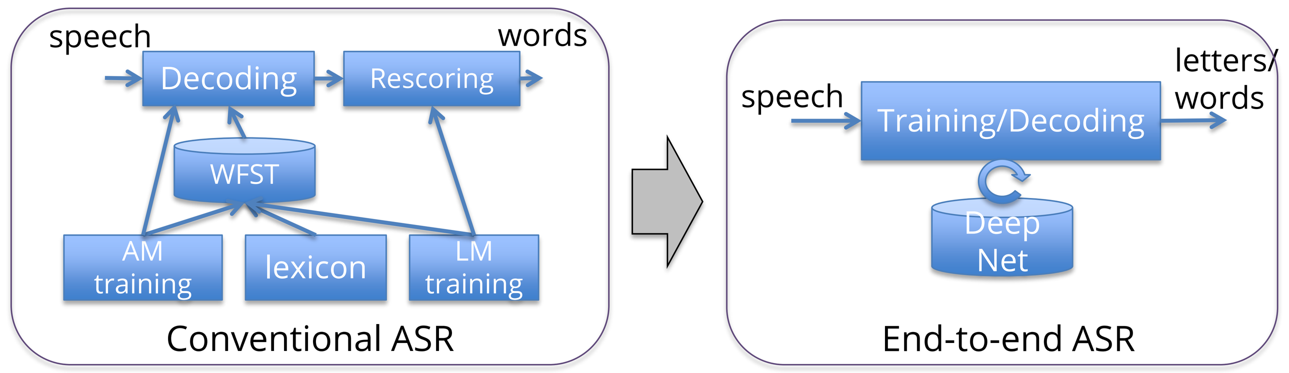
A PURELY END-TO-END SYSTEM FOR MULTI-SPEAKER SPEECH RECOGNITION

Hiroshi Seki (Toyohashi Univ.), Takaaki Hori, Shinji Watanabe, Jonathan Le Roux, John R. Hershey (MERL)

End-to-end automatic speech recognition (ASR)

- Prior to the deep learning revolution, speech processing tasks required a variety of different modules and were difficult to integrate
- Within speech recognition, end-to-end architectures have unified conventional modules into a single neural network system with no need for expert knowledge
- Easier to build accurate ASR systems for new tasks

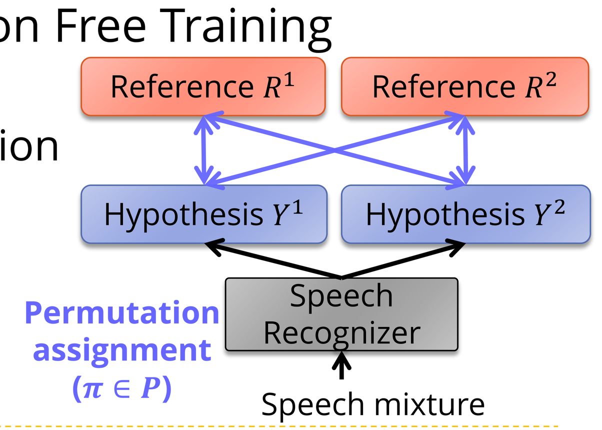


1

Multi-speaker speech recognition

- Generation of multiple transcriptions from a single-channel mixture of multiple speakers' speech.
- Permutation Problem
  - Correspondence between outputs of an algorithm and references is an arbitrary permutation.
- Transcription-level Permutation Free Training
  - One-to-many mapping by selecting the proper permutation of hypotheses and references.
  - Loss for S speaker mixtures:

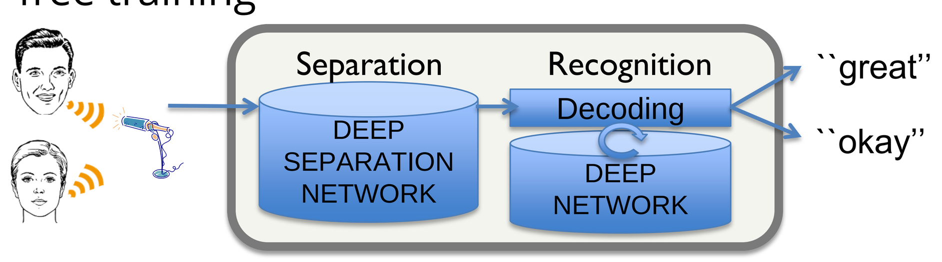
$$L = \min_{\pi \in P} \sum_{s=1}^{S} \mathrm{Loss}(Y^s, R^{\pi(s)})$$



2

Problem of conventional approach

- Preparation of explicit intermediate representation for efficient training.
  1. Explicit separation and recognition approach [Isik 2016, Settle 2018]
     × Pairwise unmixed speech for signal-level permutation free training



  2. (Non end-to-end) Implicit separation approach [Qian 2017]
     × Phonetic alignment information for transcription-level permutation free training

This work: End-to-end architecture without requiring explicit separation module and intermediate representation

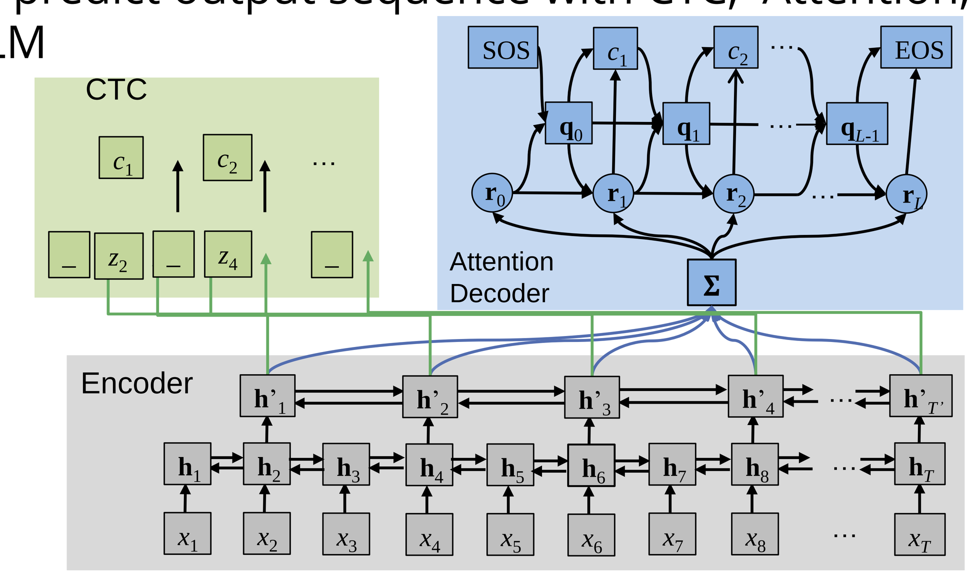
3

[Settle 2018] Joint optimization of separation and recognition modules based on ASR loss under end-to-end framework

Joint CTC/attention architecture

[Hori, et al. 2017]

- Jointly predict output sequence with CTC, Attention, and RNN-LM



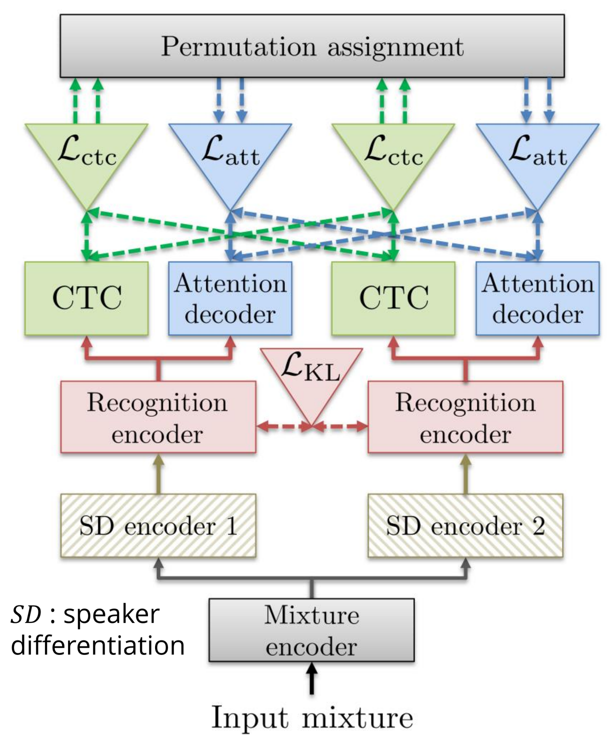
$$\mathcal{L}_{ctc} = \mathrm{Loss}_{ctc}(Y_{ctc}, R) = -\log p_{ctc}(Y_{ctc} = R|X)$$

$$p_{att}(Y_{att}|X) = \prod_n p_{att}(y_n|H, y_{1:n-1})$$

$$\mathcal{L}_{att} = \mathrm{Loss}_{att}(Y_{att}, R) = -\log p_{att}(Y_{att} = R|X)$$

4

Proposed end-to-end permutation free training



Encoder submodules

$$H_{mix} = \mathrm{Encoder}_{mix}(X)$$
$$H_{SD}^s = \mathrm{Encoder}_{SD}^s(H_{mix}), s = \{1, \cdots, S\}$$
$$H_{Rec}^s = \mathrm{Encoder}_{Rec}(H_{SD}^s), s = \{1, \cdots, S\}$$

CTC loss

$$\mathcal{L}_{ctc} = \min_{\pi \in P} \sum_{s=1}^{S} \mathrm{Loss}_{ctc}(Y_{ctc}^s, R^{\pi(s)})$$

- s: encoder index
- $\pi(s)$: reference index

Attention decoder loss

$$\mathcal{L}_{att} = \min_{\pi \in P} \sum_{s=1}^{S} \mathrm{Loss}_{att}(Y_{att}^{s,\pi(s)}, R^{\pi(s)})$$

5

Reduction of permutation cost

- Synchronous output
  - Decision of best permutation based on the CTC loss alone.

$$\hat{\pi} = \mathrm{argmin}_{\pi \in P} \sum_{s}^{S} \mathrm{Loss}_{ctc}(Y_{ctc}^s, R^{\pi(s)})$$

$$\mathcal{L}_{ctc} = \min_{\pi \in P} \sum_{s=1}^{S} \mathrm{Loss}_{ctc}(Y_{ctc}^s, R^{\hat{\pi}(s)})$$

$$\mathcal{L}_{att} = \cancel{\min_{\pi \in P}} \sum_{s=1}^{S} \mathrm{Loss}_{att}(Y_{att}^{s,\hat{\pi}(s)}, R^{\hat{\pi}(s)})$$

*Permutation based on CTC was 16.3 times faster than that based on the decoder network

6

Promoting separation of hidden vectors

- Generation of multiple label sequences based on single decoder network
  - Frame-wise negative KL loss

$$L_{KL} = -\eta \sum_l \{\mathrm{KL}\left(\overline{H_{Rec}^1}(l) || \overline{H_{Rec}^2}(l)\right) + \mathrm{KL}\left(\overline{H_{Rec}^2}(l) || \overline{H_{Rec}^1}(l)\right)\}$$

$$\overline{H_{Rec}^s} = \left(\mathrm{softmax}(H_{Rec}^s(l))\right| l = 1, \cdots L)$$

Encouragement of hidden vectors to avoid generating similar hypotheses.

7

Experiments (1/2)

- Corpus1: Wall Street Journal (WSJ)
- Corpus2: Corpus of Spontaneous Japanese (CSJ)

Duration (hours) of unmixed and mixed corpora

Mixed: mixture of 2 speakers between 0~5 dB

	Training	Development	Evaluation
WSJ (unmixed)	81.5	1.1	0.7
WSJ (mixed)	98.5	1.3	0.8
CSJ (unmixed)	583.8	6.6	5.2
CSJ (mixed)	826.9	9.1	7.5

- Input / Output
  - Input: 80 dim. mel-filterbank + pitch feature (+delta, delta delta)
  - Output (WSJ): 49 labels (alphabets and special tokens)
  - Output (CSJ): 3,315 labels (Japanese Kanji/Hiragana/Katakana characters and special tokens)

8

Experiments (2/2)

- Baseline model for single-speaker ASR
  - Encoder: 6-layer CNN + 7-layer BLSTM (320 cells)
  - Decoder: 1-layer LSTM (320 cells) with location-based attention mechanism
- Proposed models for multi-speaker ASR
  - 2 encoder architectures and (# layers):

Split by	$\mathrm{Encoder}_{Mix}$	$\mathrm{Encoder}_{SD}$	$\mathrm{Encoder}_{Rec}$
No (baseline)	VGG (6)	—	BLSTM (7)
VGG	VGG (4)	VGG (2)	BLSTM (7)
BLSTM	VGG (6)	BLSTM (2)	BLSTM (5)

- Joint decoding with RNN-LM

9

Results

- Evaluation of unmixed speech

Task	Avg. Char. error rate [%]
WSJ	2.6
CSJ	7.8

- Character Error Rate (CER) [%] of mixed speech for WSJ task

Split by	High E. Spk	Low E. Spk	Avg.
No	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

- Character Error Rate (CER) [%] of mixed speech for CSJ task

Split by	High E. Spk	Low E. Spk	Avg.
No	93.3	92.1	92.7
BLSTM	11.0	18.8	14.9

10

Comparison with other approaches

- Explicit separation and recognition approach

Method	Word Error Rate (%)
Deep clustering + ASR [Isik 2016]	30.8
This work	28.2

- End-to-end explicit separation and recognition approach

Method	Character Error Rate (%)
End-to-end Deep clustering + ASR [Settle 2018]	13.2
This work	14.0

Comparable performance to the end-to-end explicit separation and recognition network, without having to pre-train using clean signal training references.

11

Conclusions

- Proposed an approach to directly convert an input speech mixture into multiple label sequences under the end-to-end framework
- Eliminated the necessity to prepare explicit intermediate representation, e.g. phonetic alignment information or pairwise unmixed speech.
- Achieved comparable performance with an end-to-end system featuring explicit separation and recognition modules.

12