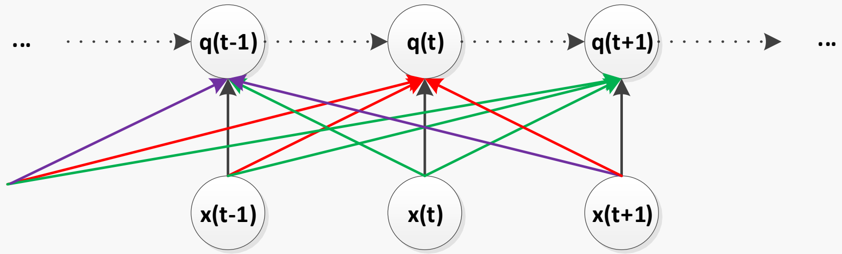
# Improvement of CTC / LFMMI




- Attention CTC
  - **Motivation:**
  - 1. hard align -> soft align
  - 2. change modeling but not criterion
  - **Method:**
  - 1. Chunk based
  - 2. time convolution to obtain g_t
  - 3. output z_u to replace h_u in obtaining attention weight \alpha
  - 4. diff weight \alpha for diff dimension of g_t

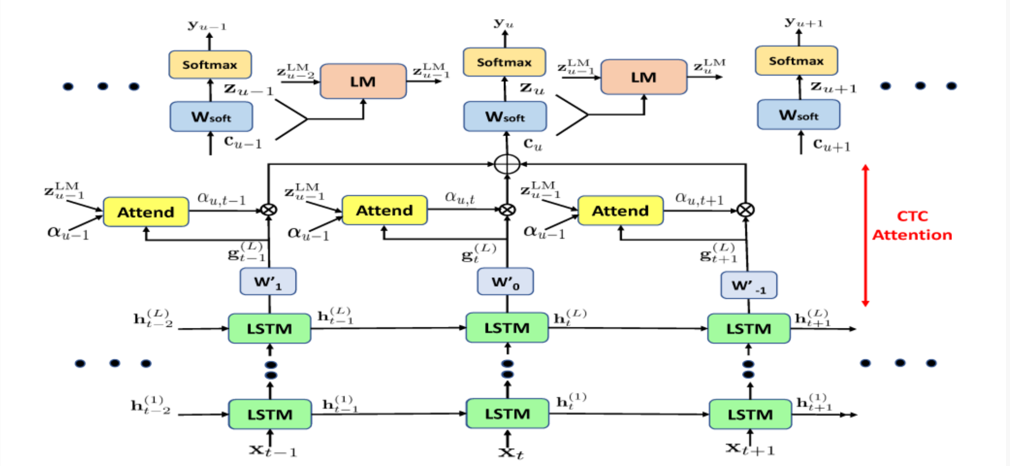


# Improvement of CTC / LFMMI

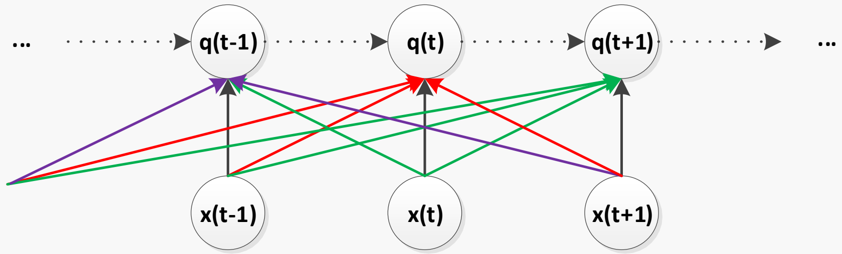



- Attention CTC
  - 1. Chunk based
  - 2. time convolution to obtain g_t
  - 3. output z_u to replace h_u in obtaining attention weight \alpha
  - 4. diff weight \alpha for diff dimension of g_t

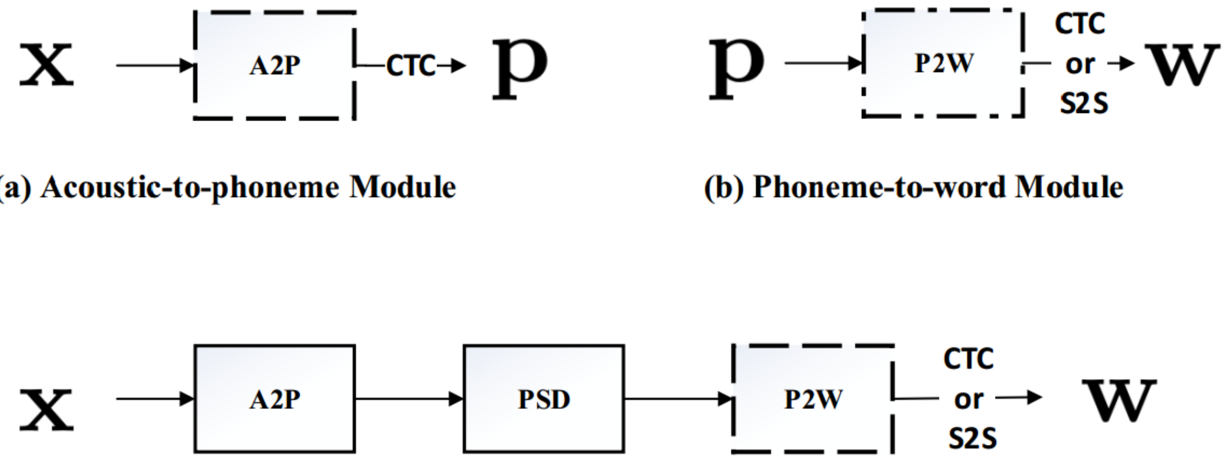
- Add language model as a "decoder"



# Improvement of word CTC




- Modular Training

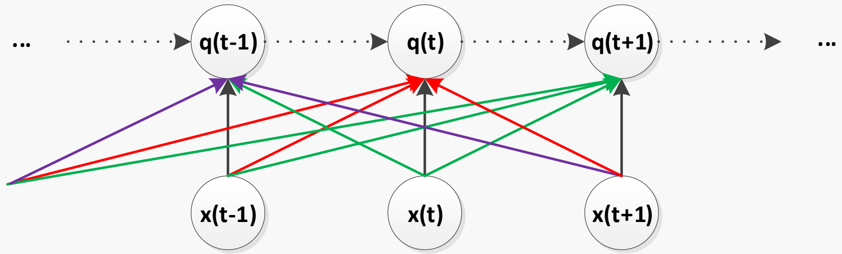



(a) Acoustic-to-phoneme Module

(b) Phoneme-to-word Module

(c) PSD-based Joint Training

# Improvement of word CTC

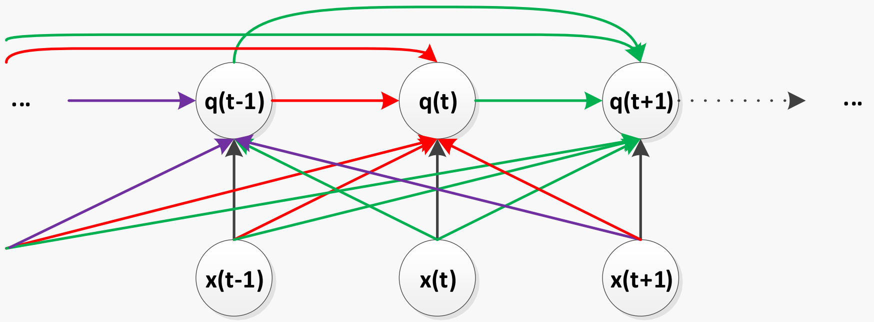


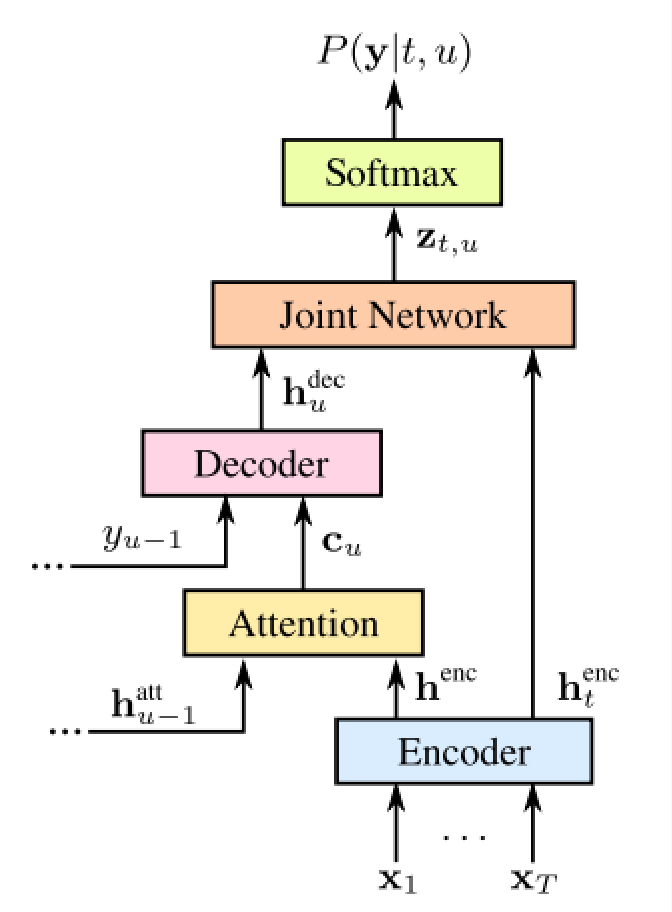

- Modular Training
- Data augmentation, structure & training tricks
- Cope with OOV / words seldom existing in training
  - Multi-task
  - Joint inference in single output, e.g.: A P P L E <APPLE>
  - Word-piece

# Improvement of RNN-T

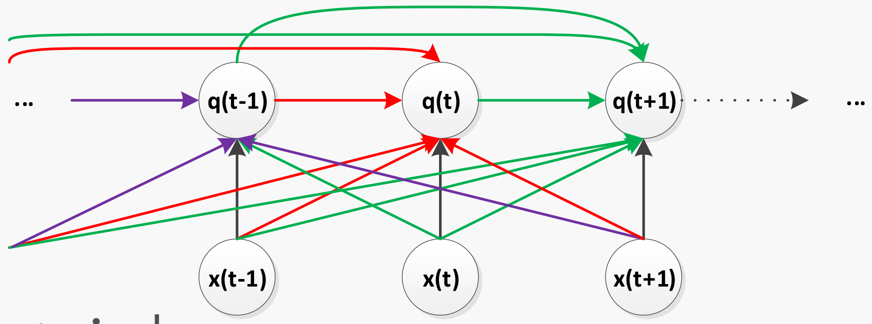
- Attention RNN-T
  - The decoder network depend on the entire encoder representation
  - Criterion is the same
  - Still frame-synchronous decoding

- Language model initialization
- Improve Decoding (see next slide)



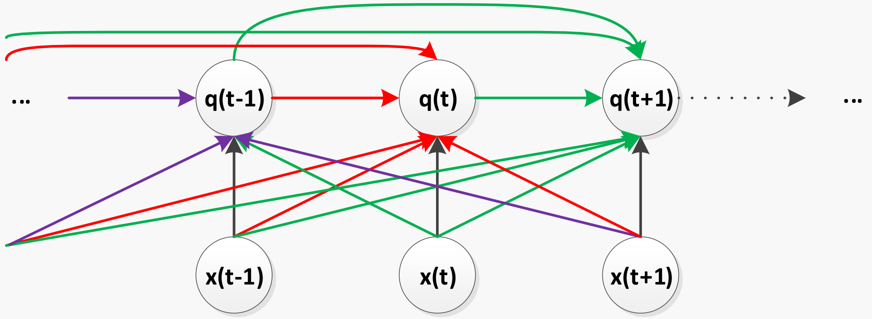
# Improvement of S2S

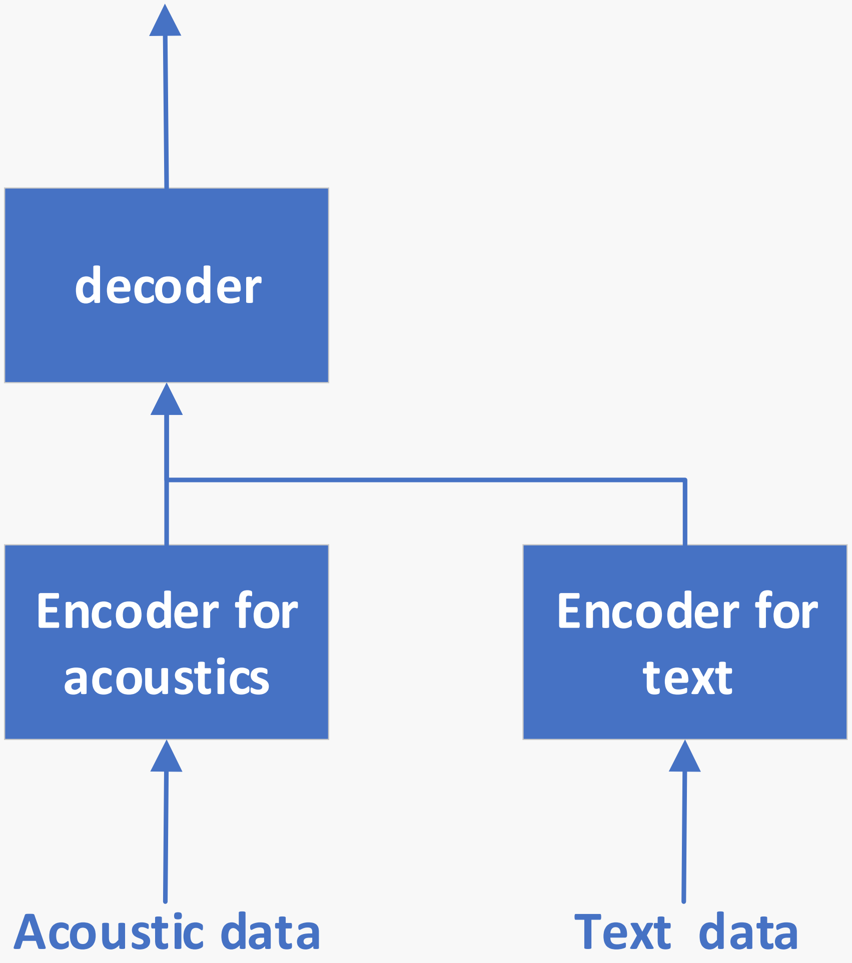



- Data augmentation, structure & training tricks
- Add language model (see next slide)
- Improve Decoding
  - Schedule sampling
  - Lattice-to-Sequence Models for Uncertain Inputs
  - Discriminative training (better sequential and discriminative modeling)
  - Reinforcement learning (minimum risk training for neural machine translation)
    - agent: S2S model;
    - state: concatenation of context & hidden state in S2S;
    - Action: output label set
    - Reward: WER variants; change to temporal distributed reward

# Add language model

- Cope with OOV (as discussed above)
  - Multi-task
  - Joint inference in single output, e.g.: A P P L E <APPLE>
  - Word-piece
- Multi-task/view framework:
  - LM & AM using shared layers
  - Using text and acoustic data to train AM & LM respectively
  - Add synthetic data designed for LM: synthesized input generated from large text corpora by some duration models / rules
- External RNNLM joint training
- How to adapt the LM?



# Outline

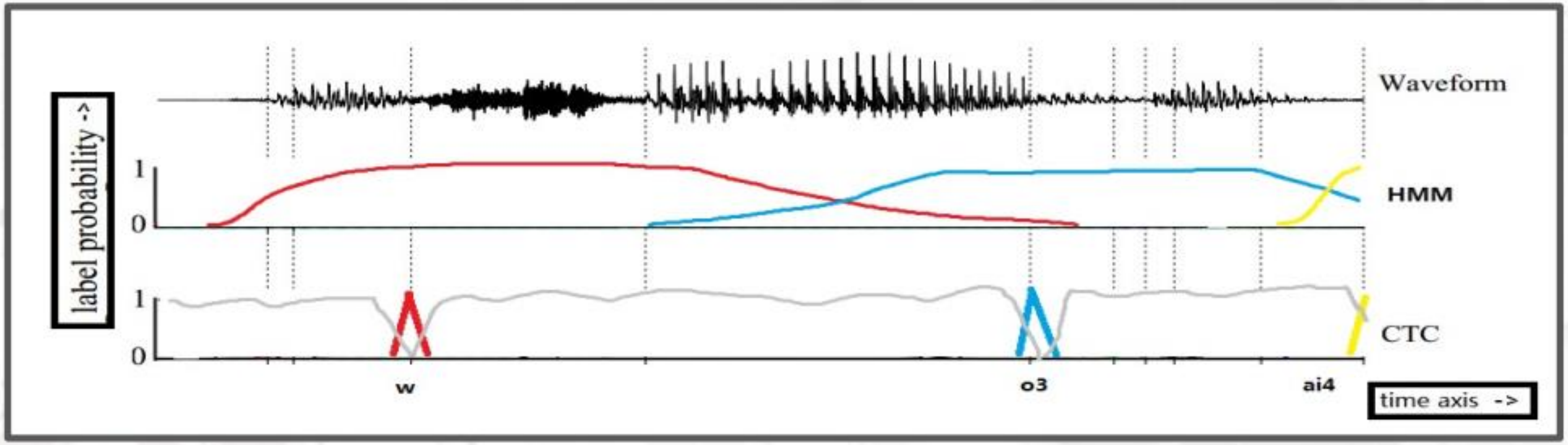
- End-to-end Modeling
  - CTC
  - LF-MMI
  - RNN-transducer
  - Sequence-to-sequence

- End-to-end Inference
  - **Phone level PSD**
  - **Word level PSD**
  - **Reducing WFST sizes**

# Phone level PSD

## CTC 的概率尖峰特性

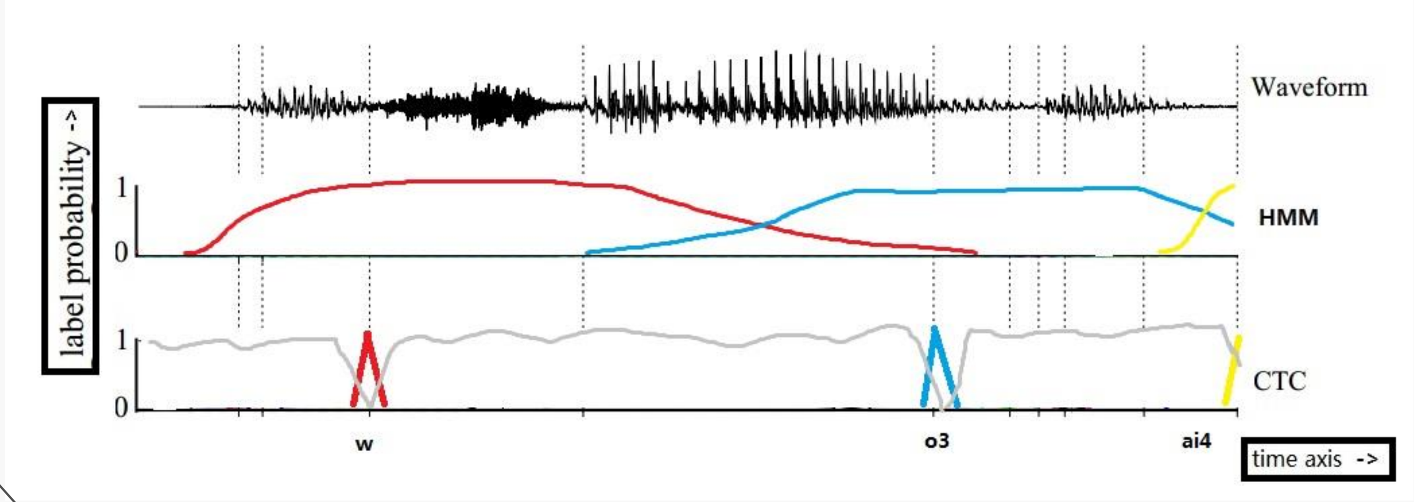
（ Connectionist Temporal Classification ）

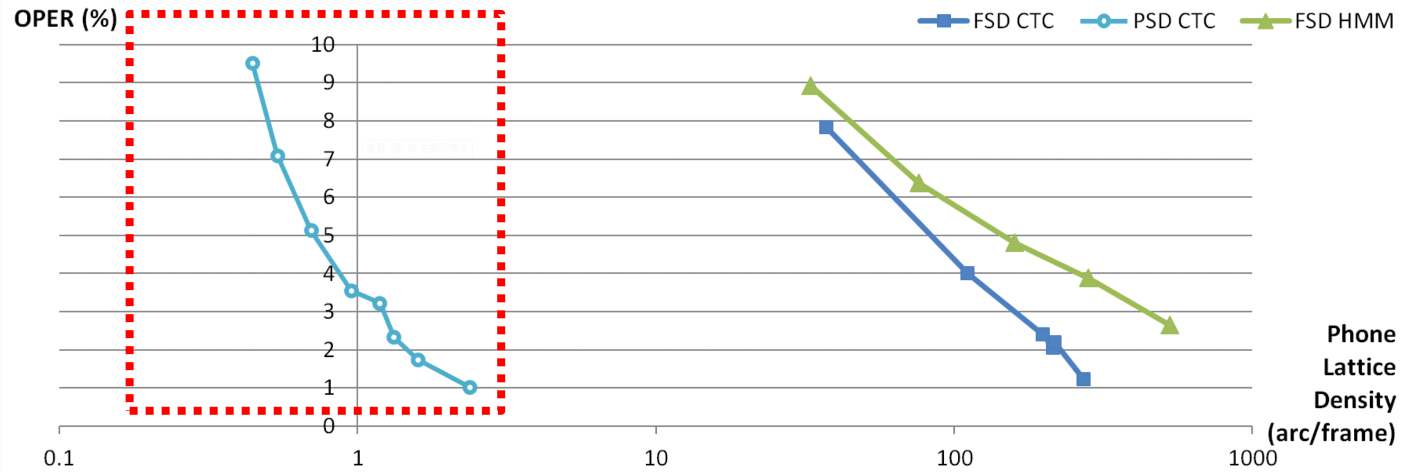



- blank 后验概率在绝大多数情况下占据主导
- 音素（phone）概率被训练过程集中推成尖峰

# Phone level PSD

Reduce information rate without precision loss



# Phone level PSD

## CTC在解码中的应用：音素同步解码

### 传统帧同步Viterbi 解码

$$\mathbf{w}^* = \operatorname*{argmax}_{\mathbf{w}}\{P(\mathbf{w})p(\mathbf{x}|\mathbf{w})\} = \operatorname*{argmax}_{\mathbf{w}}\{P(\mathbf{w})p(\mathbf{x}|\mathbf{l_w})\}$$

$$= \operatorname*{argmax}_{\mathbf{w}}\left\{P(\mathbf{w})\max_{\mathbf{l_w}}\frac{P(\mathbf{l_w}|\mathbf{x})}{P(\mathbf{l_w})}\right\}$$

$$\cong \operatorname*{argmax}_{\mathbf{w}}\left\{P(\mathbf{w})\max_{\pi:\pi\in L',\mathcal{B}(\pi_{1:T})=\mathbf{l_w}}\frac{1}{P(\mathbf{l_w})}\prod_{t=1}^{\boxed{T}}y^t_{\pi_t}\right\}$$

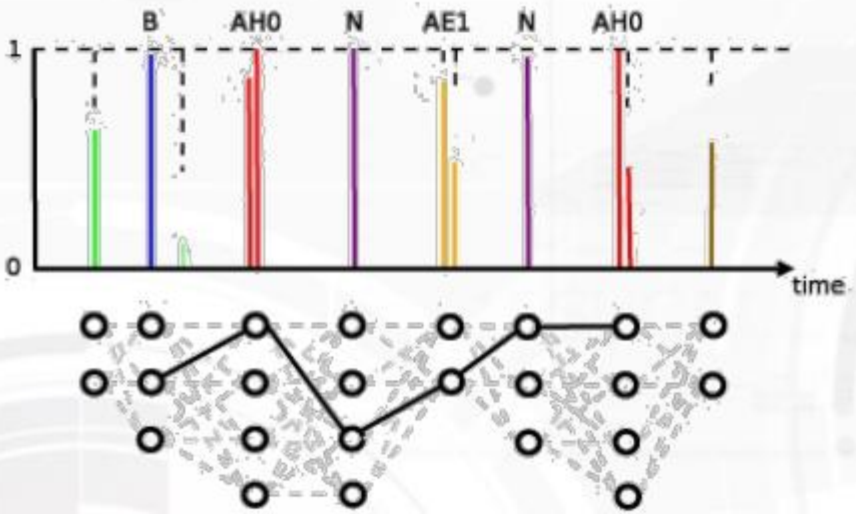
### 从帧同步到音素同步

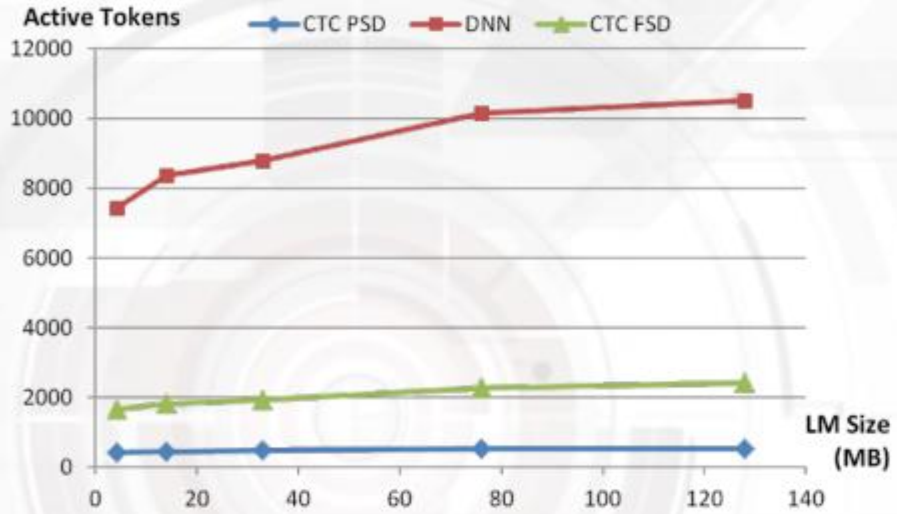
$$\mathbf{w}^* \cong \operatorname*{argmax}_{\mathbf{w}}\left\{P(\mathbf{w})\max_{\pi:\pi\in L',\mathcal{B}(\pi_{1:T})=\mathbf{l_w}}\frac{1}{P(\mathbf{l_w})}\left\{U=\{u:y^u_{\mathtt{blank}}\simeq 1\}\prod_{t\notin U}y^t_{\pi_t}\cdot\cancel{\prod_{t\in U}y^t_{\mathtt{blank}}}\right\}\right\} \quad (4)$$

$$= \operatorname*{argmax}_{\mathbf{w}}\left\{P(\mathbf{w})\max_{\pi':\pi'\in L,\mathcal{B}(\pi'_{1:J})=\mathbf{l_w}}\frac{1}{P(\mathbf{l_w})}\prod_{j=1}^{\boxed{J}}y^{t_j}_{\pi'_j}\right\} \quad (6) \qquad J = T-|U|$$

# Phone level PSD

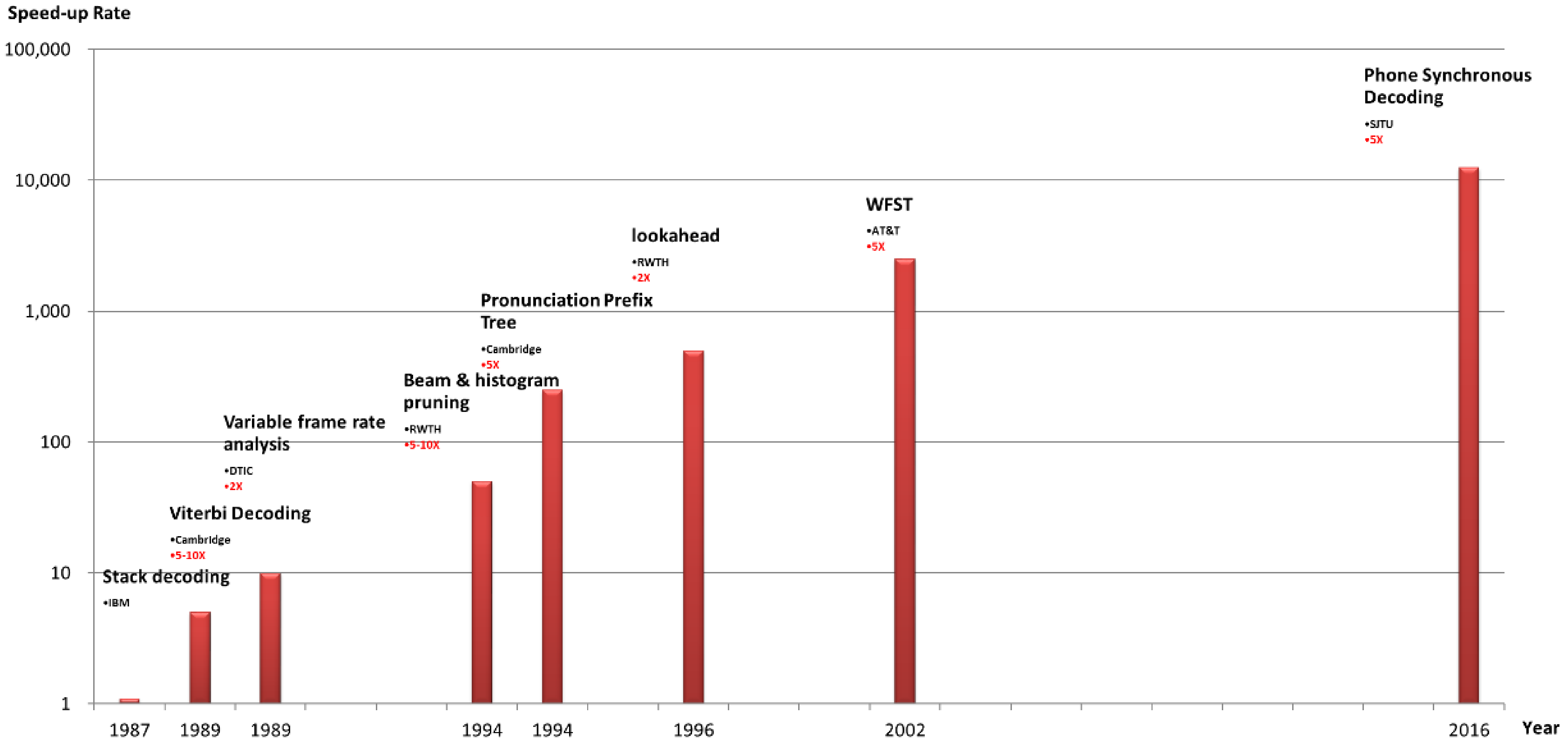
## 解码加速




| Model | Search Step | CER | RTF |
|---|---|---|---|
| HMM | Frame | 13.3 | 0.32 |
| CTC | Frame<br>Phone | 10.2<br>10.1 | **0.044(7.3X)**<br>**0.016(20X)** |

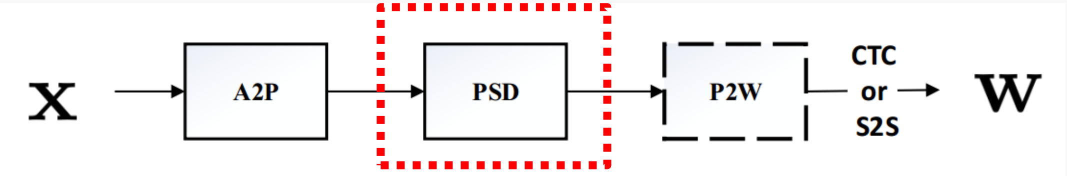
# Phone level PSD



# Word level PSD

- Motivation:
  - Different information rate in acoustics and phoneme
  - long sequence is hard for S2S (for speech, avg. 500 tokens)
  - Speedup training and decoding
- Results:
  - Better speed
  - Better performance

- Procedure:
  - A2P inference
  - PSD sub-sampling
  - P2W inference
  - Back propagation
  - fine-tune P2W only




(c) PSD-based Joint Training

# Reducing WFST sizes

| Exp-ID | Model | Unidi | 1st pass Model Size |
|---|---|---|---|
| E8 | Proposed | **5.8** | **0.4 GB** |
| E9 | Conventional LFR system | 6.7 | 0.1 GB (AM) + 2.2 GB (PM) + 4.9 GB (LM) = 7.2GB |

**Table 5:** The improved LAS outperforms the conventional LFR system while being more compact. Both models use second-pass rescoring.

- Especially in multi-dialect ASR, which needs a respective WFST for each dialect