MITSUBISHI ELECTRIC
Changes for the Better

for a greener tomorrow

MITSUBISHI ELECTRIC RESEARCH LABORATORIES
Cambridge, Massachusetts

Advanced topics in end-to-end speech recognition

Hybrid CTC/Attention Architecture for End-to-End Speech Recognition

Takaaki Hori, Shinji Watanabe,
Jonathan Le Roux, John Hershey,
MERL interns (Suyoun Kim, Tomoki Hayashi, Shane Settle, Hiroshi Seki)
External collaborators (Yu Zhang, William Chan)

2018 JHU Summer School on Human Language Technology
Wednesday, June 20, 2018

Outline

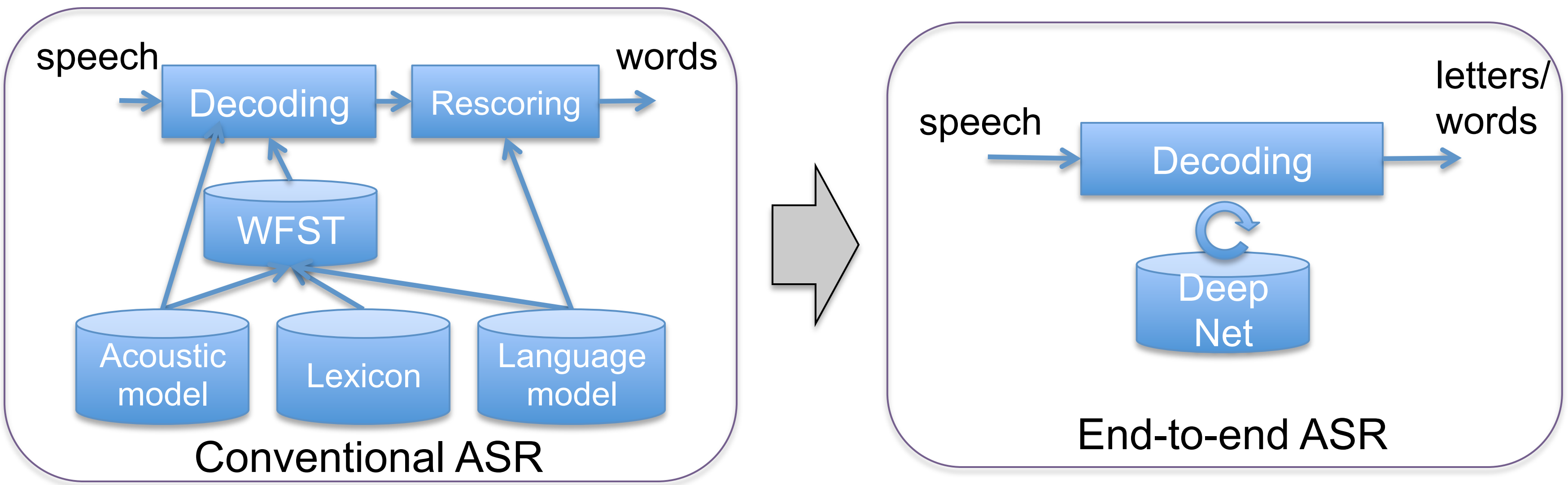
- End-to-end speech recognition

- Hybrid CTC/attention-based end-to-end speech recognition
  - Multi-task CTC/attention learning (ICASSP'17)
  - Joint CTC/attention decoding (ACL'17)
  - Integration with a deep CNN and an RNN-LM (Interspeech'17)
  - Multi-level language modeling and decoding (ASRU'17)

- Multi-lingual multi-speaker end-to-end speech recognition
  - Multi-lingual end-to-end speech recognition (ASRU'17, ICASSP'18)
  - Multi-speaker end-to-end speech recognition (ICASSP'18, ACL'18)

End-to-end Speech Recognition

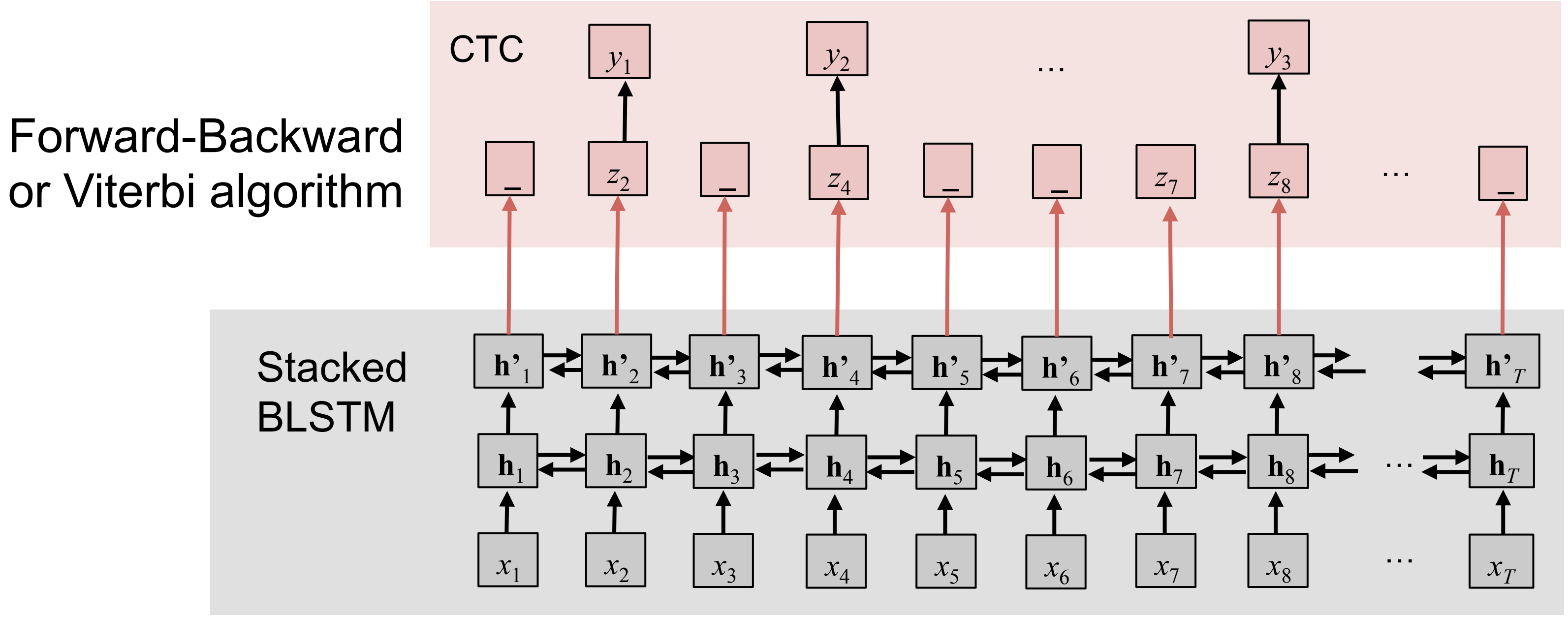
- Train a deep network that directly maps speech signal to the target letter/word sequence
- Greatly simplify the complicated model-building/decoding process
- Easy to build ASR systems for new tasks without expert knowledge
- Potential to outperform conventional ASR by optimizing the entire network with a single objective function



End-to-end ASR (1)
Connectionist temporal classification (CTC)

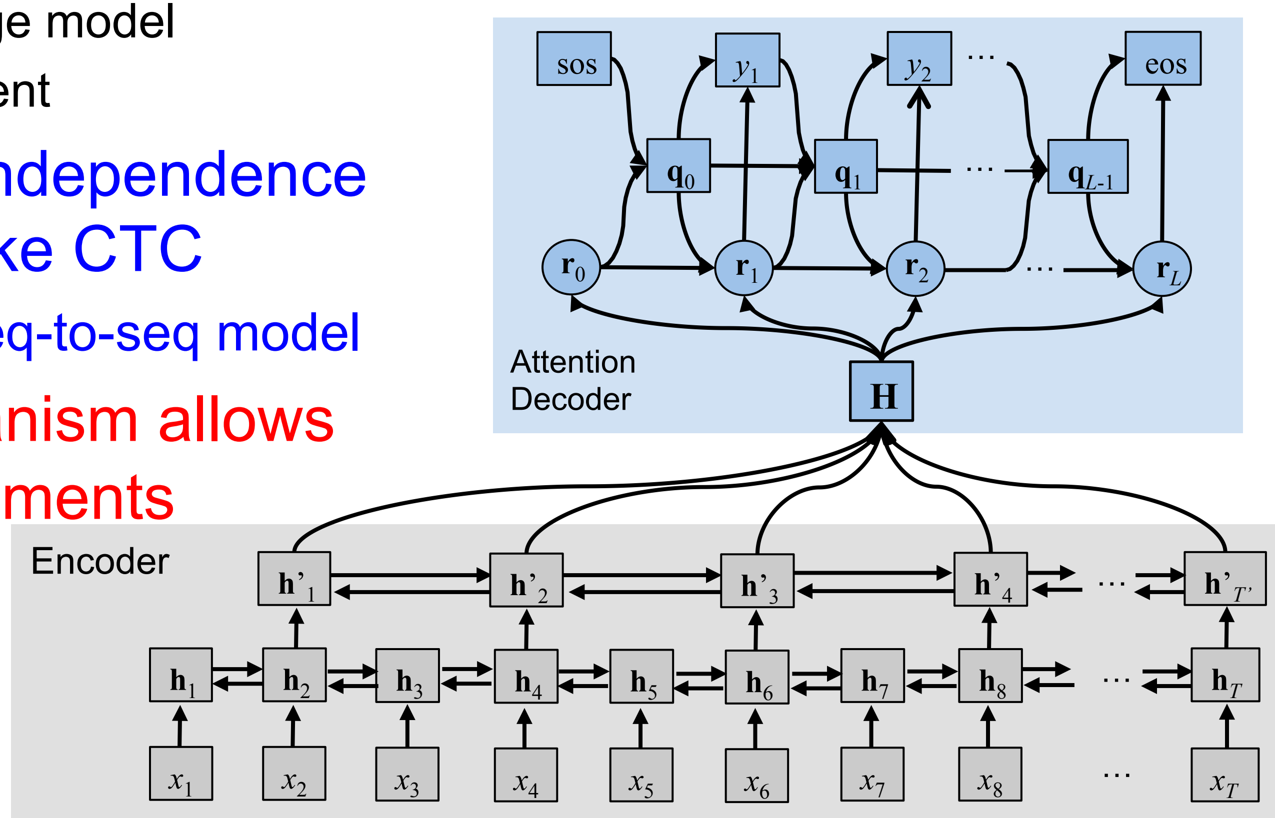
[Graves+ 2006, Graves+ 2014, Miao+ 2015]

- Use bidirectional RNNs to predict frame-based labels including blanks
- Find alignments between X and Y using dynamic programming
- Relying on conditional independence assumptions
- Output sequence is not well modeled



End-to-end ASR (2)
Attention-based encoder decoder [Chorowski+ 2014, Chan+ 2015]

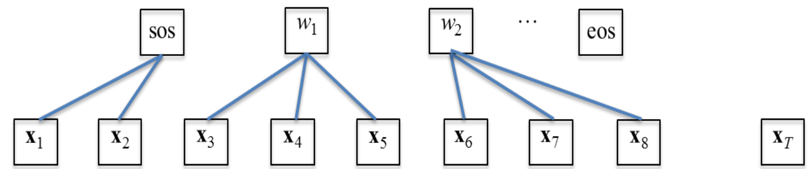
- Combine acoustic and language models in a single architecture
  - Encoder: acoustic model
  - Decoder: language model
  - Attention: alignment
- No conditional independence assumption unlike CTC
  - More precise seq-to-seq model
- Attention mechanism allows too flexible alignments
  - Hard to train the model from scratch



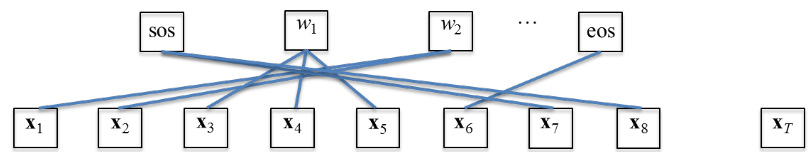
Input/output alignment by temporal attention

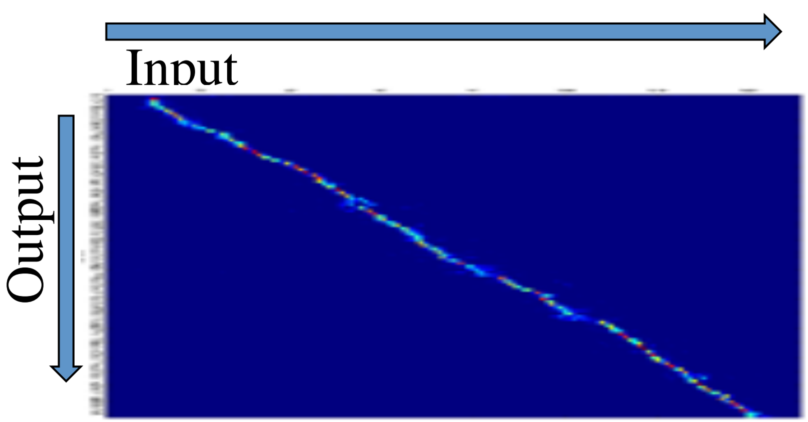
- Unlike CTC, attention model does not preserve order of inputs
- Our desired alignment in ASR task is **monotonic**
- Not regularized alignment makes the model **hard to learn** from scratch

HMM or CTC case

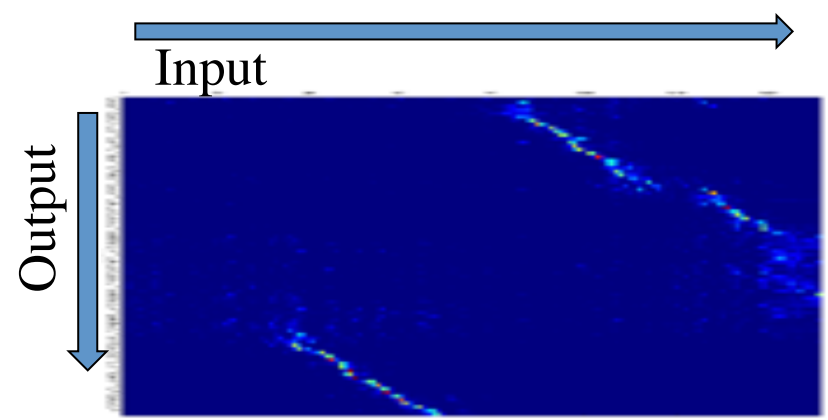


Attention model case





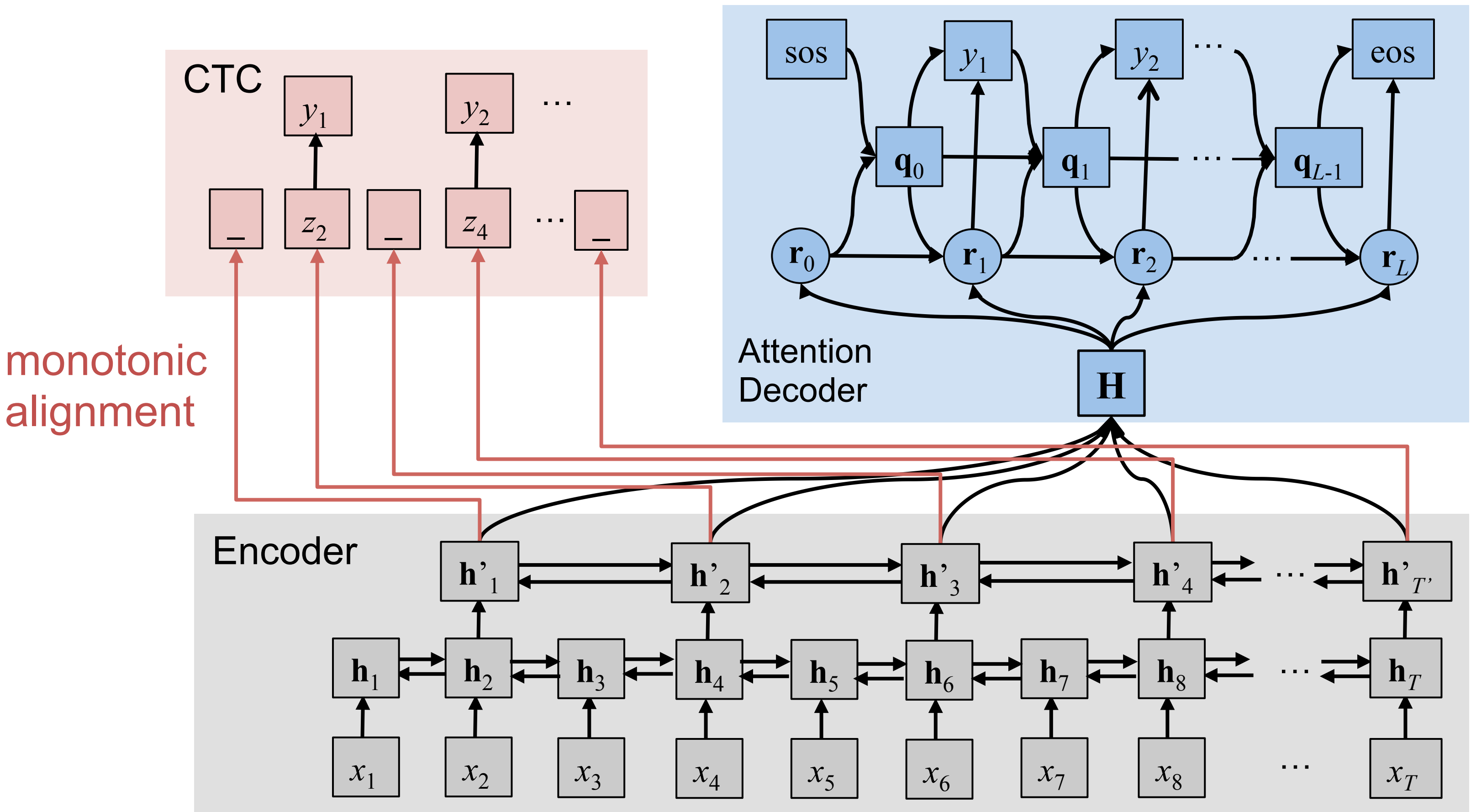
Example of monotonic alignment



Example of distorted alignment

Hybrid CTC/attention network [Kim+'17]

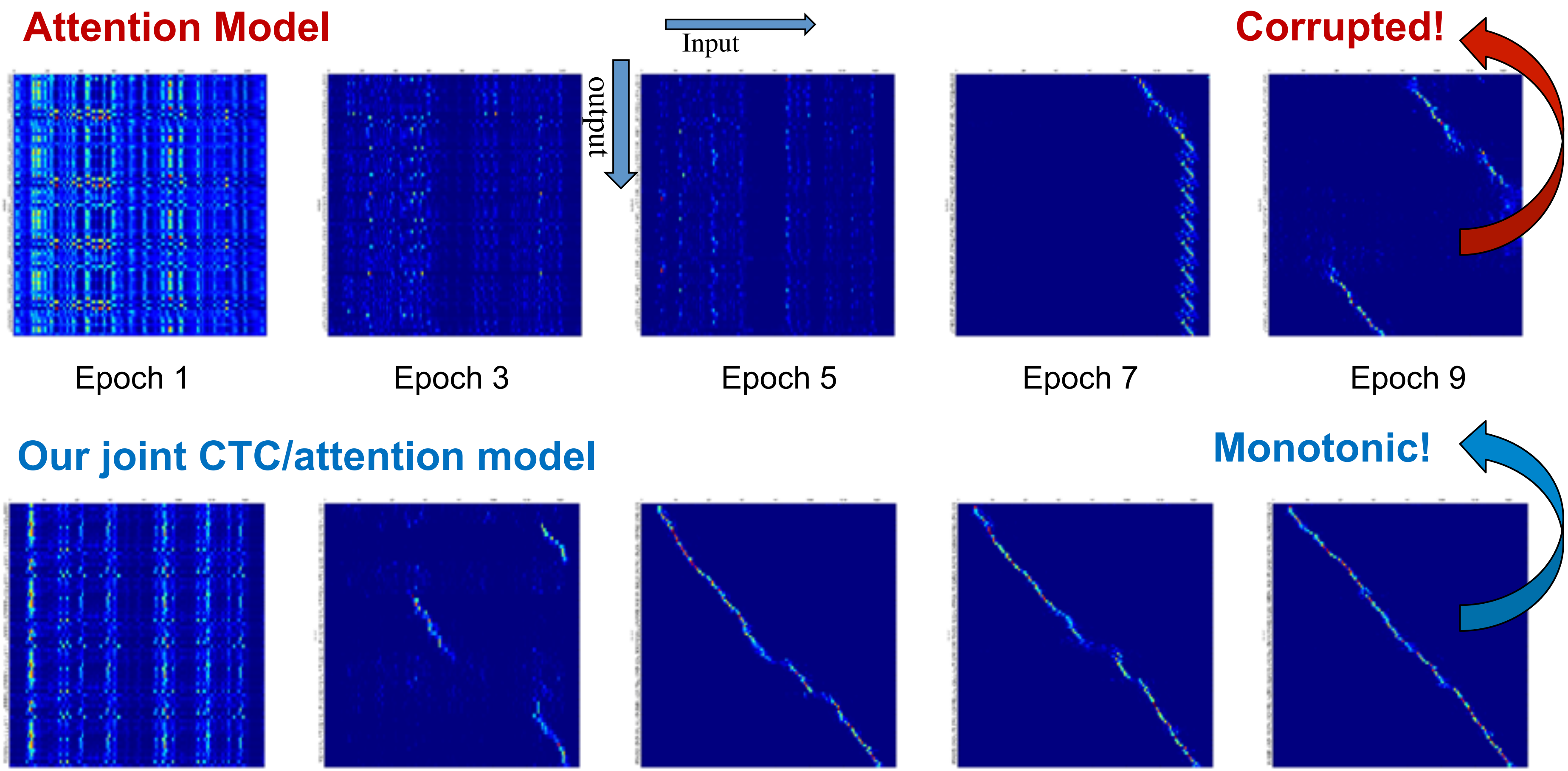
Multitask learning: $\mathcal{L}_{\mathrm{MTL}} = \lambda \mathcal{L}_{\mathrm{CTC}} + (1-\lambda)\mathcal{L}_{\mathrm{Attention}}$ λ: CTC weight



CTC guides attention alignment to be monotonic

More robust input/output alignment of attention

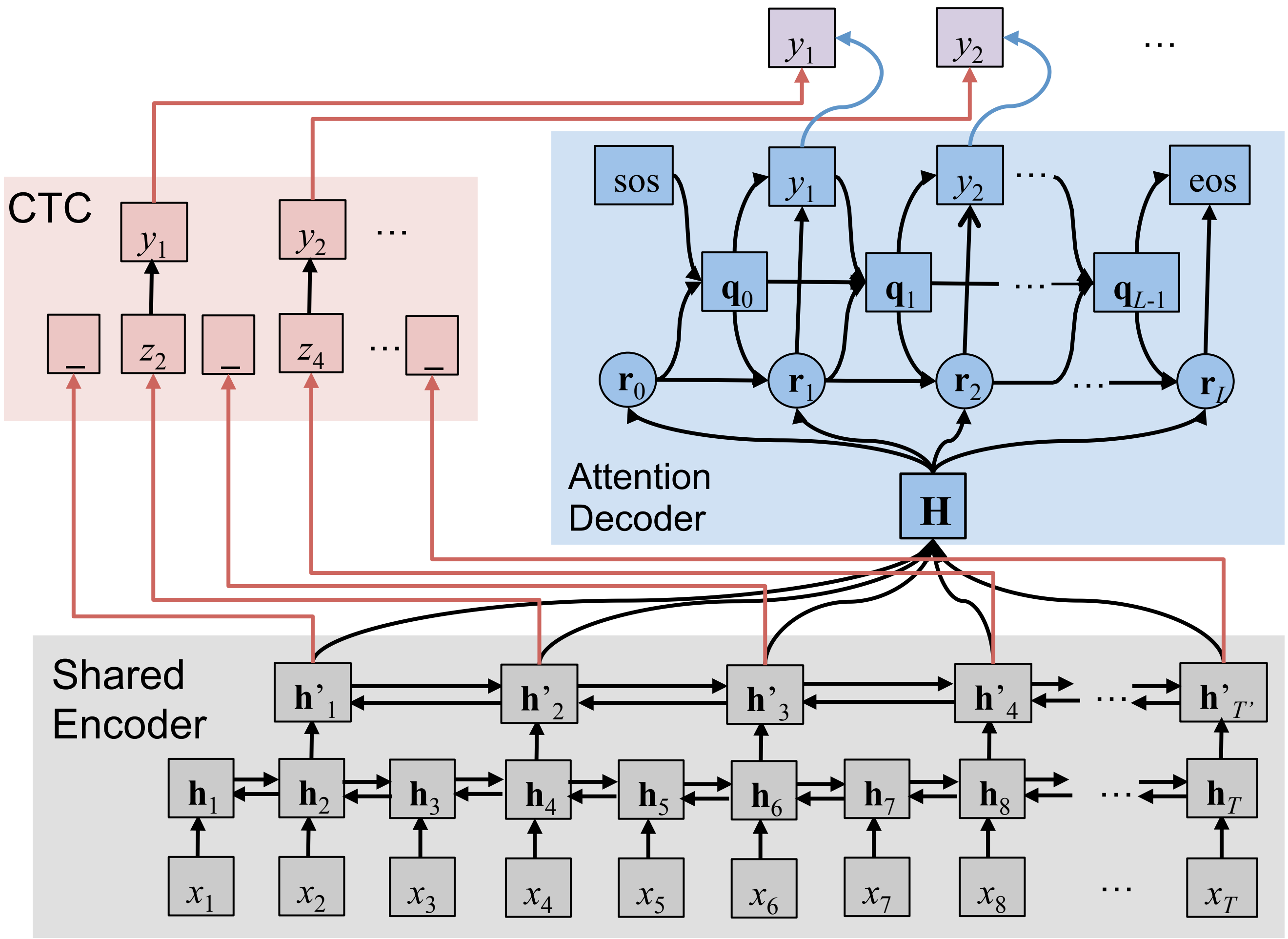
- Alignment of one selected utterance from CHiME4 task



Faster convergence

Joint CTC/attention decoding [Hori+'17]

Use CTC for decoding together with the attention decoder



Joint CTC/attention decoding

- Decoding objective is changed to the CTC/attention probability

$$\hat{Y} = \arg\max_{Y \in \mathcal{V}^*} \log p_{\mathrm{att}}(Y|X)$$

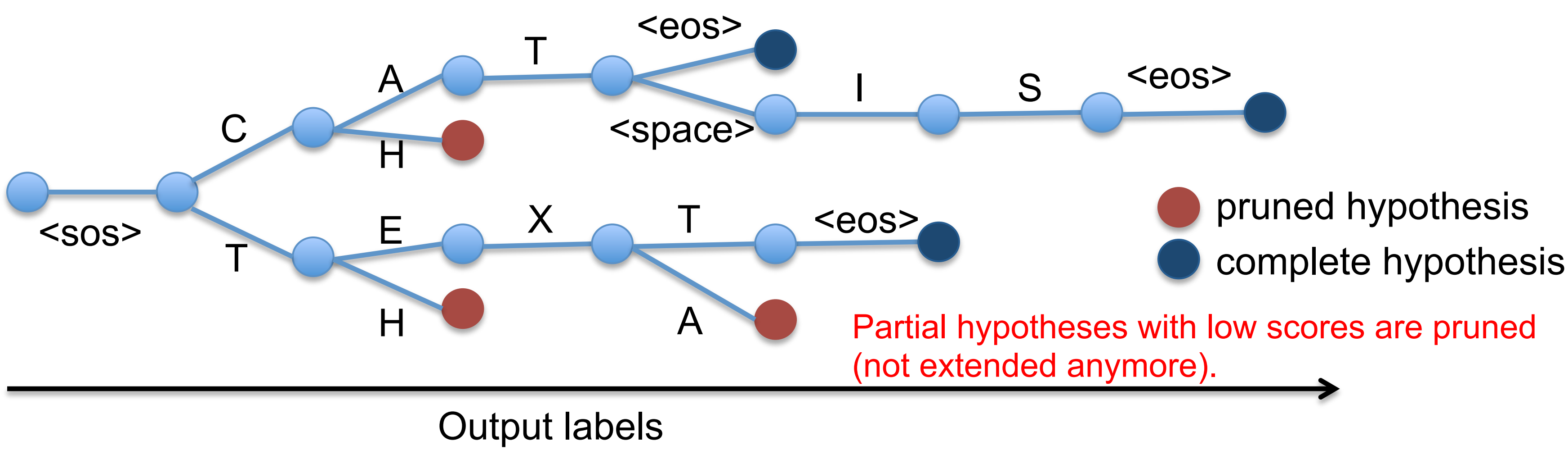
$\mathcal{V}$: vocabulary

$$\hat{Y} = \arg\max_{Y \in \mathcal{V}^*} \{\lambda \log \boxed{p_{\mathrm{ctc}}(Y|X)} + (1-\lambda) \log p_{\mathrm{att}}(Y|X)\}$$

λ: CTC weight

- CTC helps select better hypotheses in the decoding phase

Label-synchronous beam search for decoding



Attention model score of partial hypothesis h

$$\alpha_{\mathrm{att}}(h) = \alpha_{\mathrm{att}}(g) + \log p_{\mathrm{att}}(c|g, X)$$

$$h = g \cdot c \quad (g: \text{previous hypothesis},\ c: \text{next character})$$

Recognition output

$$\hat{Y} = \arg\max_{Y \in \Phi} \alpha_{\mathrm{att}}(Y)$$

Φ: set of complete hypotheses

Decoding strategies

- Rescoring approach
  - 1st pass employs the attention decoder using a beam search technique
  - 2nd-pass rescores the N-best hypotheses Φ_N with hybrid CTC/attention probabilities and select the best hypothesis

$$\hat{Y} = \arg\max_{Y \in \Phi_N} \{\lambda \log p_{\mathrm{ctc}}(Y|X) + (1-\lambda) \log p_{\mathrm{att}}(Y|X)\}$$

  - Can not save the hypotheses pruned in the 1st pass

- One-pass approach
  - Use the joint CTC/attention probabilities from the beginning of the search

$$\alpha_{\mathrm{joint}}(h) = \lambda \underline{\alpha_{\mathrm{ctc}}(h)} + (1-\lambda) \underline{\alpha_{\mathrm{att}}(h)}$$

  - Hopefully work with less pruning errors, but we don't know how to compute $\alpha_{\mathrm{ctc}}(h)$

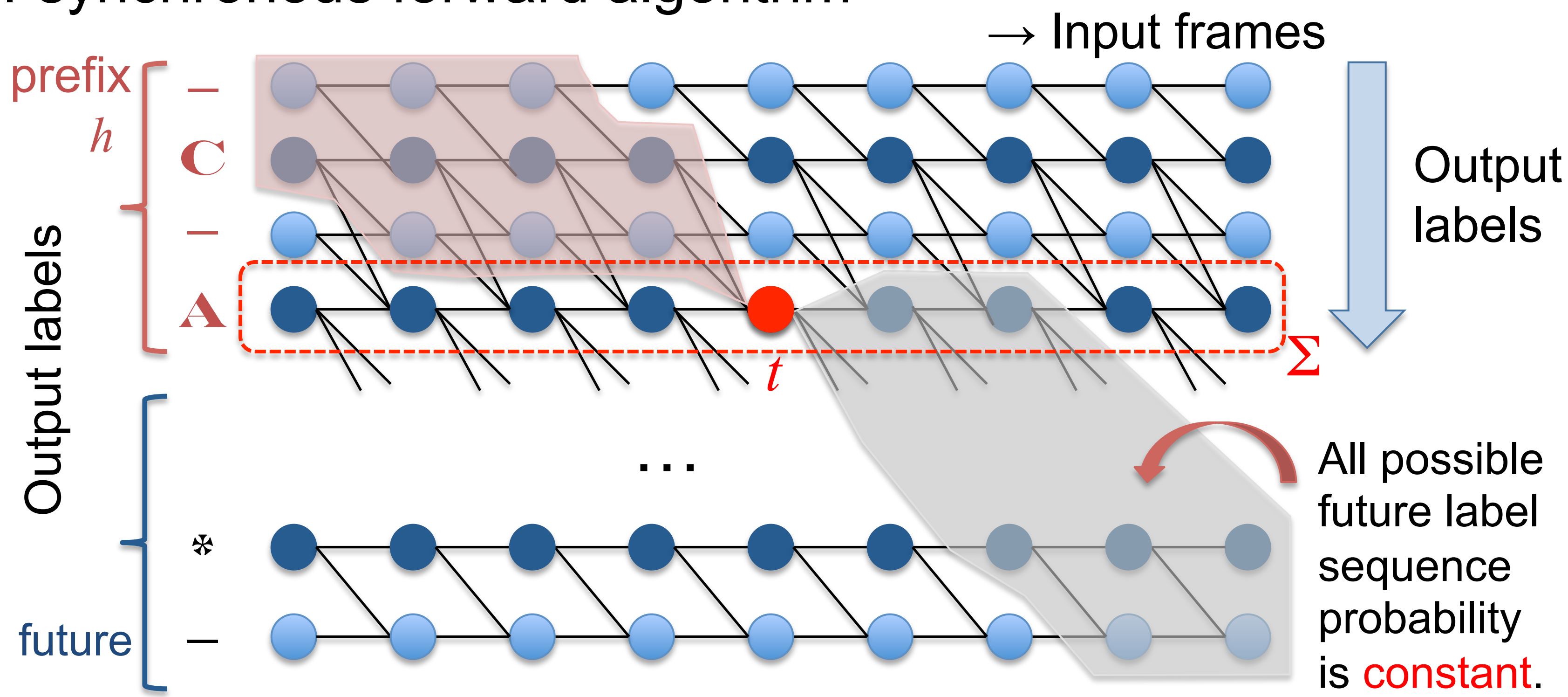
CTC-based hypothesis score

CTC prefix probability [Graves'08]

$$\alpha_{\mathrm{ctc}}(h) \triangleq \log \sum_{\nu \in (\mathcal{U} \cup \{\texttt{<eos>}\})^+} p_{\mathrm{ctc}}(h \cdot \nu | X)$$

Cumulative probability of all label sequences that prefix is h

Label-synchronous forward algorithm



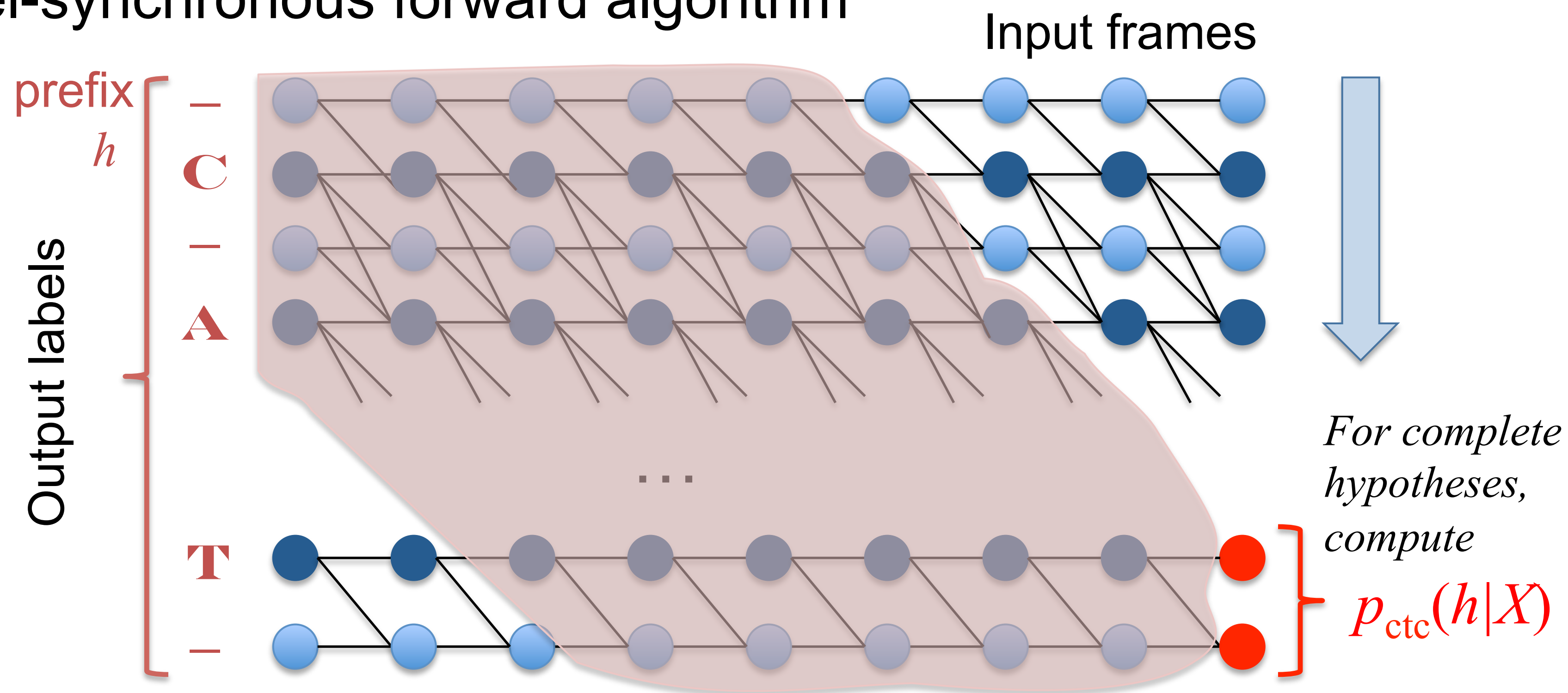
CTC-based hypothesis score

CTC prefix probability [Graves'08]

$$\alpha_{\mathrm{ctc}}(h) \triangleq \log \sum_{\nu \in (\mathcal{U} \cup \{\texttt{<eos>}\})^+} p_{\mathrm{ctc}}(h \cdot \nu | X)$$

Cumulative probability of all label sequences that prefix is h

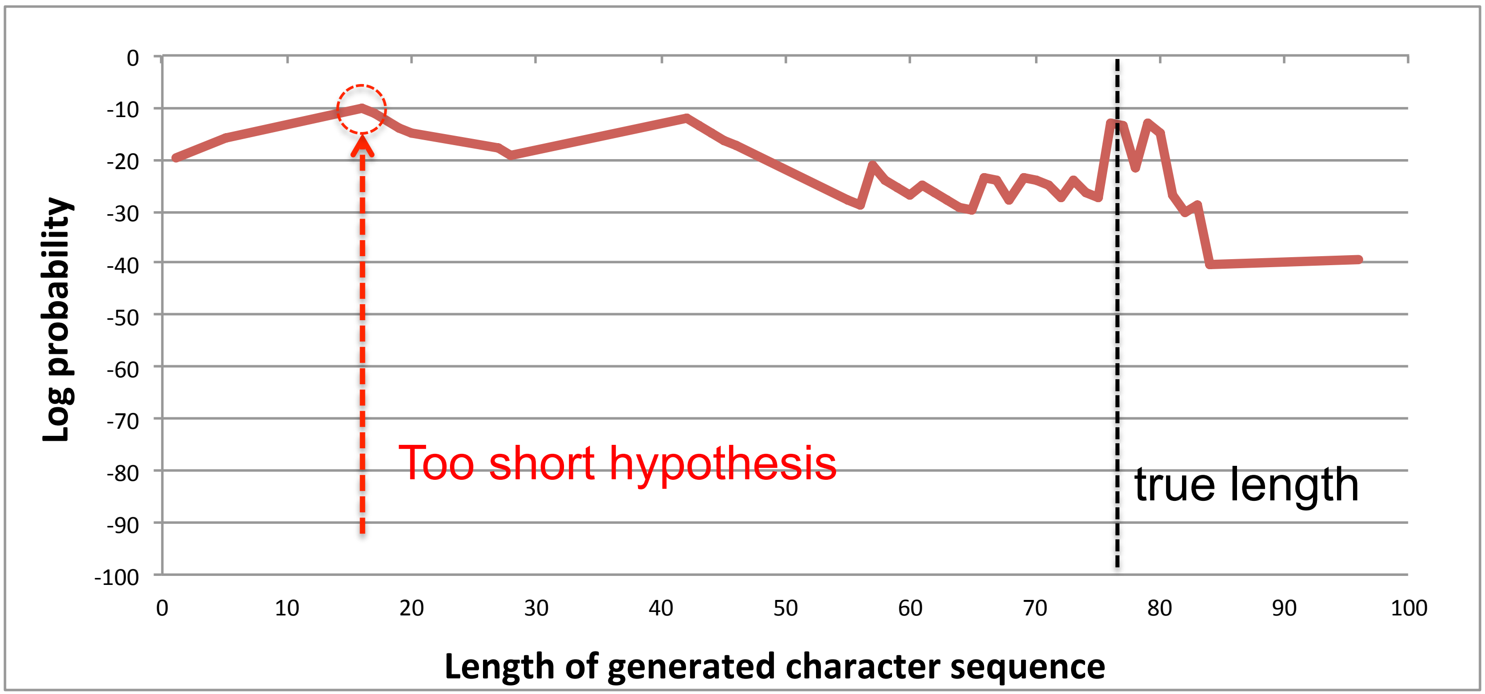
Label-synchronous forward algorithm



End detection problem

- Attention decoder fails to detect the end of sequence

Attention-based hypothesis scores

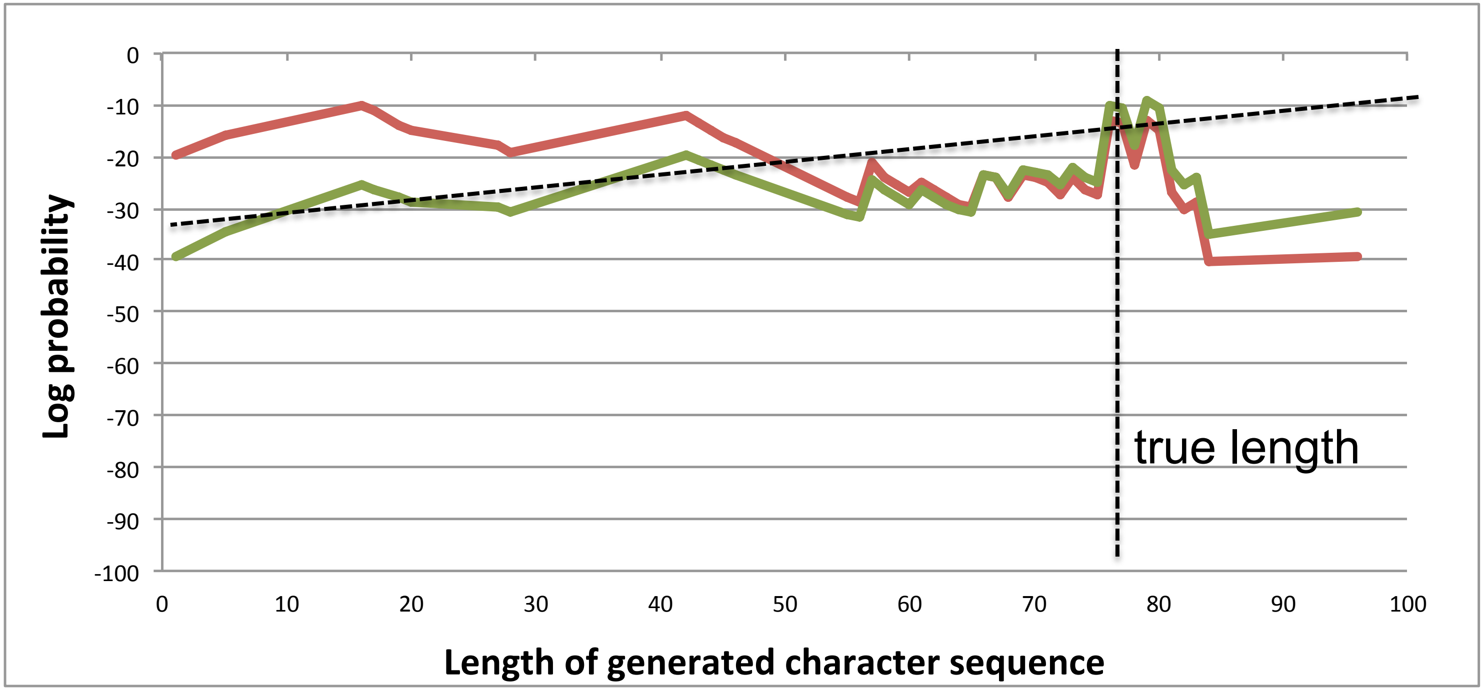


End detection problem

- Attention decoder fails to detect the end of sequence

Attention-based hypothesis scores with a length penalty

$$\alpha'_{\mathrm{att}}(h) = \alpha_{\mathrm{att}}(h) + \rho|h|$$

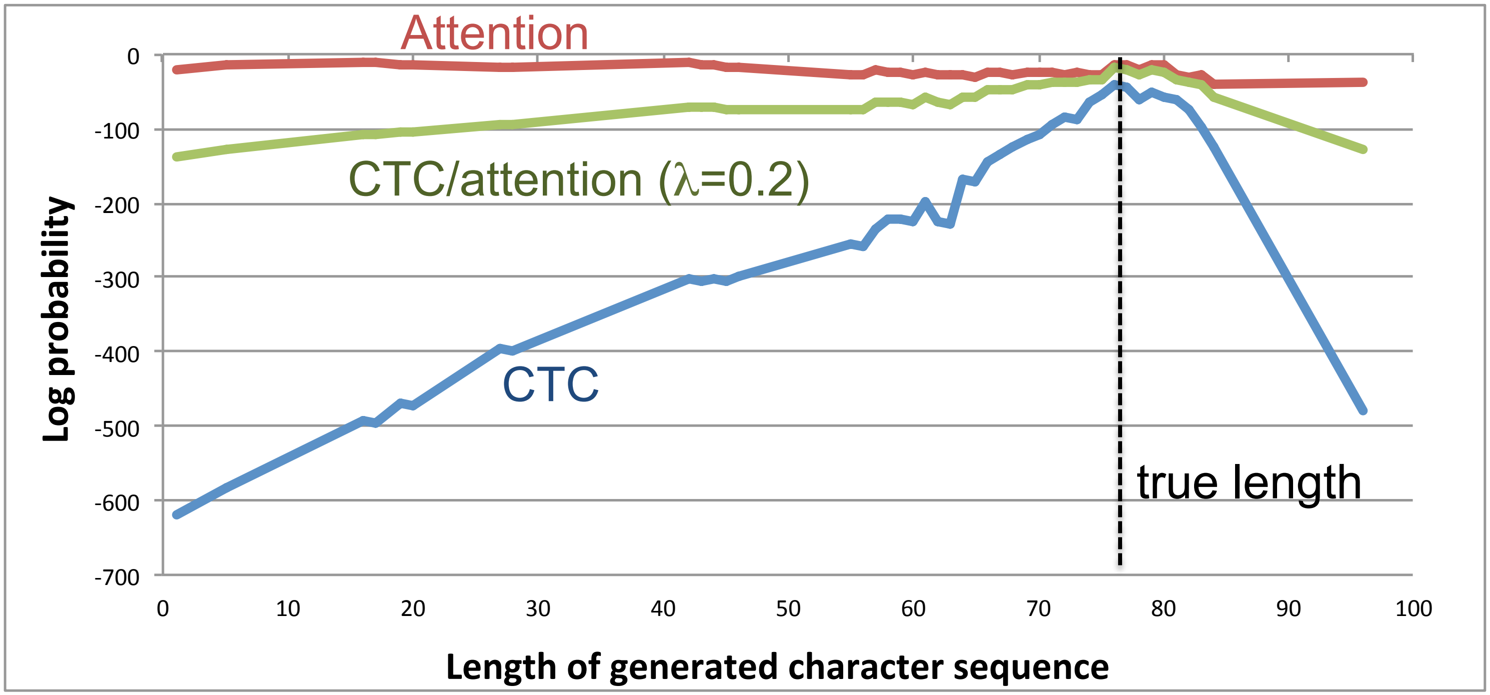


- Need to carefully tune the length penalty, max/min lengths…

End detection problem

- CTC is a good end-point estimator

CTC and CTC/attention scores



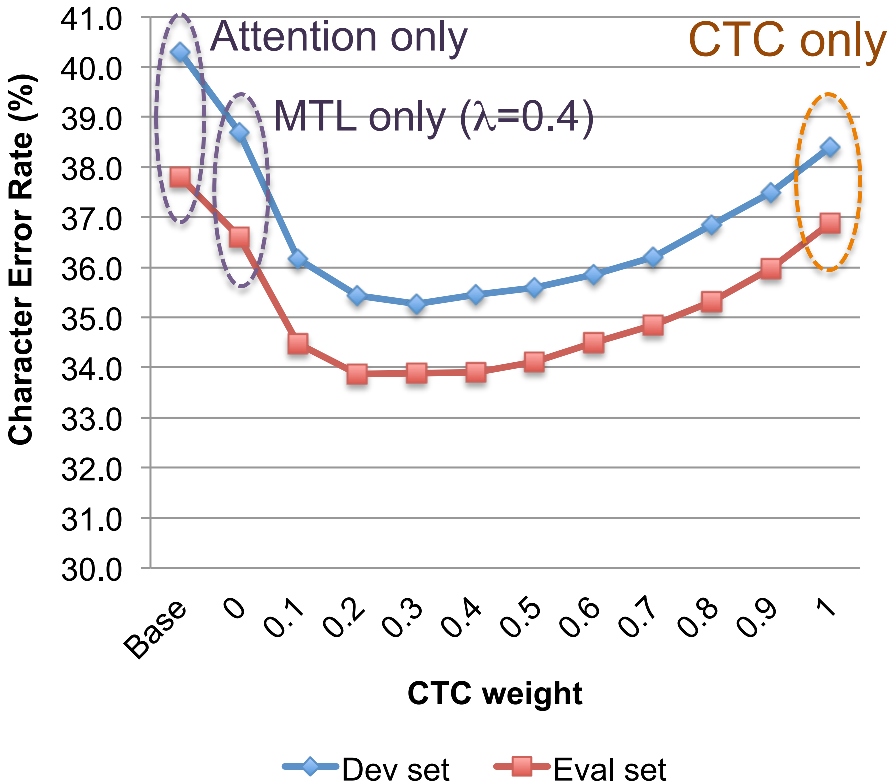
- CTC/attention decoding does not need any length control
- Can terminate decoding earlier by detecting the score peak

Experiments

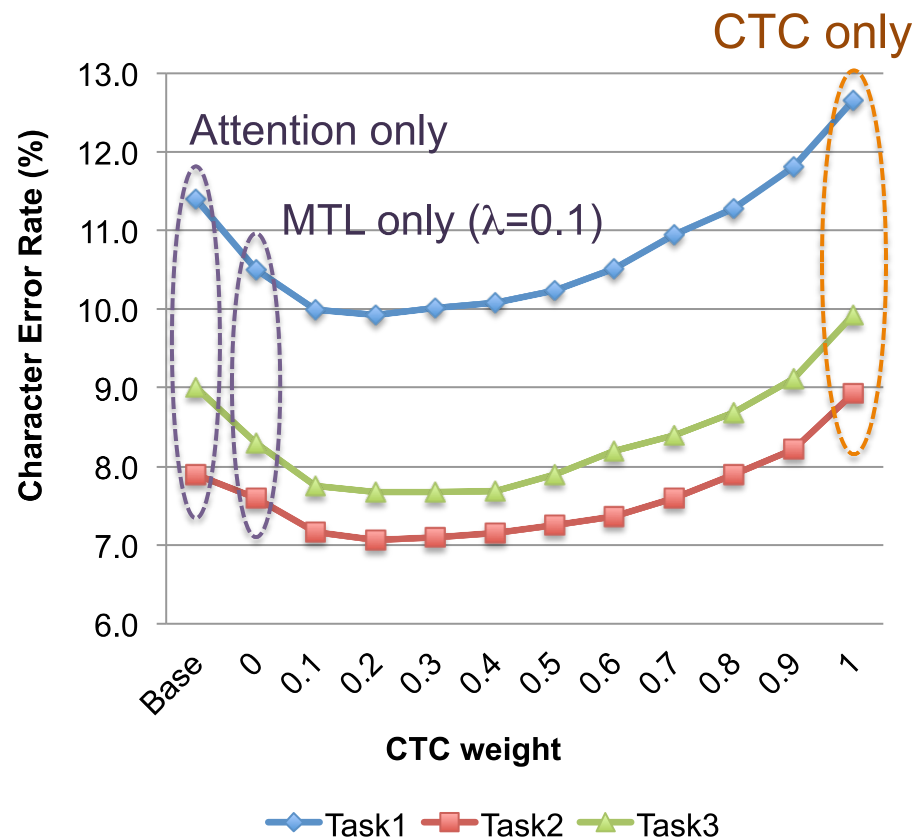
- Data sets
  - HKUST: Mandarin Chinese conversational telephone speech recognition
    - Training 167 hours, Development 4.8 hours, Evaluation 4.9 hours
    - Input feature: 80 dim. mel-filterbank + pitch feature
    - Output labels: 3653
  - CSJ: Japanese lecture speech transcription task
    - Training 581 hours, Evaluation: task1: 1.9 hours, task2: 2.0 hours, task3: 1.3 hours
    - Input feature: 40 dim. mel-filterbank + delta + delta-delta
    - Output labels: 3315
- Models
  - Encoder – 4 layer BLSTM (320 cells)
  - Decoder – 1 layer LSTM (320 cells) with location-based attention mechanism

Impact of CTC weight

- CTC weight (λ) vs. Character error rate



HKUST task



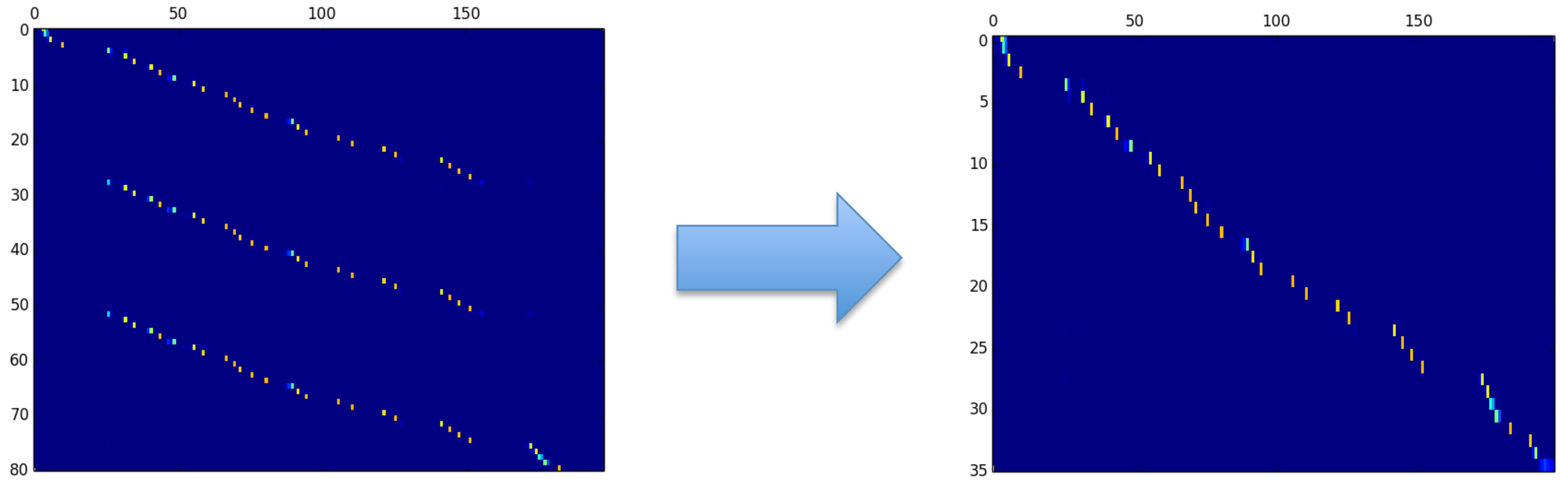
CSJ task

Example of recovering insertion errors (HKUST)

id: (20040717_152947_A010409_B010408-A-057045-057837)
Reference
但 是 如 果 你 想 想 如 果 回 到 了 过 去 你 如 果 带 着 这 个 现 在 的 记 忆 是 不 是 很 痛 苦 啊

Hybrid CTC/attention (w/o joint decoding)
Scores: (#Correctness #Substitution #Deletion #Insertion) 28 2 3 45
但 是 如 果 你 想 想 如 果 回 到 了 过 去 你 如 果 带 着 这 个 现 在 的 节 如 果 你 想 想 如 果 回 到 了 过 去 你 如 果 带 着 这 个 现 在 的 节 如 果 你 想 想 如 果 回 到 了 过 去 你 如 果 带 着 这 个 现 在 的 机 是 不 是 很 ・ ・ ・

w/ Joint decoding
Scores: (#Correctness #Substitution #Deletion #Insertion) 31 1 1 0
HYP: 但 是 如 果 你 想 想 如 果 回 到 了 过 去 你 如 果 带 着 这 个 现 在 的 ・ 机 是 不 是 很 痛 苦 啊



Example of recovering deletion errors (CSJ)

id: (A01F0001_0844951_0854386)

Reference

ま た え 飛 行 時 の エ コ ー ロ ケ ー シ ョ ン 機 能 を よ り 詳 細 に 解 明 す る 為 に 超 小 型 マ イ
ク ロ ホ ン お よ び 生 体 ア ン プ を コ ウ モ リ に 搭 載 す る こ と を 考 え て お り ま す そ う す
る こ と に よ っ て

Hybrid CTC/attention (w/o joint decoding)

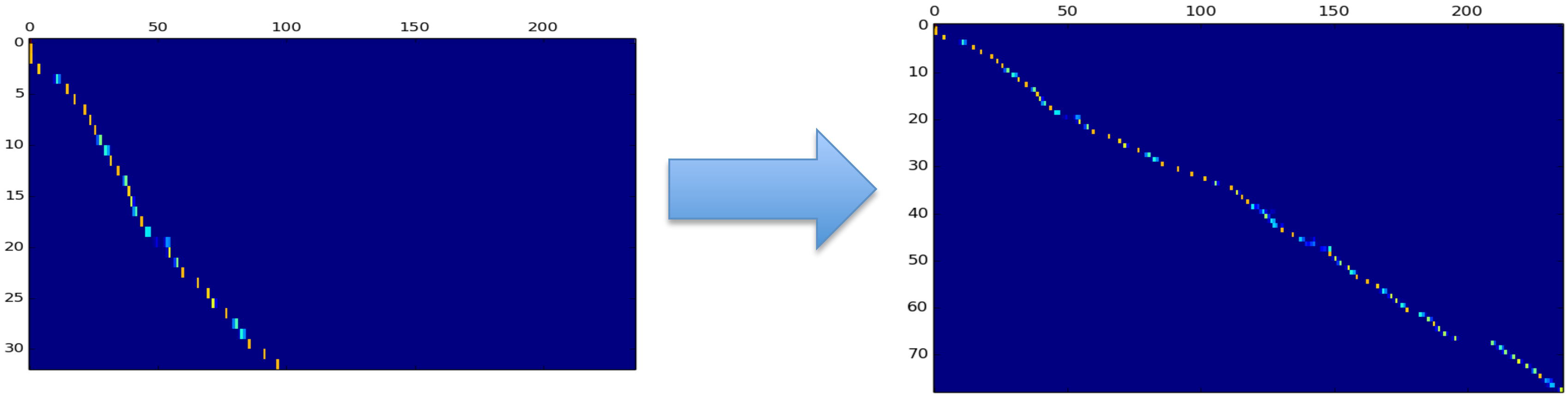
Scores: (#Correctness #Substitution #Deletion #Insertion) 30 0 47 0

ま た え 飛 行 時 の エ コ ー ロ ケ ー シ ョ ン 機 能 を よ り 詳 細 に 解 明 す る 為 ・ ・ ・ ・ ・ ・
・ ・
・ ・ ・ に ・ ・ ・

w/ Joint decoding

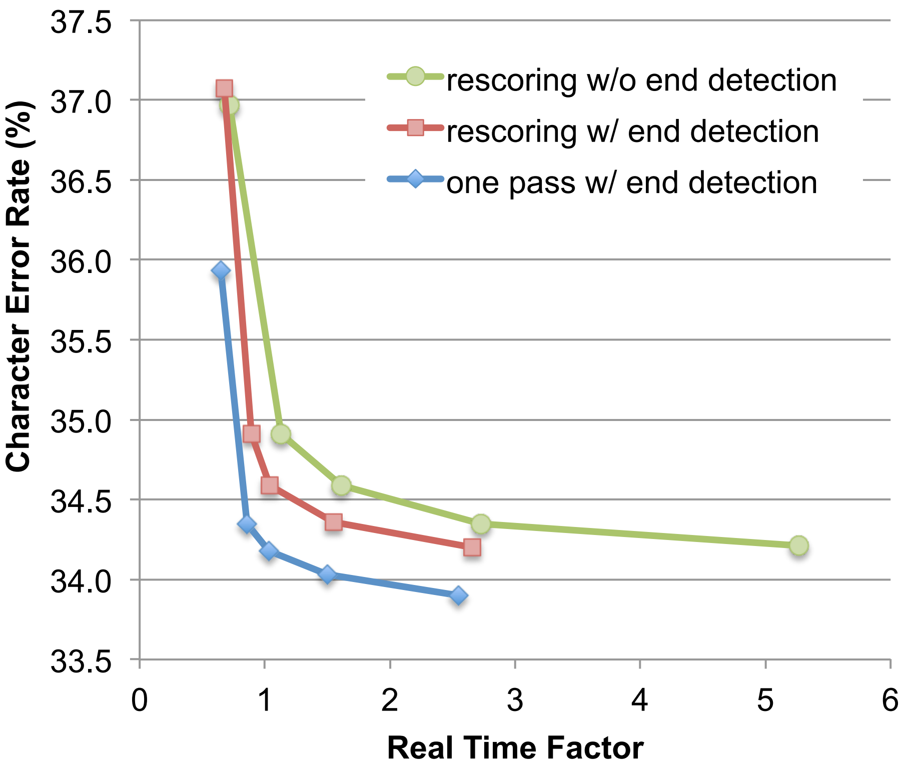
Scores: (#Correctness #Substitution #Deletion #Insertion) 67 9 1 0

ま た え 飛 行 時 の エ コ ー ロ ケ ー シ ョ ン 機 能 を よ り 詳 細 に 解 明 す る 為 に 長 国 型 マ イ
ク ロ ホ ン お ・ い く 声 単 位 方 を コ ウ モ リ に 登 載 す る こ と を 考 え て お り ま す そ う す
る こ と に よ っ て

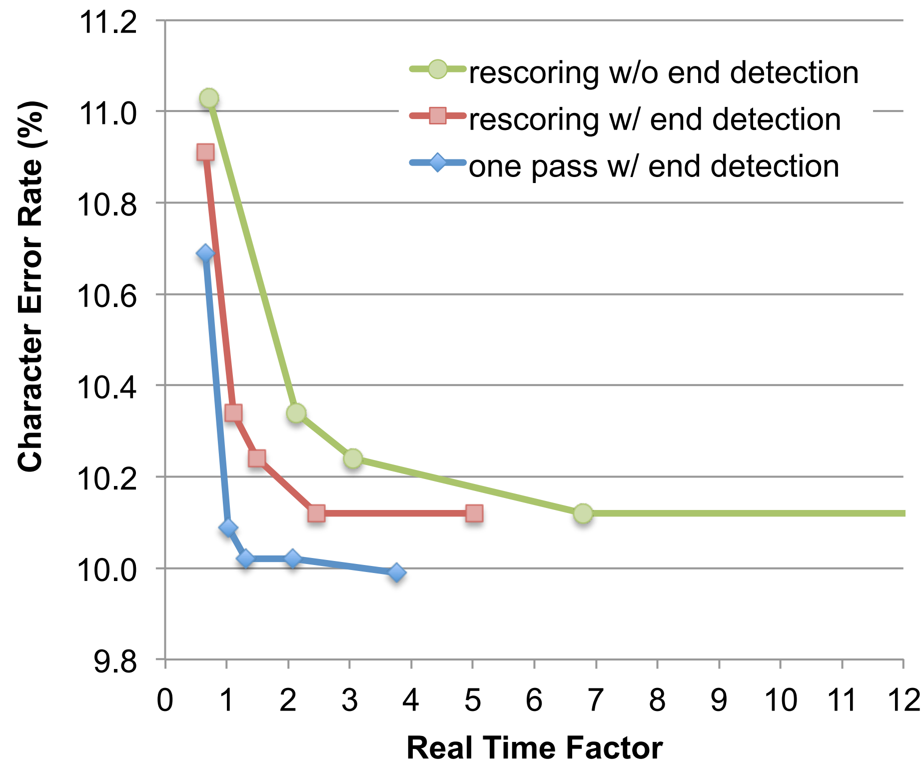


Comparison with rescoring approach

- Real-time Factor vs. Character Error Rate
  - with a single CPU (Intel(R) Xeon(R) processors, E5-2690 v3, 2.6 GHz)



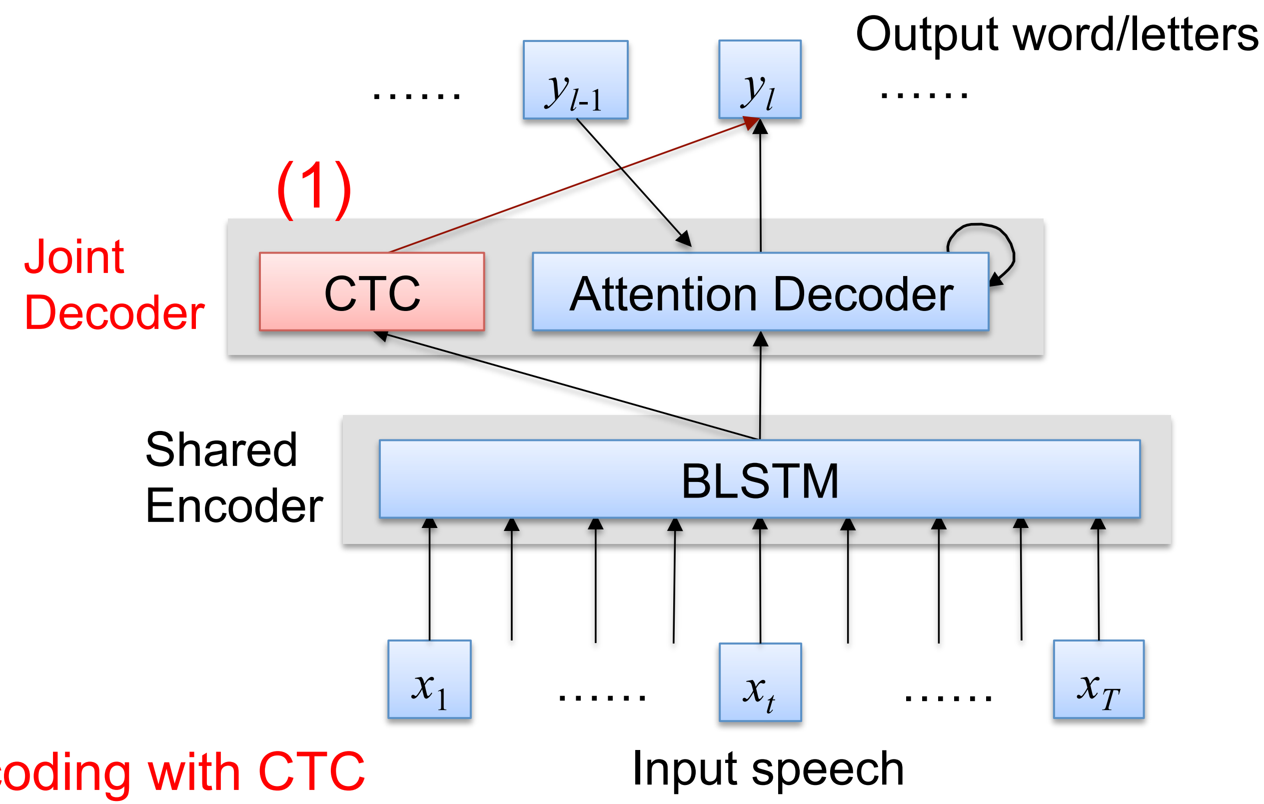
HKUST task



CSJ task

Extended CTC/attention network [Hori+'17]

(1) Connectionist Temporal Classification (CTC)

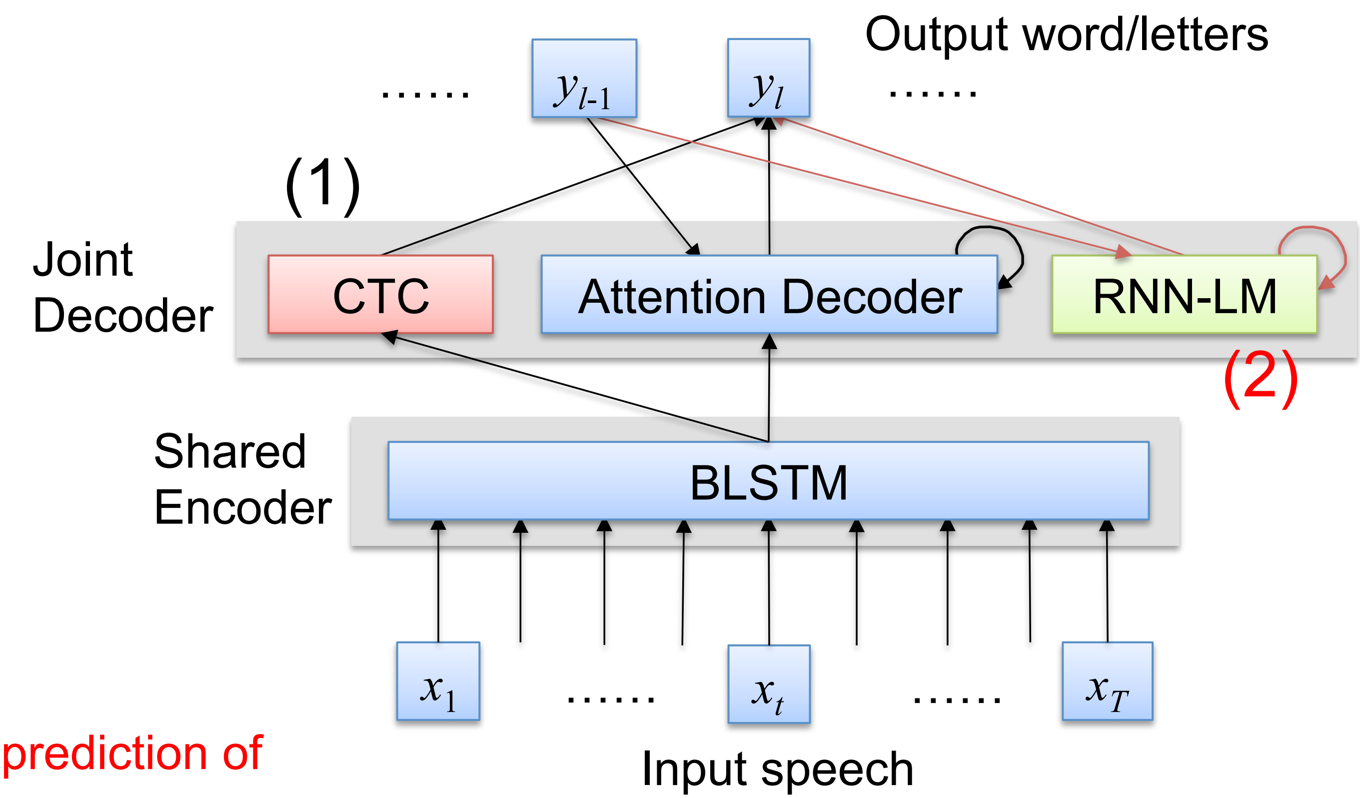


Joint training and decoding with CTC help better align input and output sequences

Extended CTC/attention network [Hori+'17]

(1) Connectionist Temporal Classification (CTC)

(2) Recurrent Neural Network Language Model (RNN-LM)



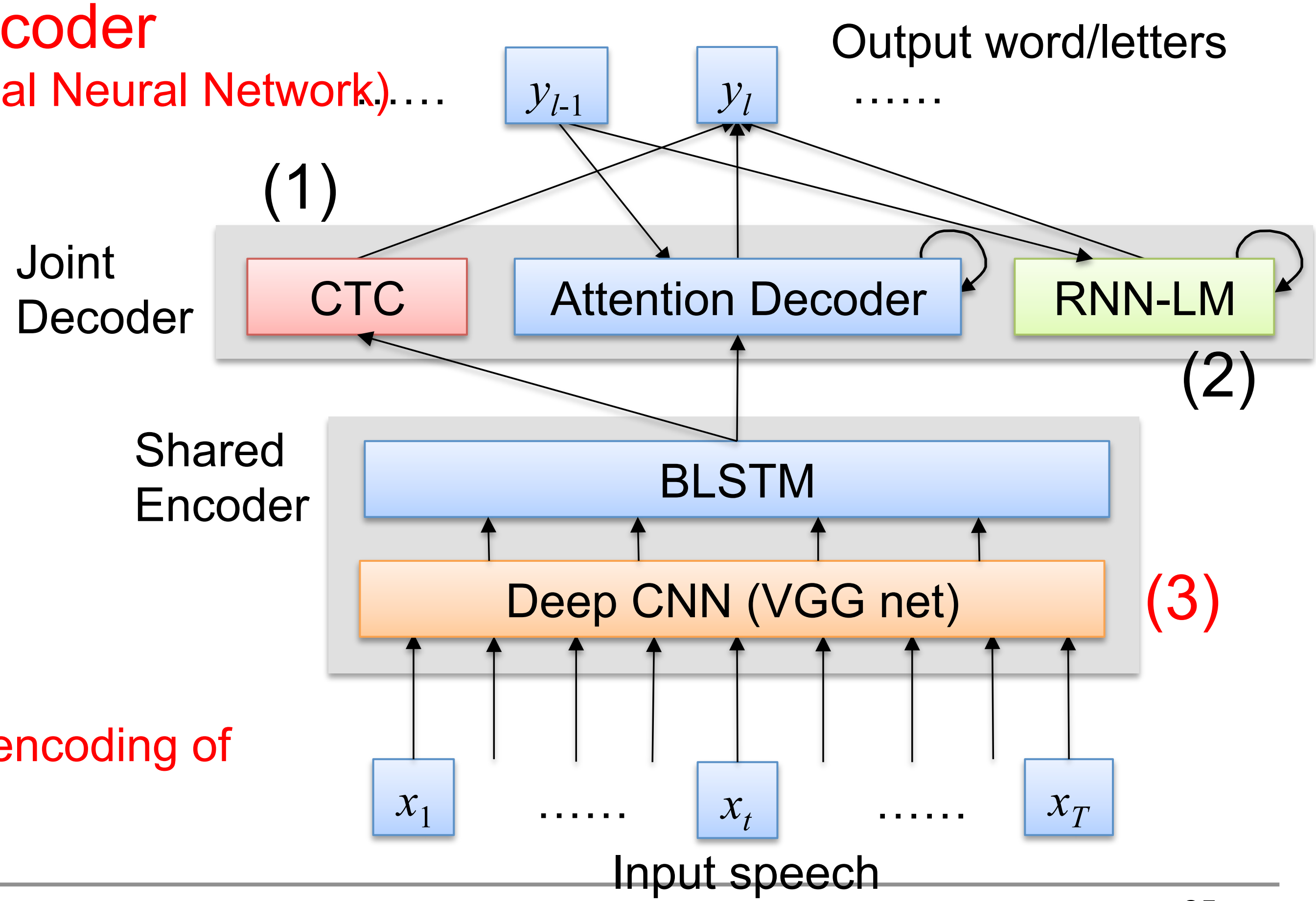
RNN-LM helps better prediction of output sequence

Extended CTC/attention network [Hori+'17]

(1) Connectionist Temporal Classification (CTC)

(2) Recurrent Neural Network Language Model (RNN-LM)

(3) Deep CNN encoder
(CNN: Convolutional Neural Network)



Deep CNN enhances encoding of input speech signals

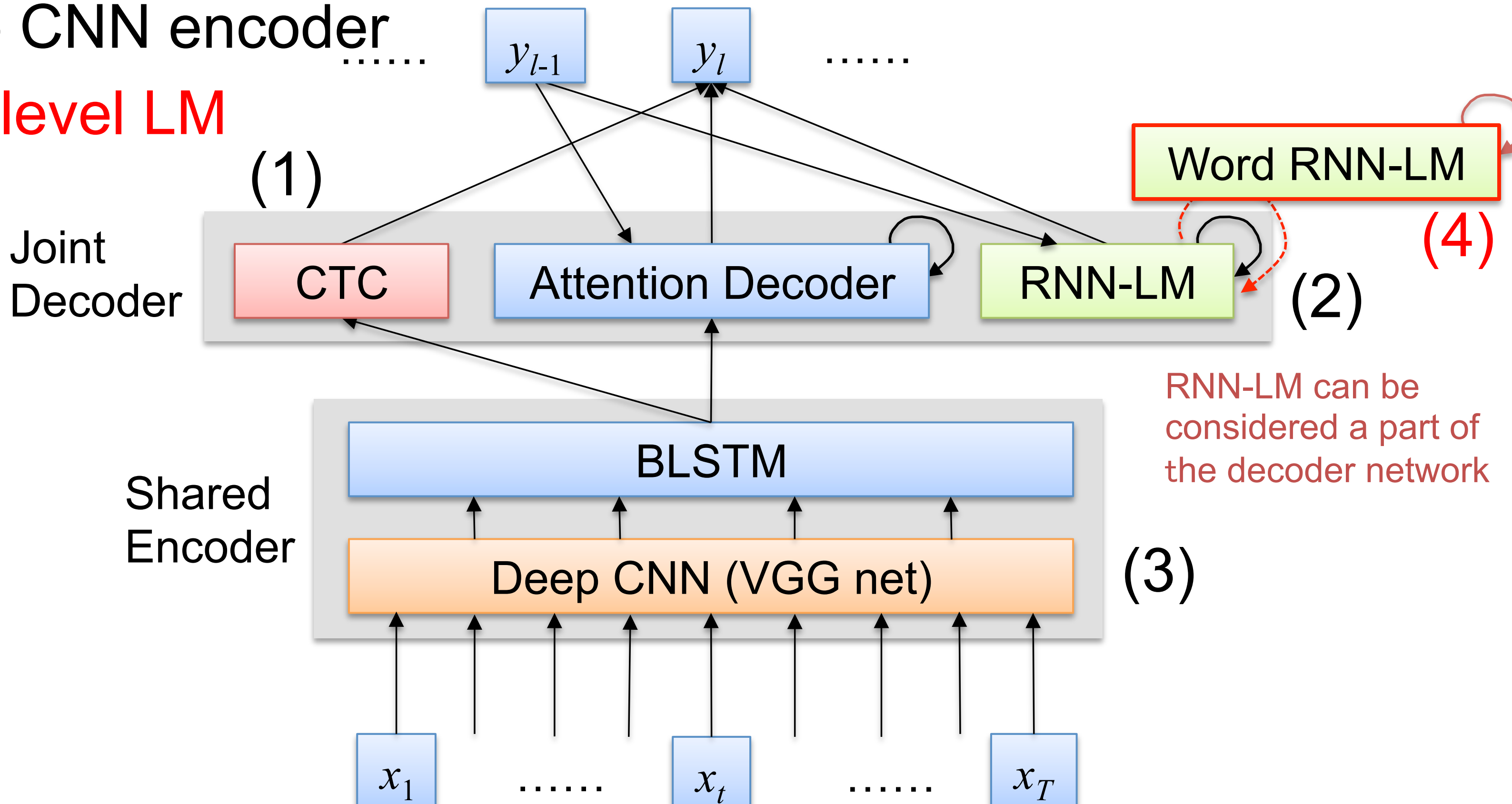
Extended CTC/attention network [Hori+'17]

(1) Connectionist Temporal Classification (CTC)

(2) Recurrent Neural Network Language Model (RNN-LM)

(3) Deep CNN encoder

(4) Multi-level LM



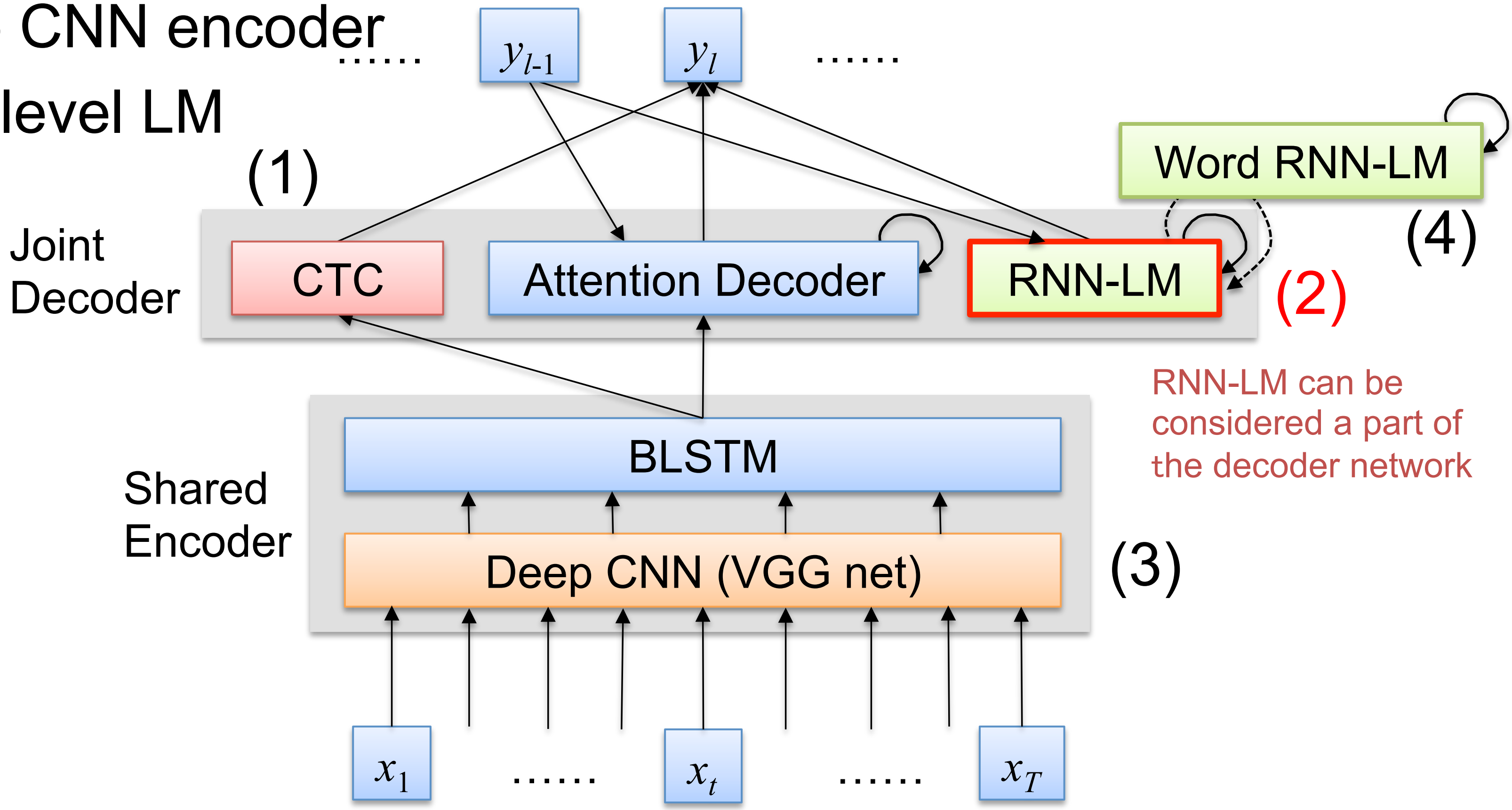
Extended CTC/attention network [Hori+'17]

(1) Connectionist Temporal Classification (CTC)

(2) Recurrent Neural Network Language Model (RNN-LM)

(3) Deep CNN encoder

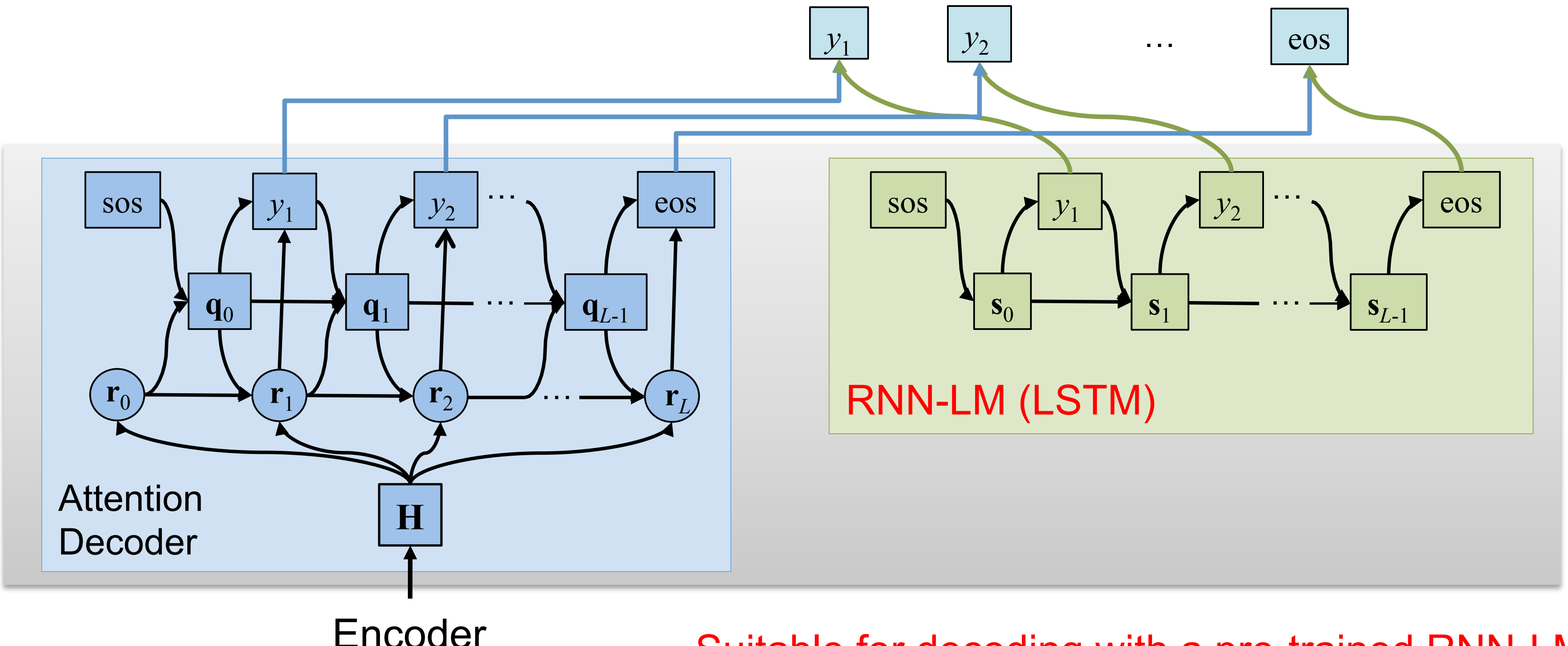
(4) Multi-level LM



(2) RNN-LM integration

- Log-probability combination with interpolation weight γ

$$\log p'_{\mathrm{att}}(y_l|y_1,\ldots,y_{l-1},X) = \gamma \log \boxed{p_{\mathrm{att}}(y_l|y_1,\ldots,y_{l-1},X)} + (1-\gamma)\log \boxed{p_{\mathrm{lm}}(y_l|y_1,\ldots,y_{l-1})}$$

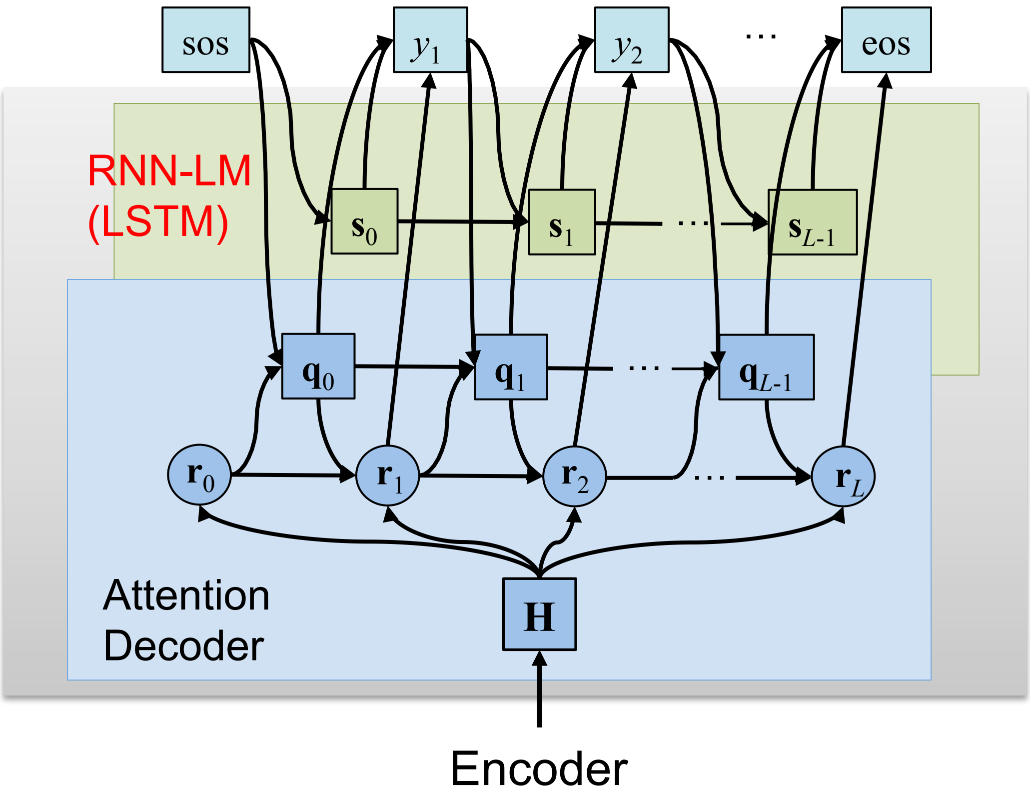


Suitable for decoding with a pre-trained RNN-LM

(2) RNN-LM integration

- Logit-level combination without interpolation weights

$$p''_{\text{att}}(y_l|y_1,\ldots,y_{l-1},X) = \text{softmax}\left(\underbrace{\mathbf{W}^{(o)}_{\text{att}}[\mathbf{q}^T_{l-1},\mathbf{r}^T_l]^T + \mathbf{b}^{(o)}_{\text{att}}}_{\text{Attention decoder}} + \underbrace{\mathbf{W}^{(o)}_{\text{lm}}\mathbf{s}_{l-1} + \mathbf{b}^{(o)}_{\text{lm}}}_{\text{RNN-LM}}\right)[y_l]$$



Logits of two networks can be balanced well by joint training

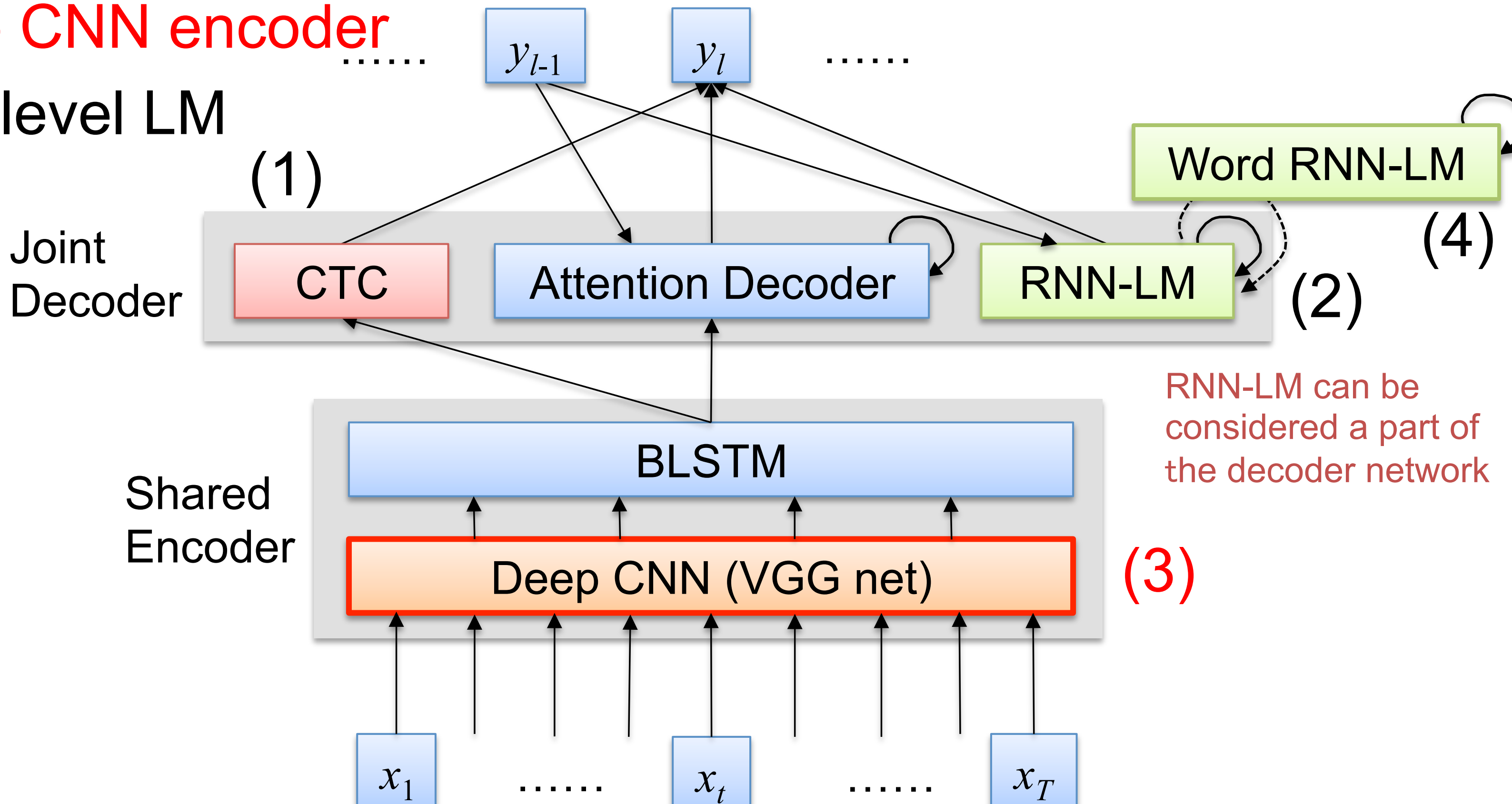
Extended CTC/attention network [Hori+'17]

(1) Connectionist Temporal Classification (CTC)

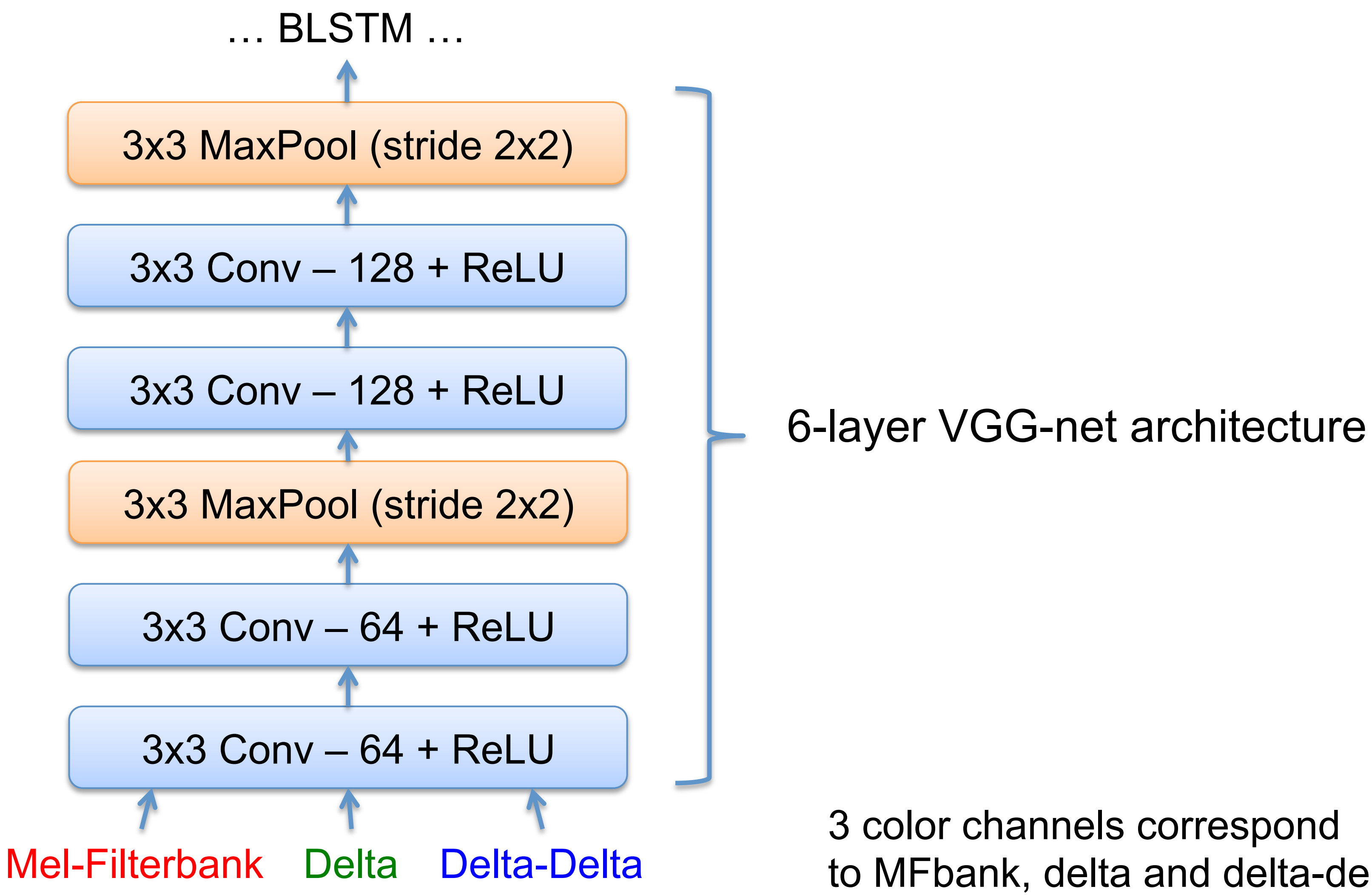
(2) Recurrent Neural Network Language Model (RNN-LM)

(3) Deep CNN encoder

(4) Multi-level LM

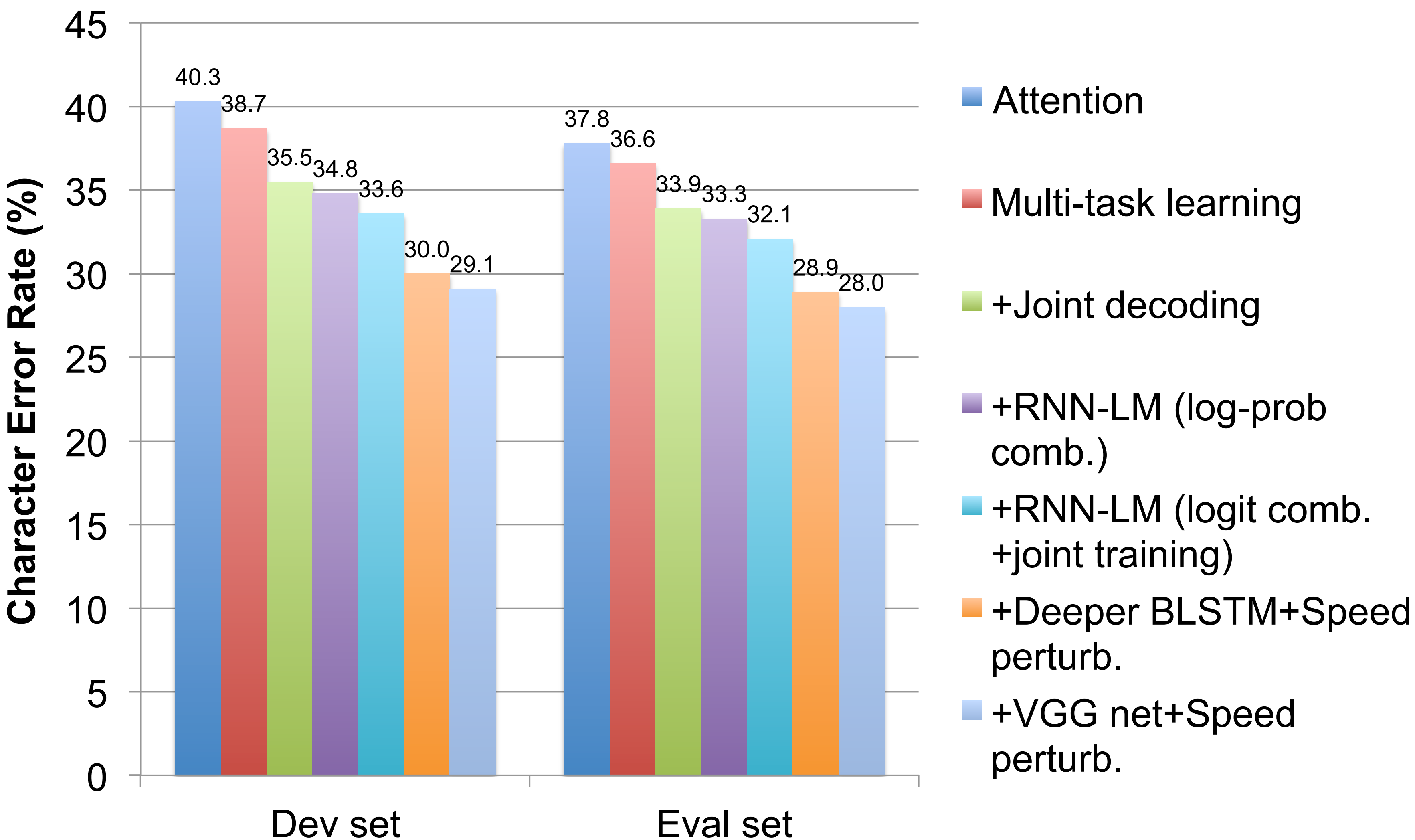


(3) Deep CNN – VGG Net [Simonyan+'14]

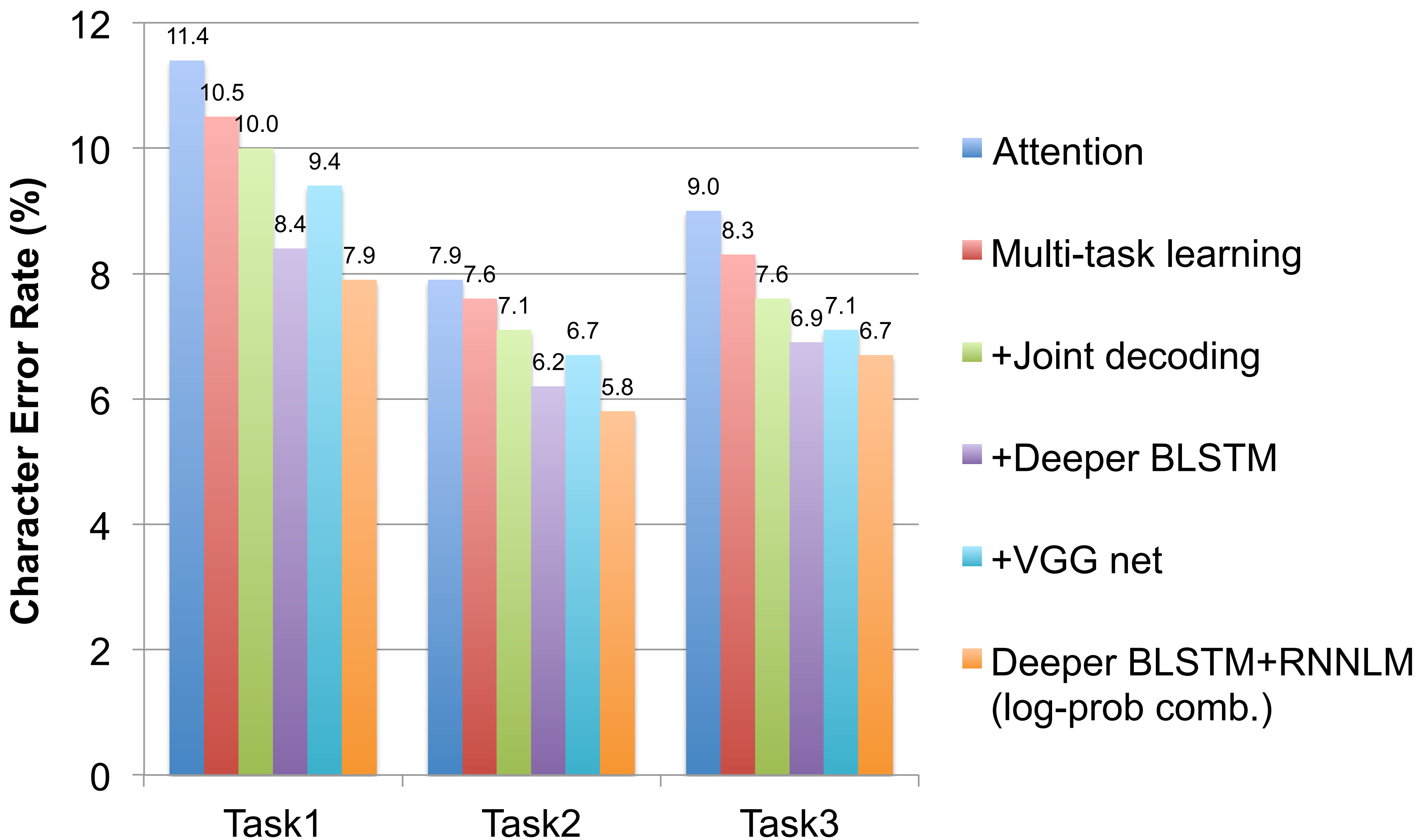


3 color channels correspond to MFbank, delta and delta-delta

Effect of extended models in HKUST task



Effect of extended models in CSJ task



Comparison with conventional ASR systems

- Character Error Rate (%) in HKUST task

Models	Dev.	Eval
Our best model	29.1	28.0
DNN/HMM	-	35.9
LSTM/HMM + speed perturb.	-	33.5
CTC with language model (Miao et al. 2016)	-	34.8
TDNN/HMM, lattice-free MMI + speed perturb. (Povey et al., 2016)	-	28.2

- Character Error Rate (%) in CSJ task

Models	Task 1	Task 2	Task 3
Our best model	7.9	5.8	6.7
DNN/HMM (Moriya et al., 2015)	9.0	7.2	9.6
CTC-syllable (Kanda et al., 2016)	9.4	7.3	7.5

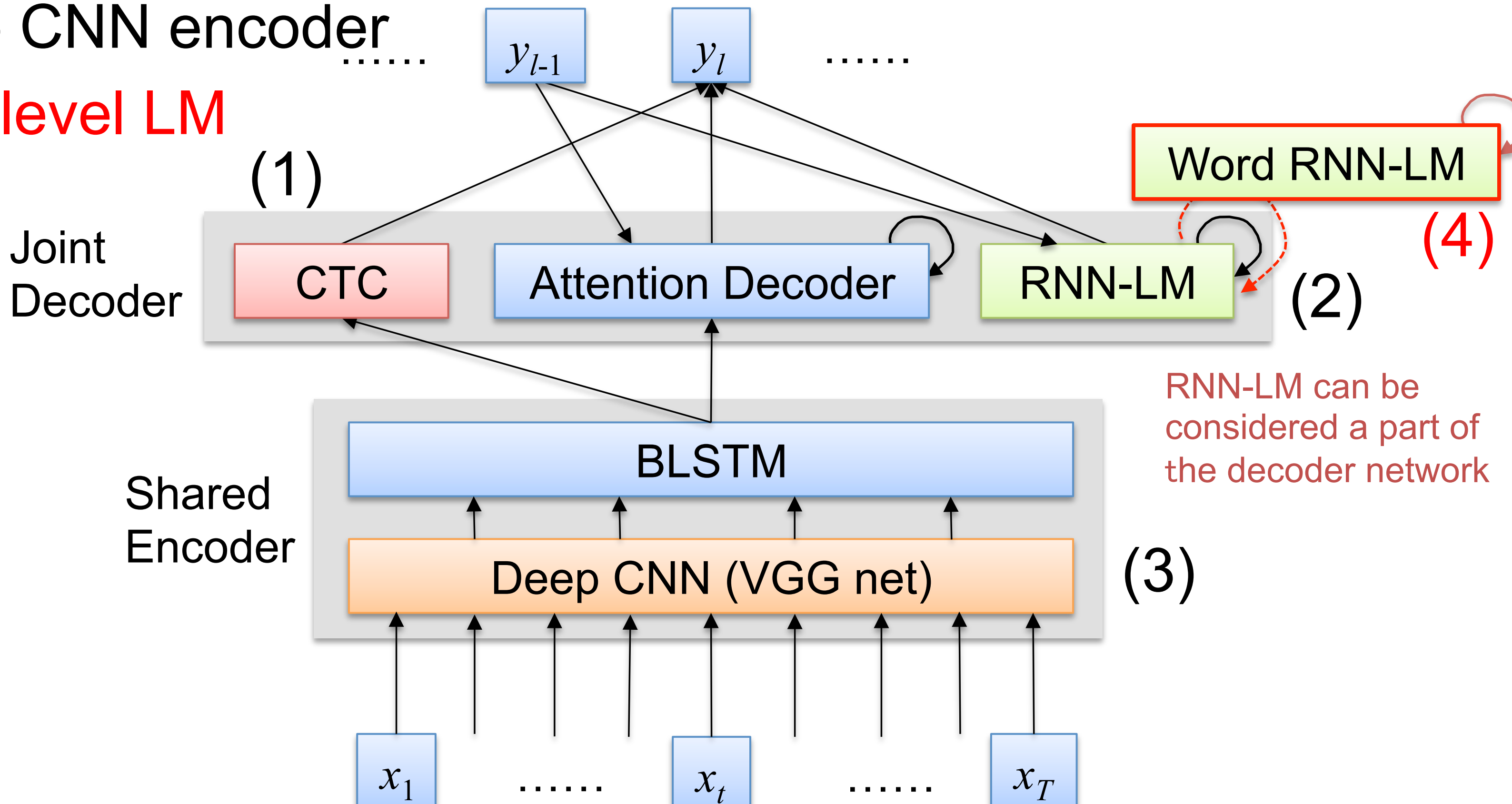
Extended CTC/attention network [Hori+'17]

(1) Connectionist Temporal Classification (CTC)

(2) Recurrent Neural Network Language Model (RNN-LM)

(3) Deep CNN encoder

(4) Multi-level LM



Problem of character-based prediction

- End-to-end ASR is usually designed to generate character sequences
  - No explicit word boundaries in some languages
  - Training acoustic-to-word mapping is hard (need huge data)
- General approaches
  - N-gram LM+ WFST approach [Miao+ 2015, Chorowsky+ 2015]
    - Difficult to incorporate RNN-LMs
    - No treatment for out-of-vocabulary (OOV) words
  - Character-based RNN-LM [Hori+ 2017]
    - Can perform open vocabulary ASR
    - Character LMs under-performs word LMs if large text corpus is available

Our approach

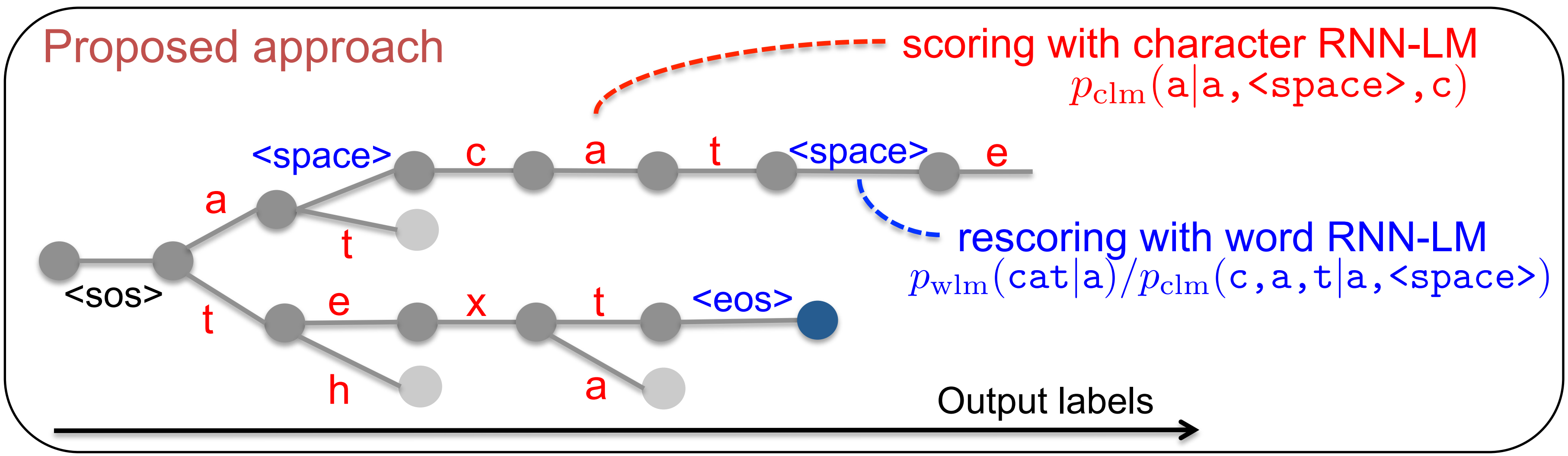
Multi-level LMs to incorporate word-based RNN-LMs while keeping open-vocabulary ASR

Basic concept

- Decoding with character-level LM scores

$$\hat{C} = \arg\max_{C \in \mathcal{U}^*} \{\lambda \log p_{\mathrm{ctc}}(C|X) + (1-\lambda)\log p_{\mathrm{att}}(C|X)$$
$$+ \gamma \log \underline{p_{\mathrm{lm}}(C)}\}$$

- Apply LM scores at both the character and word levels



Character-based probability using word-based RNN-LM

$$p_{\mathrm{lm}}(c|g) = \begin{cases} \dfrac{p_{\mathrm{wlm}}(w_g|\psi_g)}{p_{\mathrm{clm}}(w_g|\psi_g)} & \text{if } c \in S, w_g \in \mathcal{V} \\ p_{\mathrm{wlm}}(\texttt{<UNK>}|\psi_g)\tilde{\beta} & \text{if } c \in S, w_g \notin \mathcal{V} \\ p_{\mathrm{clm}}(c|g) & \text{otherwise} \end{cases}$$

S : set of word boundary characters {<space>,<eos>,...}

$\mathcal{V}$: vocabulary of word LM

g : hypothesis a,<space>,c,a,t,<space>,e,a,t,s

w_g : last word of g eats

ψ_g : history of w_g a,cat

$\tilde{\beta}$: adjustment term for OOVs

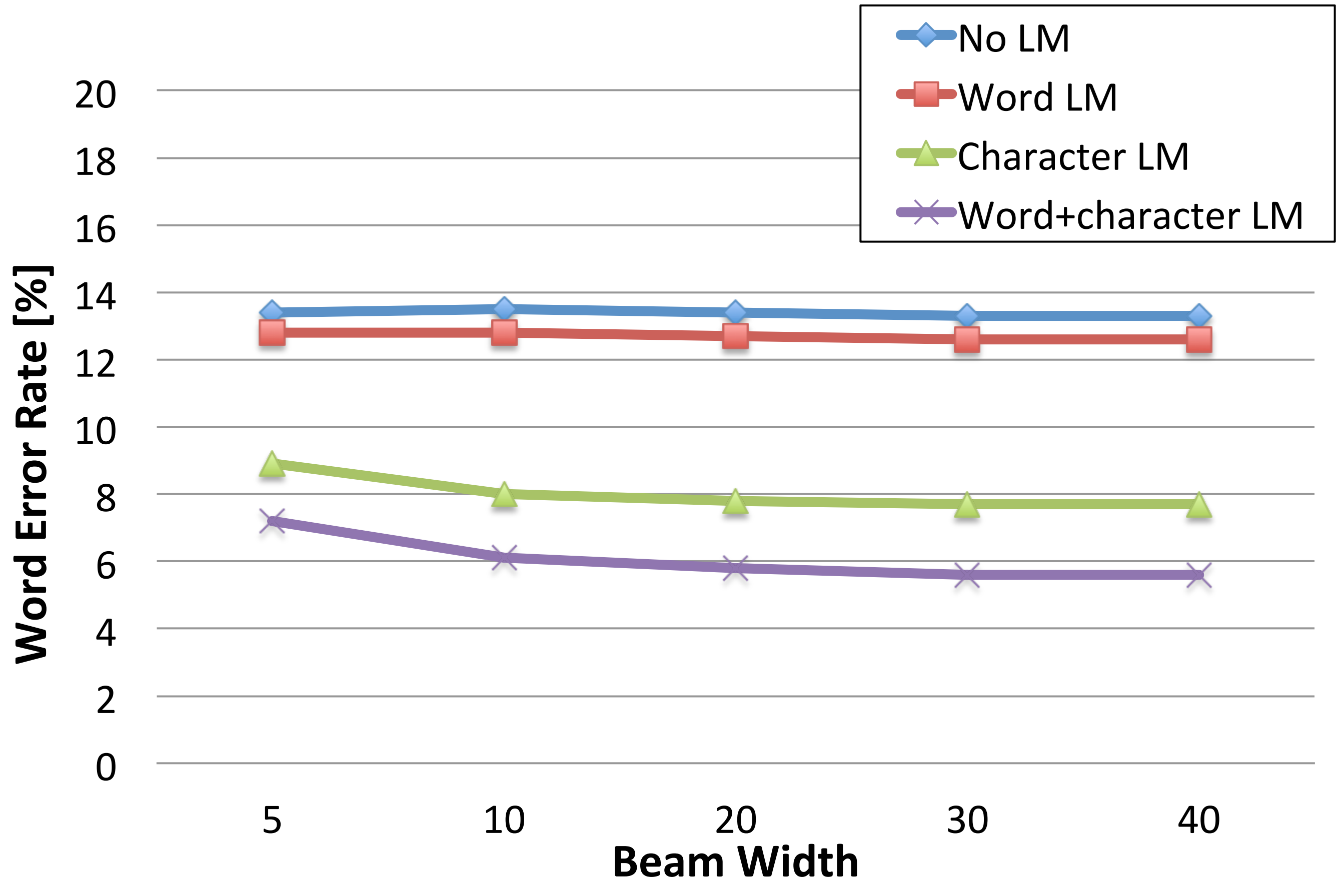
$$p_{\mathrm{wlm}}(w_{\mathrm{oov}}|\psi_g) = p_{\mathrm{wlm}}(\texttt{<UNK>}|\psi_g)p_{\mathrm{clm}}(w_{\mathrm{oov}}|\texttt{<UNK>}, \psi_g)$$

$$p_{\mathrm{clm}}(w_{\mathrm{oov}}|\texttt{<UNK>}, \psi_g) \propto p_{\mathrm{clm}}(w_{\mathrm{oov}}|\psi_g) \rightarrow \tilde{\beta}$$

Experiments

- Wall Street Journal (WSJ) corpus
  - Training: 80 hours (SI284), Development: 1.1 hours (dev93), Evaluation: 0.7 hours (eval92)
  - Input: 80 dim. mel-filterbank + pitch feature (+d, +dd)
  - Output: 32 distinct labels (26 char + apostrophe, period, …, <space>, <sos>/<eos>)
- Models
  - Encoder: 6-layer CNN + 4-layer BLSTM (320 cells)
  - Decoder: 1-layer LSTM (320 cells) with location-based attention mechanism
  - RNN-LMs: 1-layer LSTM (1000 cells), trained with WSJ text Vocabulary size of word LM: 20,000

Effect of language models



– No error reduction with only word LM even if increasing the beam width
– Character LM helps find better hypotheses for word LM

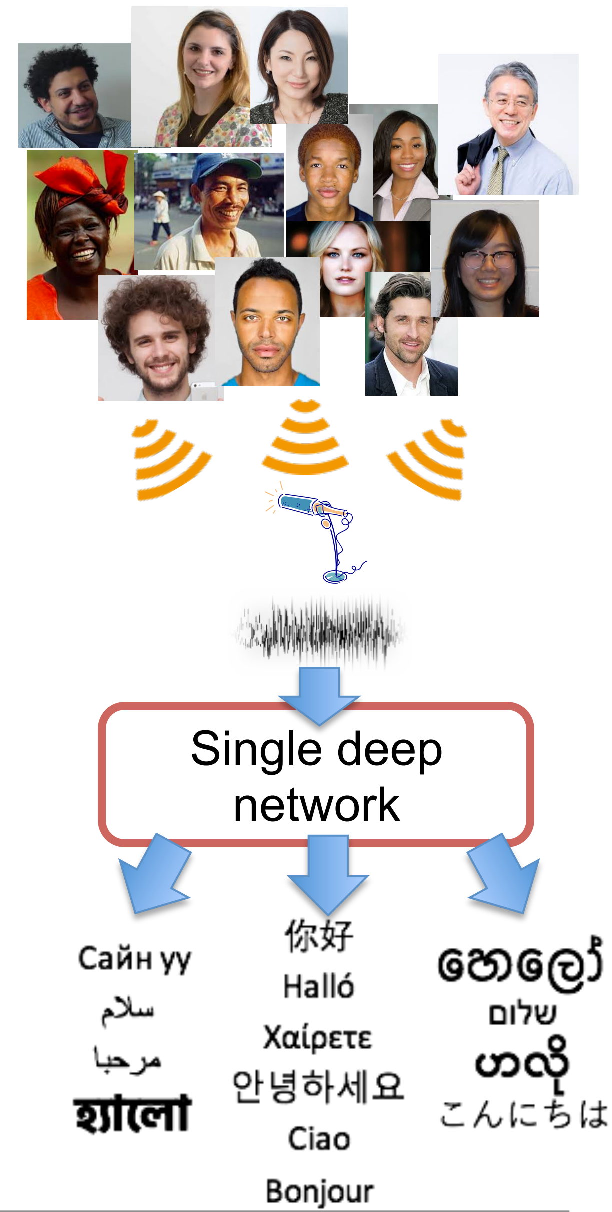
Comparison with other approaches

Models	dev93	eval92
Attention model + word 3-gram LM [Bahdanau 2016]	-	9.3
CTC + word 3-gram LM [Graves 2014]	-	8.2
CTC + word 3-gram LM [Miao 2015]	-	7.3
Attention model + word 3-gram LM [Chorowski 2016]	9.7	6.7
This work	9.6	5.6

HMM/DNN + sMBR + word 3-gram LM	6.4	3.6
HMM/DNN + sMBR + word RNN-LM	5.6	2.6

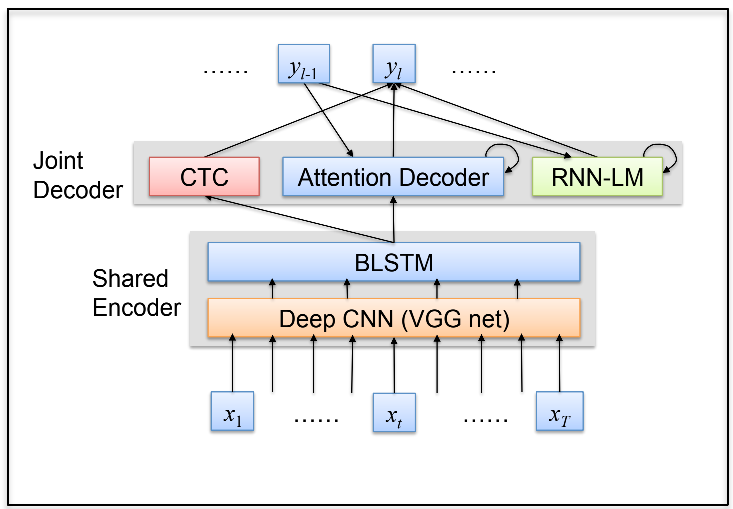
Toward multi-lingual multi-speaker end-to-end speech recognition

- End-to-end direct optimization extends the scope of potential application use cases by training the model for multiple objectives.
- Aim at single system encompassing multi-source separation and understanding
- Investigate the capability of multi-lingual multi-speaker end-to-end speech recognition



Multi-lingual end-to-end speech recognition

- Monolithic end-to-end multi-lingual ASR system
  - Build a simple, robust model without expert knowledge.

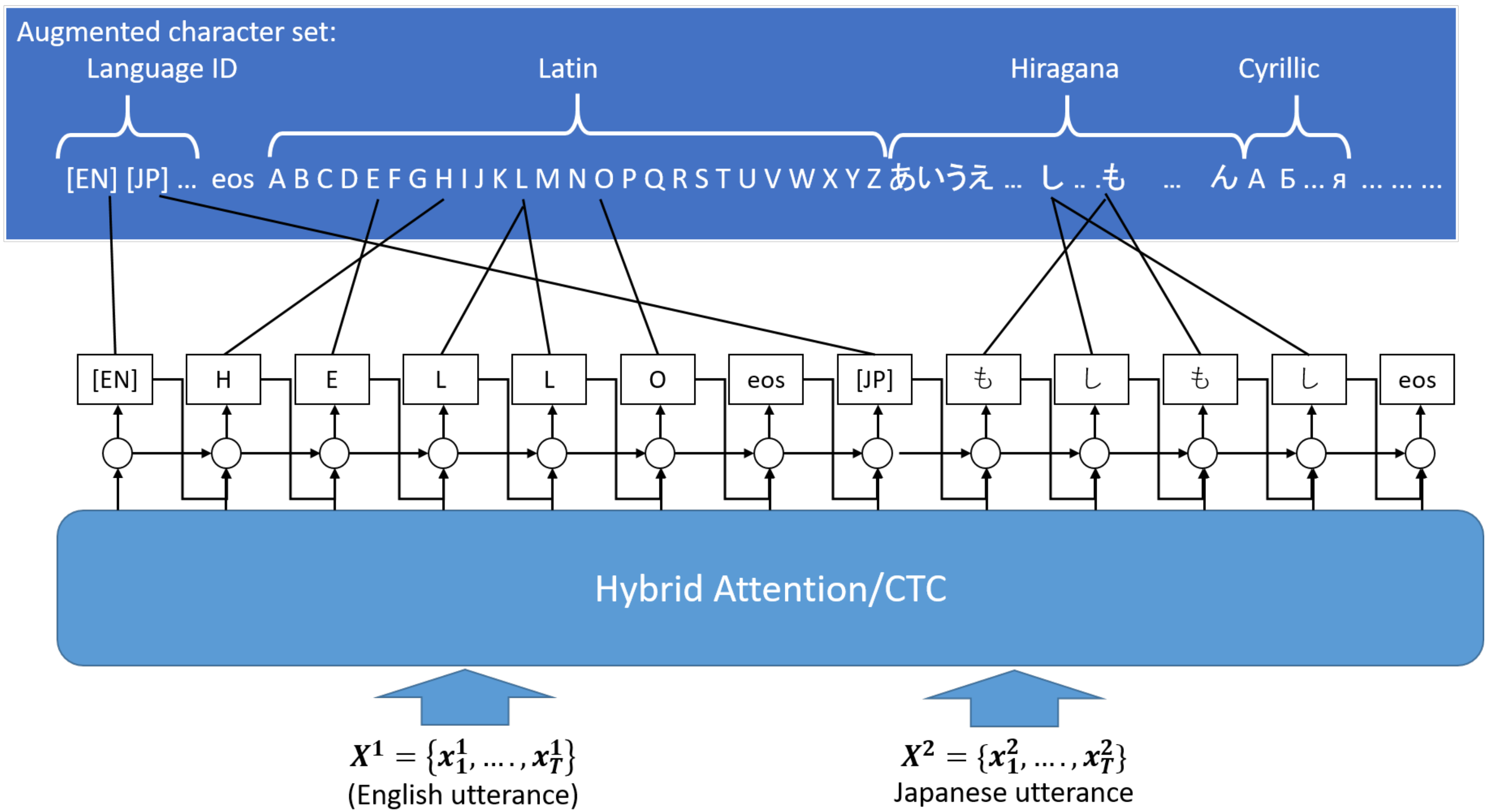


Single Deep Network

හෙලෝ
שלום
ဟလို
こんにちは
你好
Halló
Χαίρετε
안녕하세요
Ciao
Bonjour
Сайн уу
سلام
مرحبا
ह्यालो

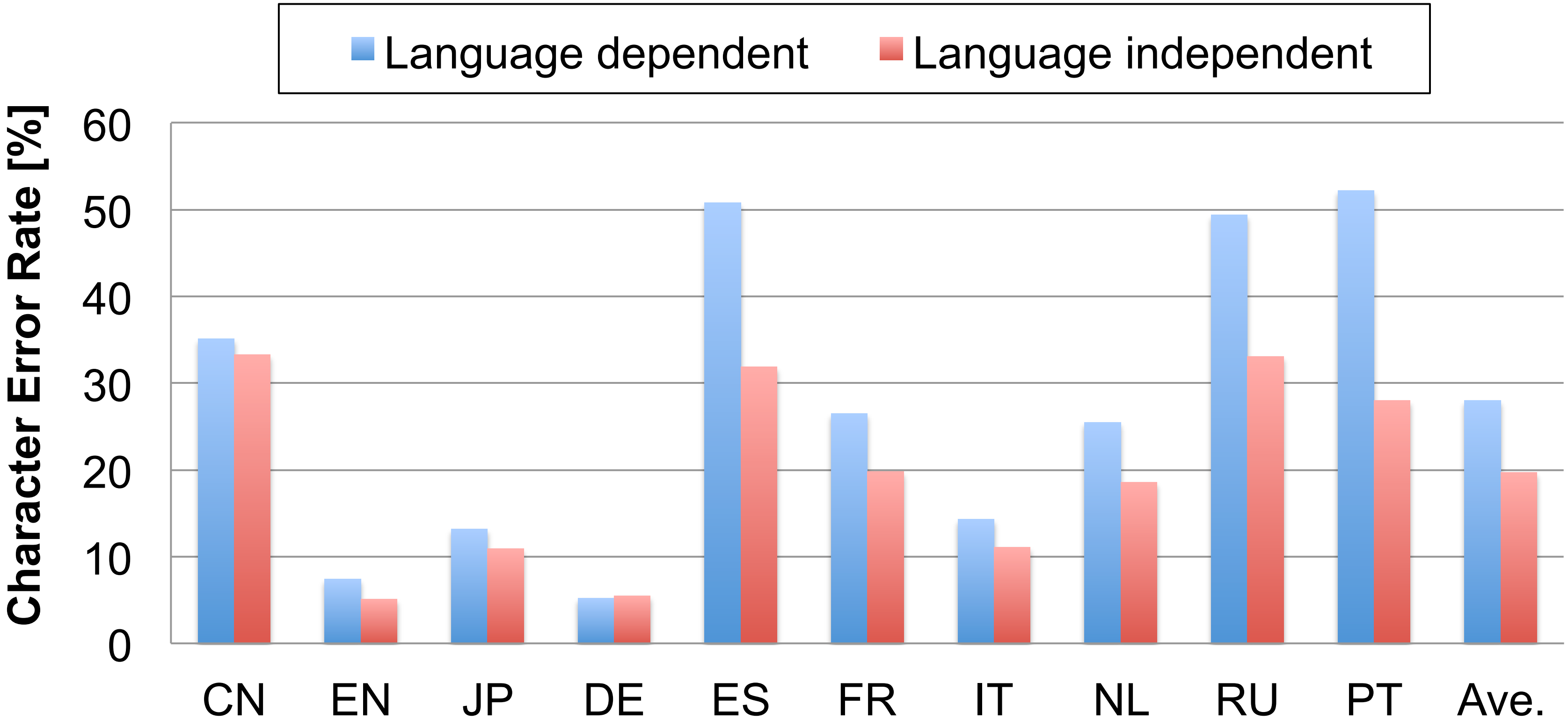
Multi-lingual end-to-end speech recognition [Watanabe+'17, Seki+'18]

- Learn a single model with multi-language data (10 languages)
- Joint language identification and speech recognition



ASR performance for 10 languages

- Comparison with language dependent systems
- One language per utterance (w/o code switching)



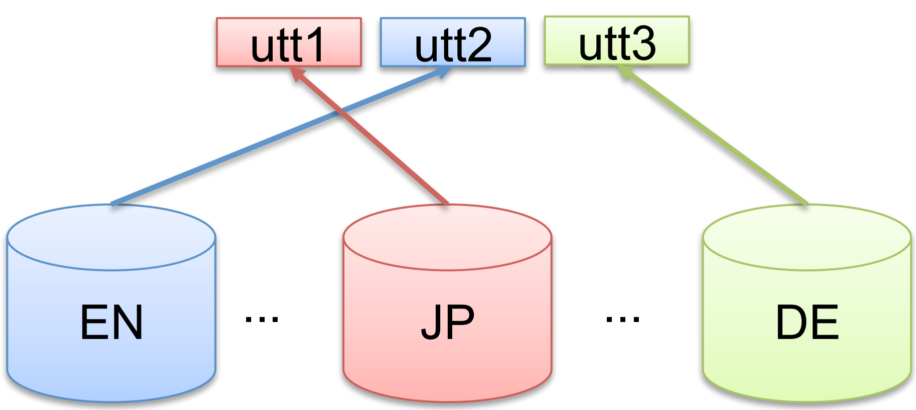
(trained with CSJ, HKUST, WSJ, Voxforge corpora)

Data generation for multi-lingual code-switching speech

Concatenation of utterances from 10 language corpora

1) Select number to concat. (1, 2, or 3)
2) Sample language and utterance:
   - P(lang): proportional to corpus duration w/ flooring
   - P(utt): uniform distribution
3) Repeat generation to reach the duration of the original corpora

Code-switching speech:

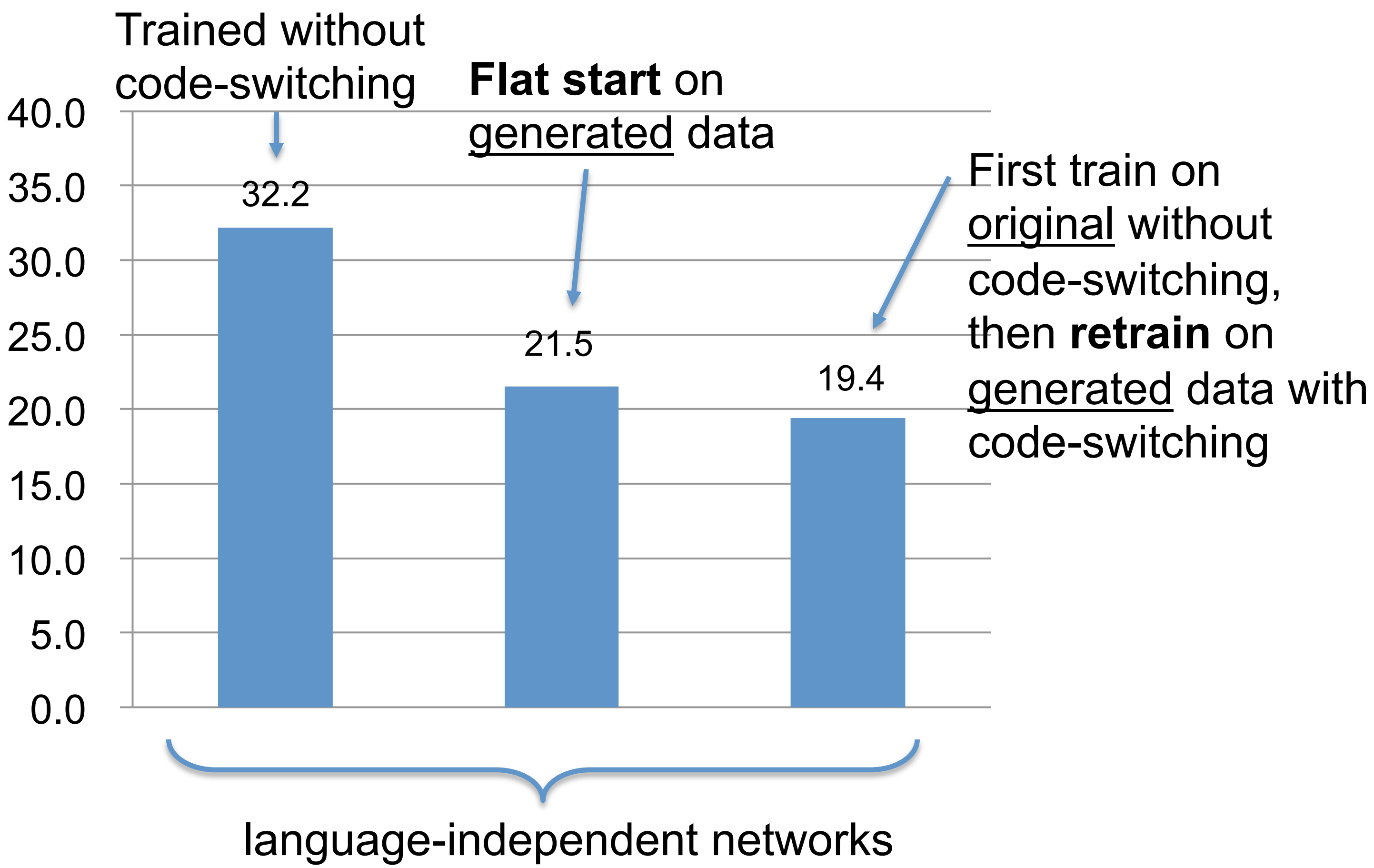


Duration (hours) of training data.

Corpus		Original	Generated
WSJ	English	81.5	87.4
CSJ	Japanese	216.3	149.1
HKUST	Mandarin	170.1	114.9
	German	45.7	64.6
	Spanish	40.3	61.6
	French	29.6	57.9
Voxforge	Italian	15.8	35.7
	Dutch	8.4	23.6
	Portuguese	3.0	9.0
	Russian	12.0	18.5
Total		622.7	622.3

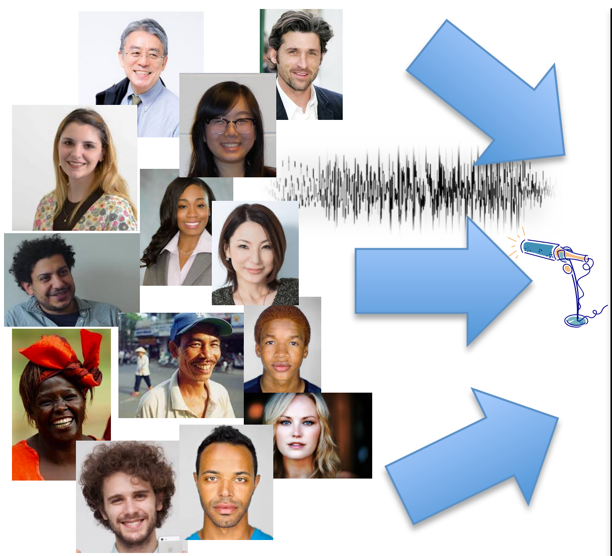
Recognition of speech with code-switching

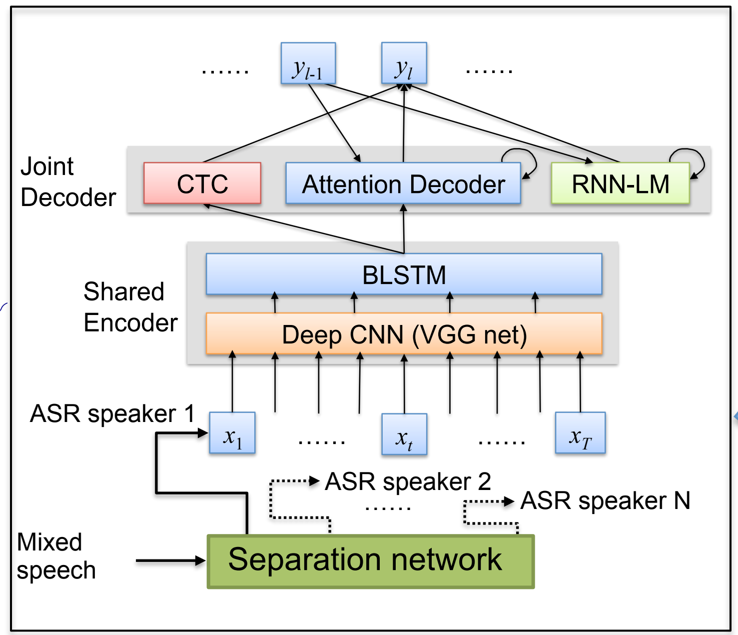
- Character Error Rate (%) on the generated evaluation set.



Multi-speaker end-to-end speech recognition

Joint separation and recognition with a single end-to-end deep network



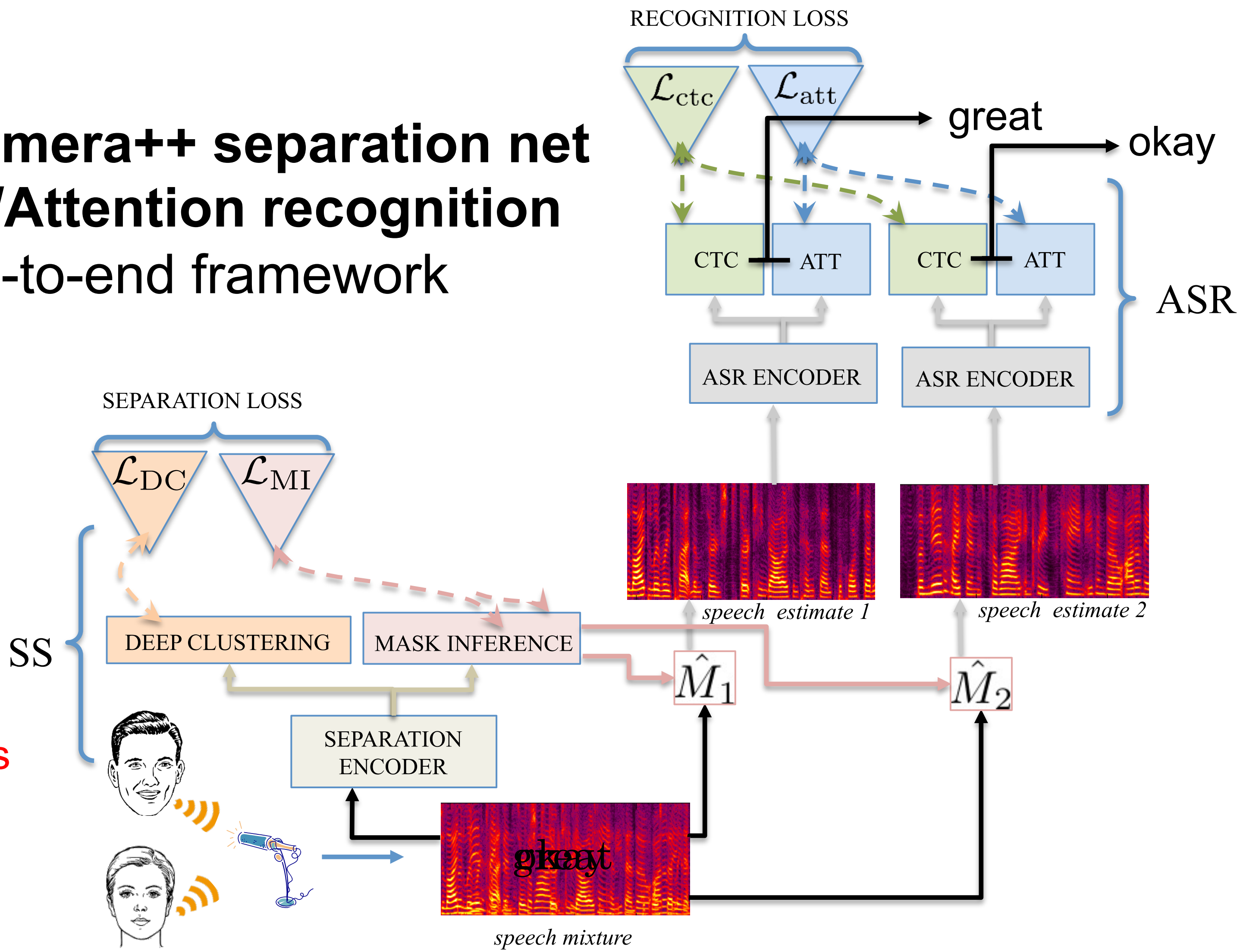


Single Deep Network

හෙලෝ
שלום
ဟလို
こんにちは
你好
Halló
Χαίρετε
안녕하세요
Ciao
Bonjour
Сайн уу
سلام
مرحبا
হ্যালো

End-to-end speech separation & recognition [Shane+18]

- Combine **Chimera++ separation net** and the **CTC/Attention recognition net** in an end-to-end framework



Use separation loss for pre-training SS and resolving permutation

Joint separation & recognition experiments

Oracle and baseline CER results (%) (w/ char LM)

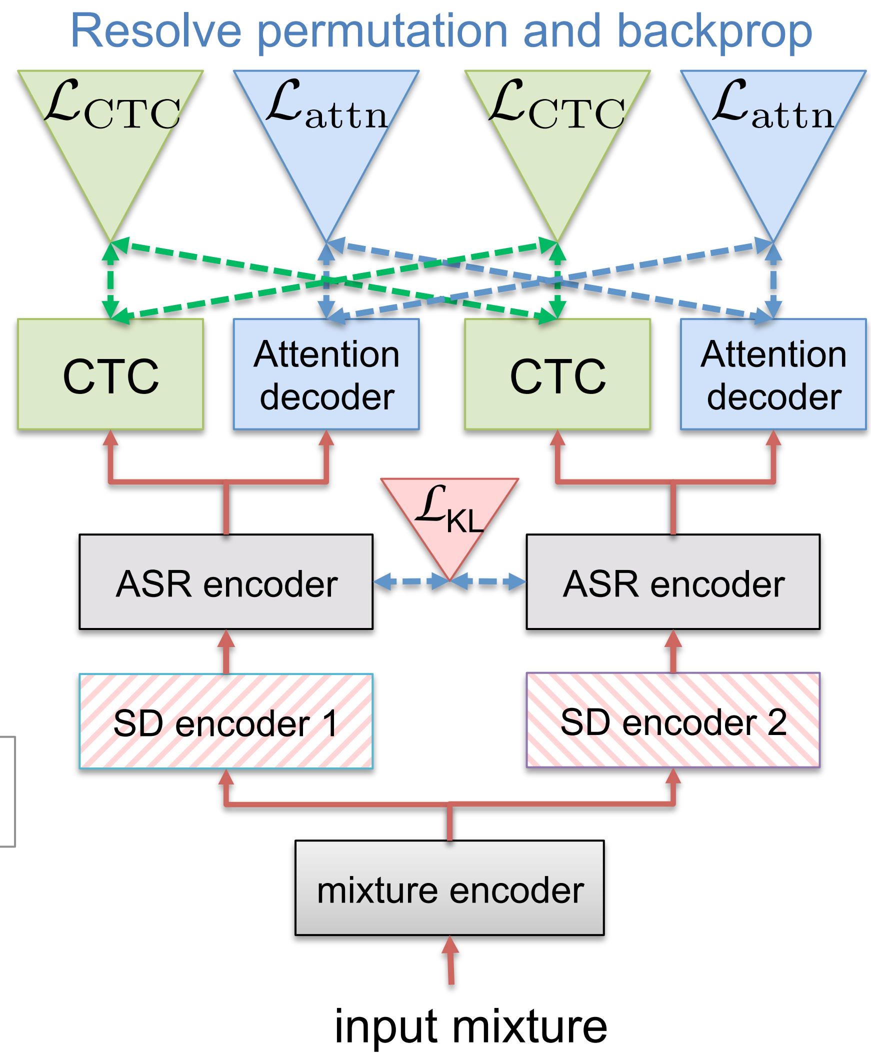
Training	Test	eval CER (%)
CLN	CLN	6.6
IBM	IBM	9.0
CLN	MIX	79.1

Proposed method, CER results (%) (w/ char LM)

Fine-tuning			CLN-ASR-PT		IBM-ASR-PT	
SS	ASR	Loss	dev CER (%)	eval CER (%)	dev CER (%)	eval CER (%)
NO	NO	-	34.1	32.0	24.2	23.1
NO	YES	ASR	18.9	18.0	18.7	17.9
YES	YES	SS+ASR	16.3	15.4	14.0	13.9
YES	YES	ASR	**13.3**	**13.2**	13.6	13.4

Purely end-to-end approach [Seki+, accepted to ACL'18]

- Not use any explicit separation network
- Incorporate implicit separation via speaker-differentiating (SD) encoders followed by a shared recognition encoder



- Transcript-level permutation-free loss

$$\mathcal{L} = \min_{\pi \in \mathcal{P}} \sum_{s=1}^{S} \text{Loss}(Y^s, R^{\pi(s)})$$

S: number of speakers	Y: network output
$\mathcal{P}$: possible permutations	R: reference

- No need for target speech in training
- Negative KL loss helps separate speaker-differentiating encodings

Purely end-to-end approach [Seki+, accepted to ACL'18]

CER (%) of mixed speech for WSJ (w/ word LM)

SPLIT	HIGH E. SPK.	LOW E. SPK.	AVG.
NO (BASELINE)	86.4	79.5	83.0
VGG	17.4	15.6	16.5
BLSTM	14.6	**13.3**	14.0
+ KL LOSS	**14.0**	**13.3**	**13.7**

Comparison with other methods

METHOD	WER (%)
DPCL + ASR (ISIK ET AL., 2016)	30.8
Proposed end-to-end ASR	**28.2**
METHOD	CER (%)
END-TO-END DPCL + ASR (CHAR LM) (SETTLE ET AL., 2018)	**13.2**
Proposed end-to-end ASR (char LM)	14.0

A bit worse than SS+ASR net but
no need for target speech in training

Multi-lingual ASR

(Supporting 10 languages: CN, EN, JP, DE, ES, FR, IT, NL, RU, PT)

ID	a04m0051_0.352274410405
	REF: [DE] bisher sind diese personen rundherum versorgt worden [EN] u. s. exports rose in the month but not nearly as much as imports ASR: [DE] bisher sind diese personen rundherum versorgt worden [EN] u. s. exports rose in the month but not nearly as much as imports

ID	csj-eval:s00m0070-0242356-0244956:voxforge-et-fr:mirage59-20120206-njp-fr-sb-570
	REF: [JP] 日本でもニュースになったと思いますが [FR] le conseil supérieur de la magistrature est présidé par le président de la république ASR: [JP] 日本でもニュースになったと思いますが [FR] le conseil supérieur de la magistrature est présidée par le président de la république

ID	voxforge-et-pt:insinfo-20120622-orb-209:voxforge-et-de:guenter-20140127-usn-de5-069:csj-eval:a01m0110-0243648-0247512
	REF: [PT] segunda feira [DE] das gilt natürlich auch für bestehende verträge [JP] え一同一人物による異なるメッセージを示しております ASR: [PT] segunda feira [DE] das gilt natürlich auch für bestehende verträge [JP] え一同一人物による異なるメッセージを示しております

Multi-speaker ASR w/ Purely E2E model

ID	445c040j_446c040f
Out[1]	REF: bids totaling six hundred fifty one million dollars were submitted ASR: bids totaling six hundred fifty one million dollars were submitted
Out[2]	REF: that's more or less what the blue chip economists expect ASR: that's more or less what the blue chip economists expect

ID	446c040j_441c0412
Out[1]	REF: this is especially true in the work of british novelists and even previously in the work of william boyd ASR: this is especially true in the work of british novelists and even previously in the work of william boyd
Out[2]	REF: as signs of a stronger economy emerge he adds long term rates are likely to drift higher ASR: a signs of a stronger economy emerge he adds long term rates are likely to drive higher

ID	440c040v_446c040n
Out[1]	REF: shamrock has interests in television and radio stations energy services real estate and venture capital ASR: chemlawn has interests in television and radio stations energy services real estate and venture capital
Out[2]	REF: as with the rest of the regime however their ideology became contaminated by the germ of corruption ASR: as with the rest of the regime however their ideology became contaminated by the jaim of corruption

Multi-lingual Multi-speaker ASR

ID	ralfherzog_1.41860235081
Out[1]	REF: [DE] eine höhere geschwindigkeit ist möglich ASR: [DE] eine höh*re geschwindigkeit ist möglich
Out[2]	REF: [JP] まずなぜこの内容を選んだかと言うと ASR: [JP] まずなぜこの内容を選んだかと言うと

ID	a02m0012_s00f0066
Out[1]	REF: [EN] grains and soybeans most corn and wheat futures prices were stronger [CN] 也是的 ASR: [EN] grains and soybeans most corn and wheat futures prices were strongk [CN] 也是的
Out[2]	REF: [JP] えーここで注目すべき点は例十十一の二重下線部に示すように [JP] アニメですとか ASR: [JP] えーここで注目すべきい点は零十十一の二十下線部に示すように [JP] アニメですとか

ID	a04m0051_0.352274410405
Out[1]	REF: [IT] economizzando le provvlste vi era da vivere per lo meno quattro glorni [EN] the warming trend may have melted the snow cover on some crops ASR: [IT] e cono mizzando le provveste vi*era da vivere per lo medo quattro gorni [EN] the warning trend may have mealtit the sno* cover on some crops
Out[2]	REF: [JP] でそれぞれの発話数え情報伝達の発話数一分当たりの発話数はえ多くなってますがえ問題解決だと少し少なくなるでディベートだとおー ASR: [JP] でそれですのでの発話スえ情報伝達の発話数一分当たり発話数はえ多くなってますがえ問題解決だと少してなくてでディベートだとおー

Conclusions

- Hybrid CTC/attention−based end-to-end speech recognition
  - Multi-task CTC/attention learning
  - Joint CTC/attention decoding
  - Extended network with a deep CNN and an RNN-LM
- Achieved good performance
  - Better than state-of-the-art ASR systems in Chinese and Japanese tasks
  - Multi-level LMs provided 5.6 %WER in WSJ task, which is the best end-to-end ASR performance
- Open source: ESPnet
  - https://github.com/espnet/espnet

Conclusions

- Multi-lingual end-to-end speech recognition
  - Trained a monolithic network with 10 languages with language IDs
  - No performance degradation compared to language dependent models
  - Effective especially for languages with small amount of training data

- Multi-speaker end-to-end speech recognition
  - Joint separation & recognition network
  - Purely end-to-end multi-speaker ASR

Thank you!