MITSUBISHI ELECTRIC
Changes for the Better

for a greener tomorrow

Transformer-based Long-context End-to-end Speech Recognition

Takaaki Hori, Niko Moritz, Chiori Hori, Jonathan Le Roux

Interspeech 2020

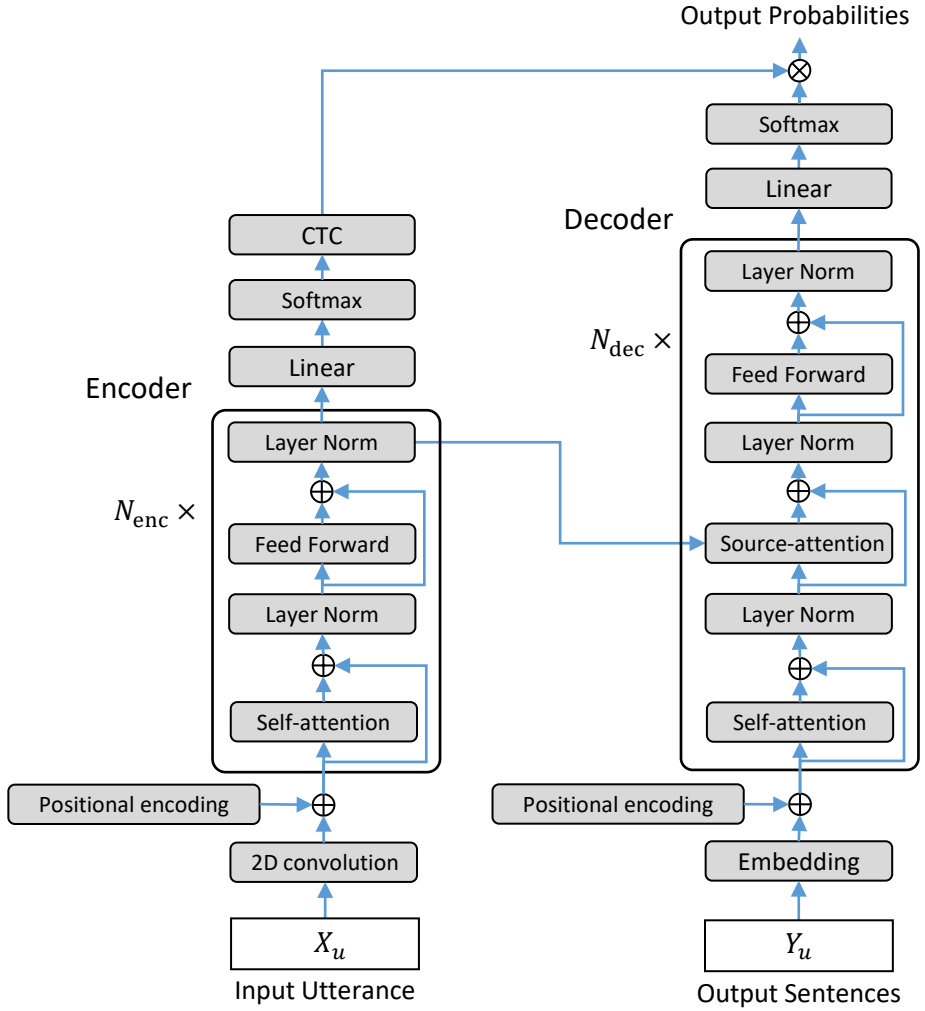
MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts, USA
http://www.merl.com

Introduction

- End-to-end ASR is becoming more widespread in the field
  - Enables to build ASR systems without expert knowledge
  - Achieves competitive or better performance than typical hybrid systems mainly by *Transformer*-based models
- Most end-to-end systems are basically designed to recognize independent utterances
- However, contextual information over multiple utterances, such as information on speaker or topic, is known to be useful for recognizing long audio recordings such as lecture and conversational speeches
- This work
  - Proposes a context-expanded Transformer that can utilize contextual information of audio and text features over multiple utterances
  - Reports substantial improvement of ASR accuracy for lecture and conversational speeches

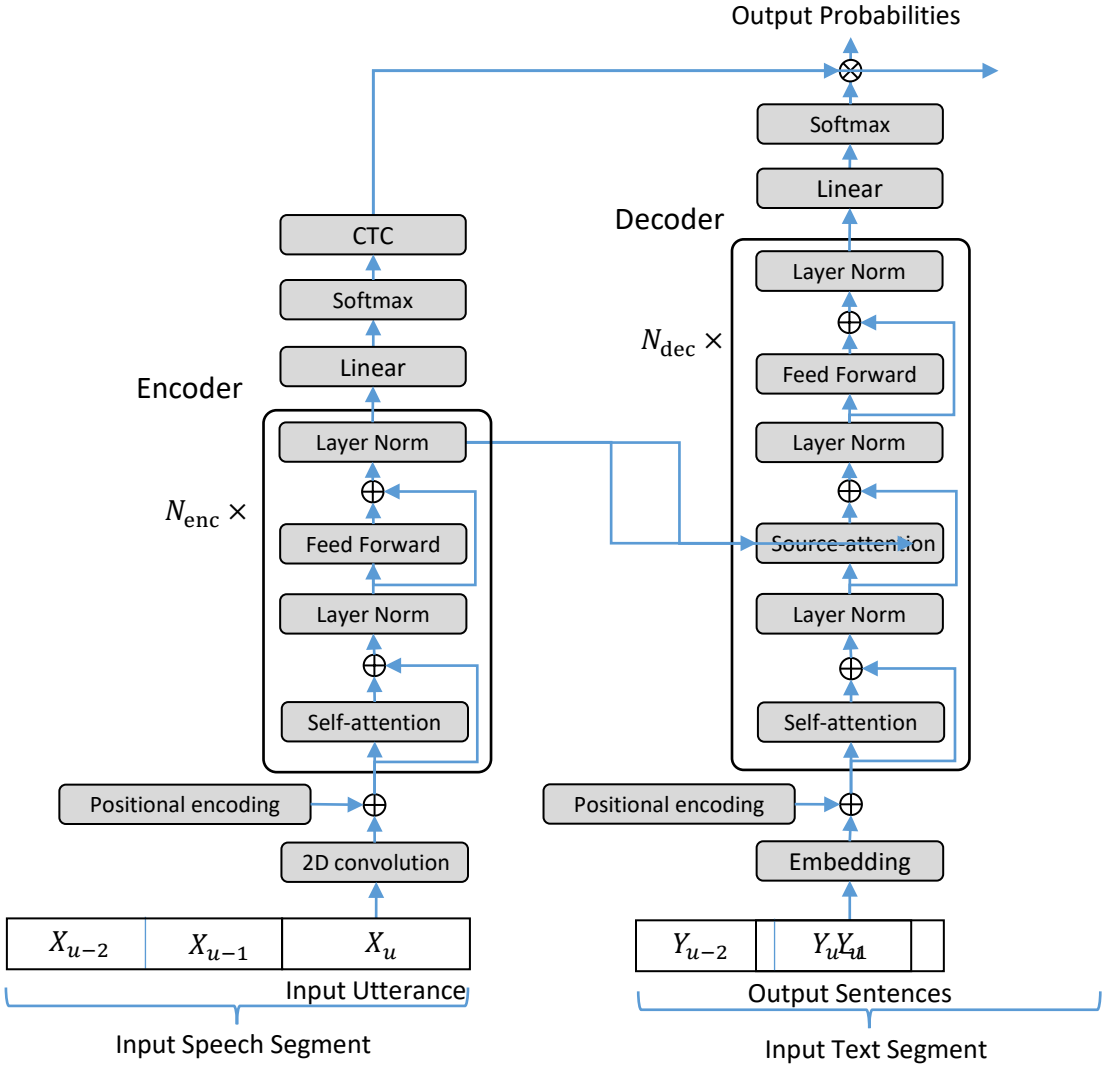
Transformer-based ASR

- Transformer [Vaswani+'17]
  - Encoder-decoder-based sequence-to-sequence model
  - Deep feed-forward architecture without recurrent connections
  - Exploits dependencies between all frames within a sentence via self/source-attention mechanisms
  - Applied to ASR [Dong+'18, Karita+'19(1)]
- Use CTC to promote monotonic alignment [Kim+'17, Hori+'17, Karita+'19(1)]
- Outperforms RNN-based models in major ASR tasks [Karita+'19(2)]
- Assumes per-utterance input/output



Context-expanded Transformer (proposed)

- Accepts speech and text segments including multiple consecutive utterances

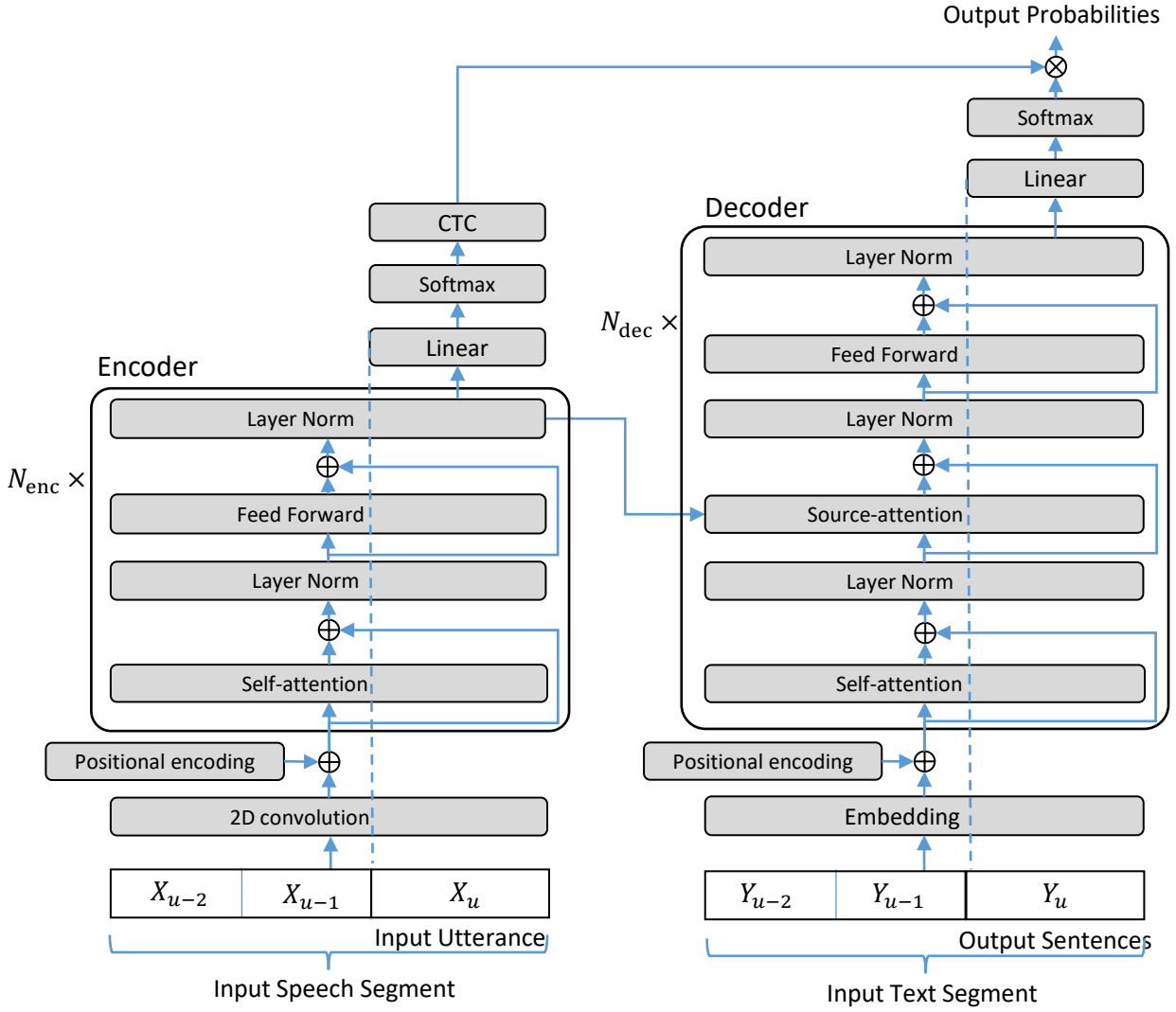


Context-expanded Transformer (proposed)

- Accepts speech and text segments including multiple consecutive utterances
- Predicts an output sequence for the last utterance

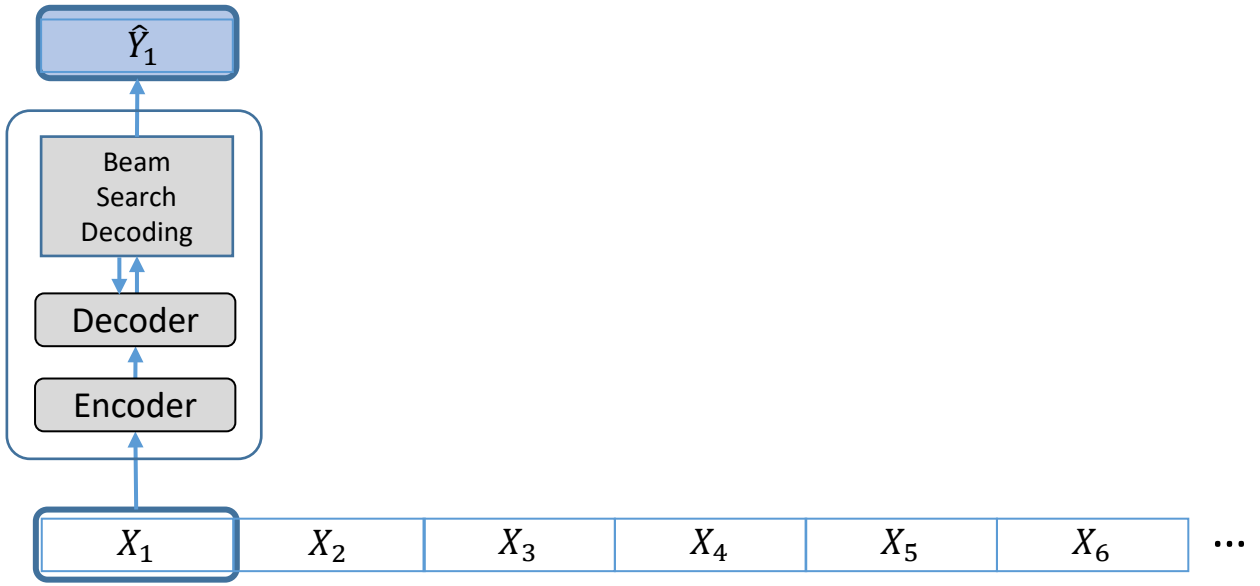
$$\hat{Y}_u = \underset{Y_u \in \mathcal{V}^*}{\operatorname{argmax}}\, p(Y_u | Y_{v:u-1}, X_{v:u})$$

- Gives position 0 to the beginning of the last utterance for positional encoding
- Applies self-attention encoder/decoder layers to the input segments
- Enables adaptation of the acoustic and language features using previous utterances at every encoder and decoder layer



Decoding long audio with the context-expanded Transformer

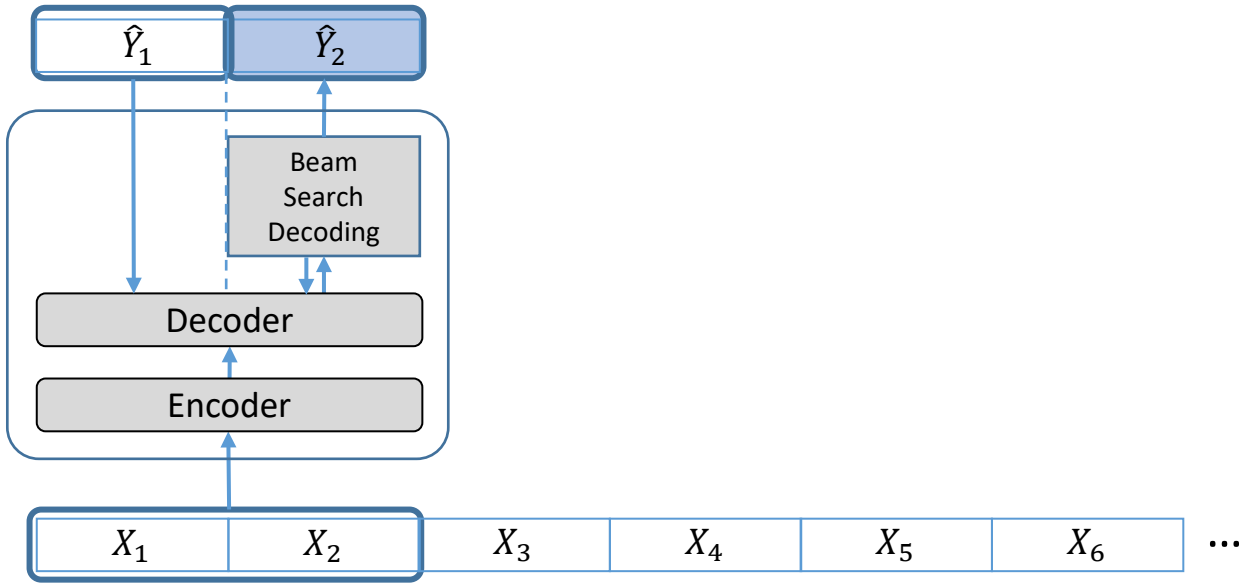
- Given the first utterance, it performs beam search decoding without any contextual information



Input utterances from long audio

Decoding long audio with the context-expanded Transformer

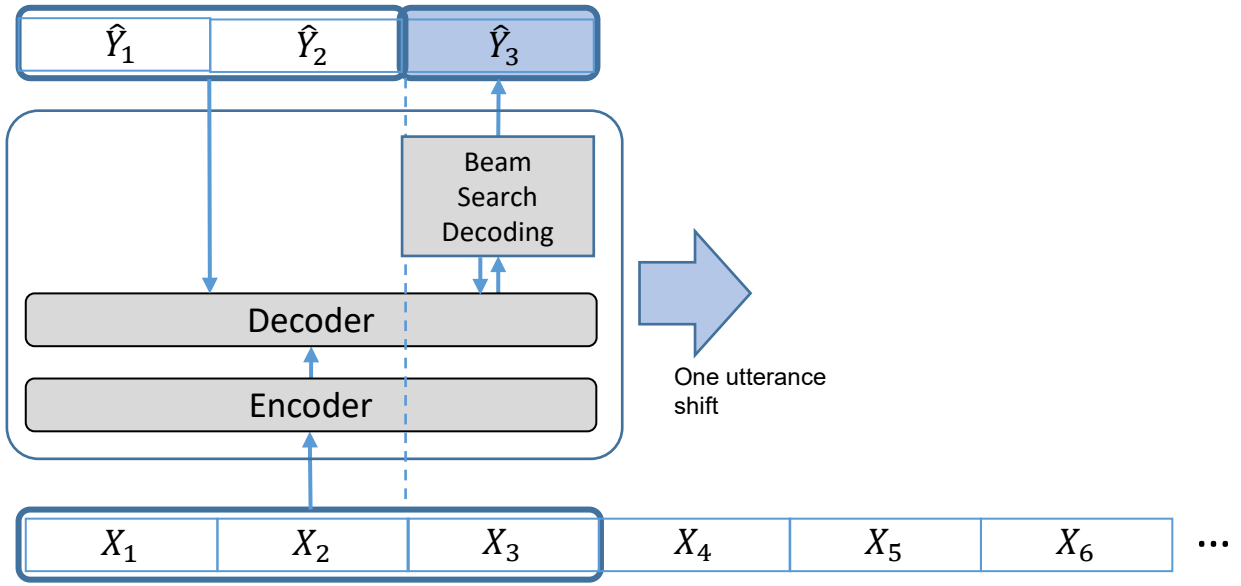
- For the second utterance, it performs decoding as well, but the input segments include the previous utterance and hypothesis



Input utterances from long audio

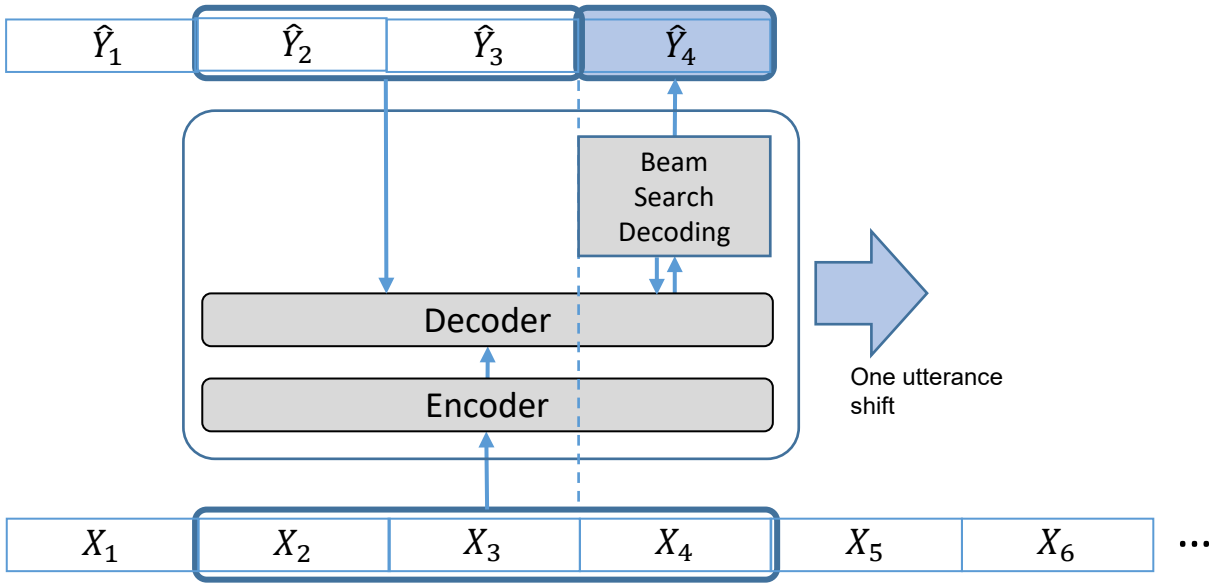
Decoding long audio with the context-expanded Transformer

- Repeats decoding in a sliding-window fashion with one-utterance shifts



Decoding long audio with the context-expanded Transformer

- Repeats decoding in a sliding-window fashion with one-utterance shifts

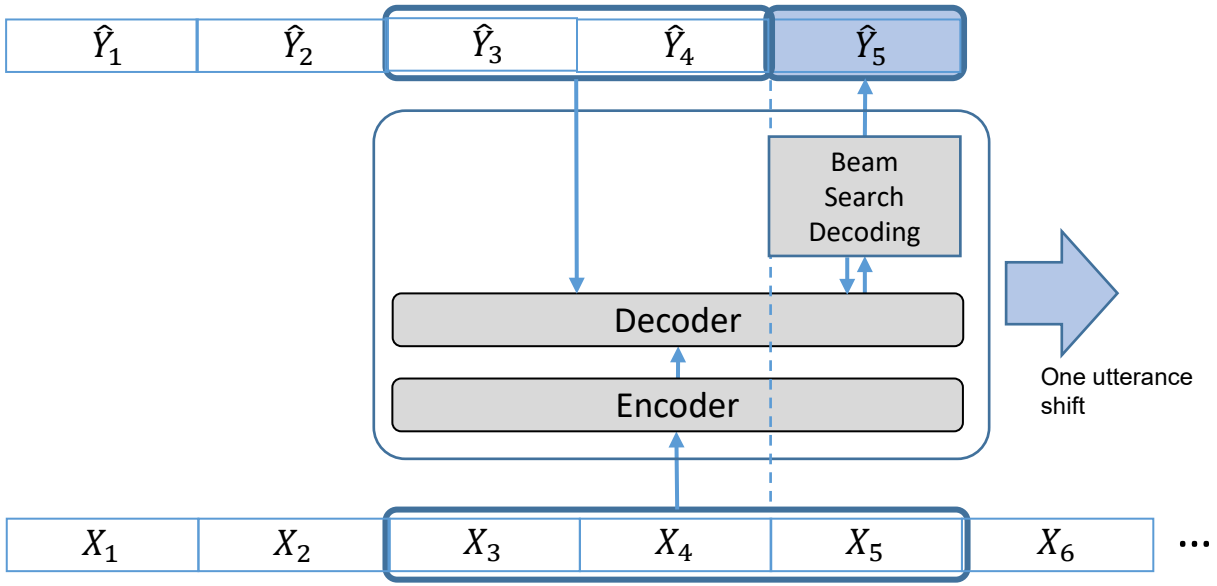


Input utterances from long audio

The input speech/text segments are truncated if the segment length is greater than a threshold (Max Segment Length)

Decoding long audio with the context-expanded Transformer

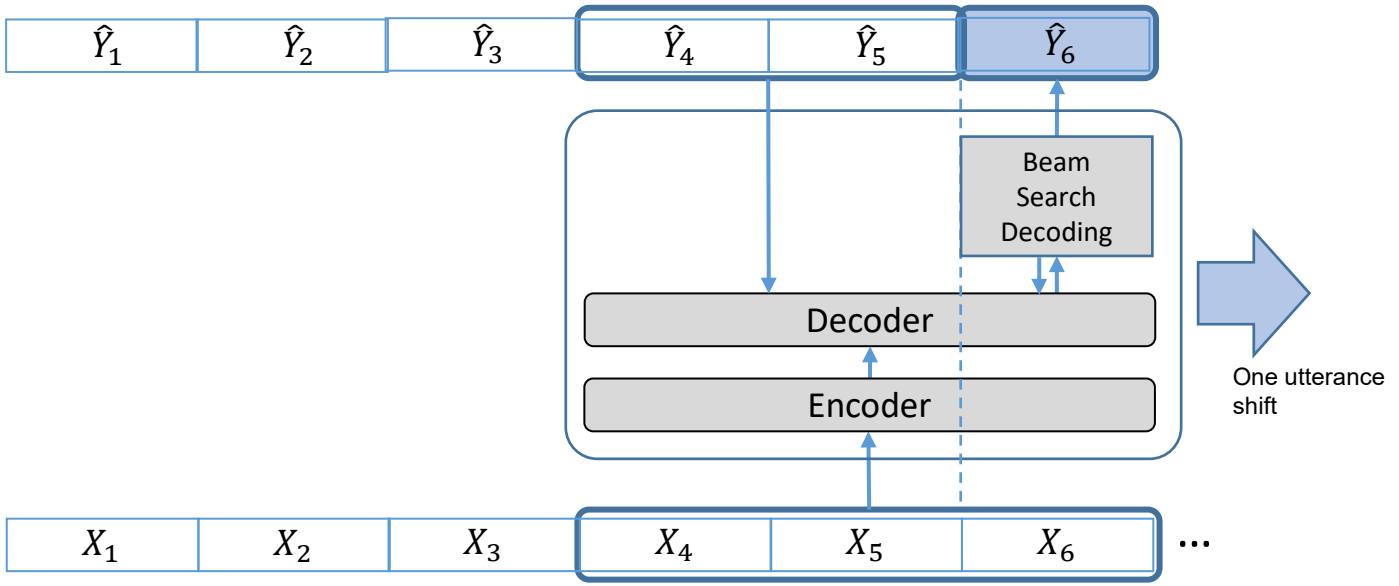
- Repeats decoding in a sliding-window fashion with one-utterance shifts



Input utterances from long audio

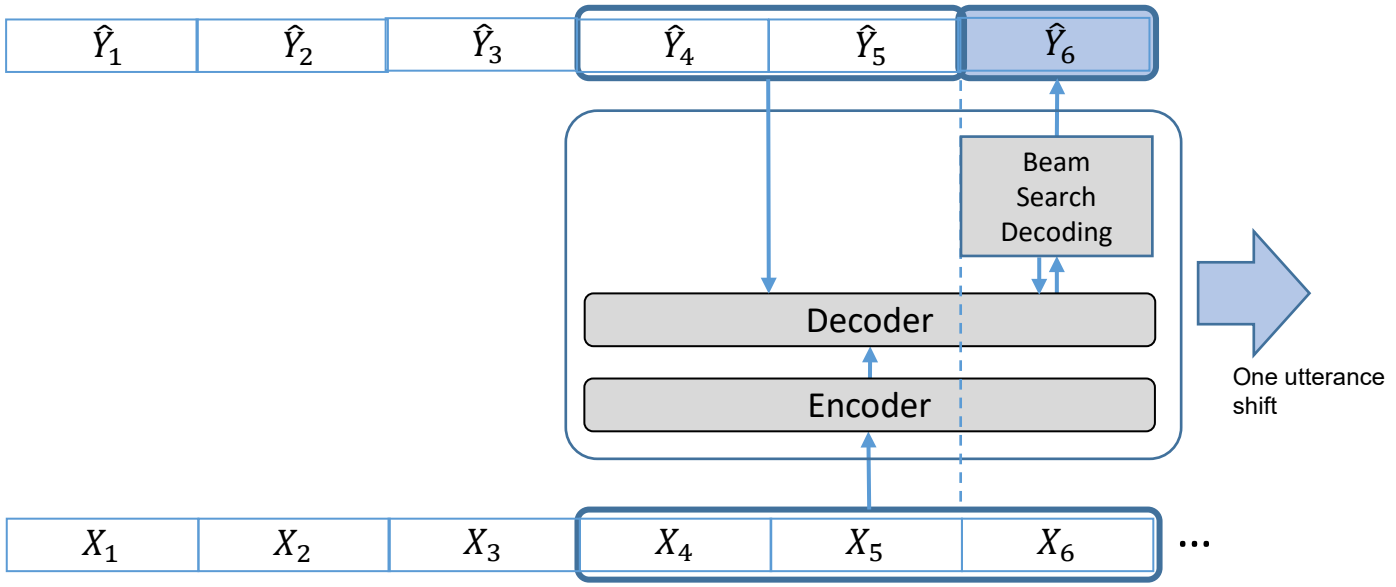
Decoding long audio with the context-expanded Transformer

- Repeats decoding in a sliding-window fashion with one-utterance shifts



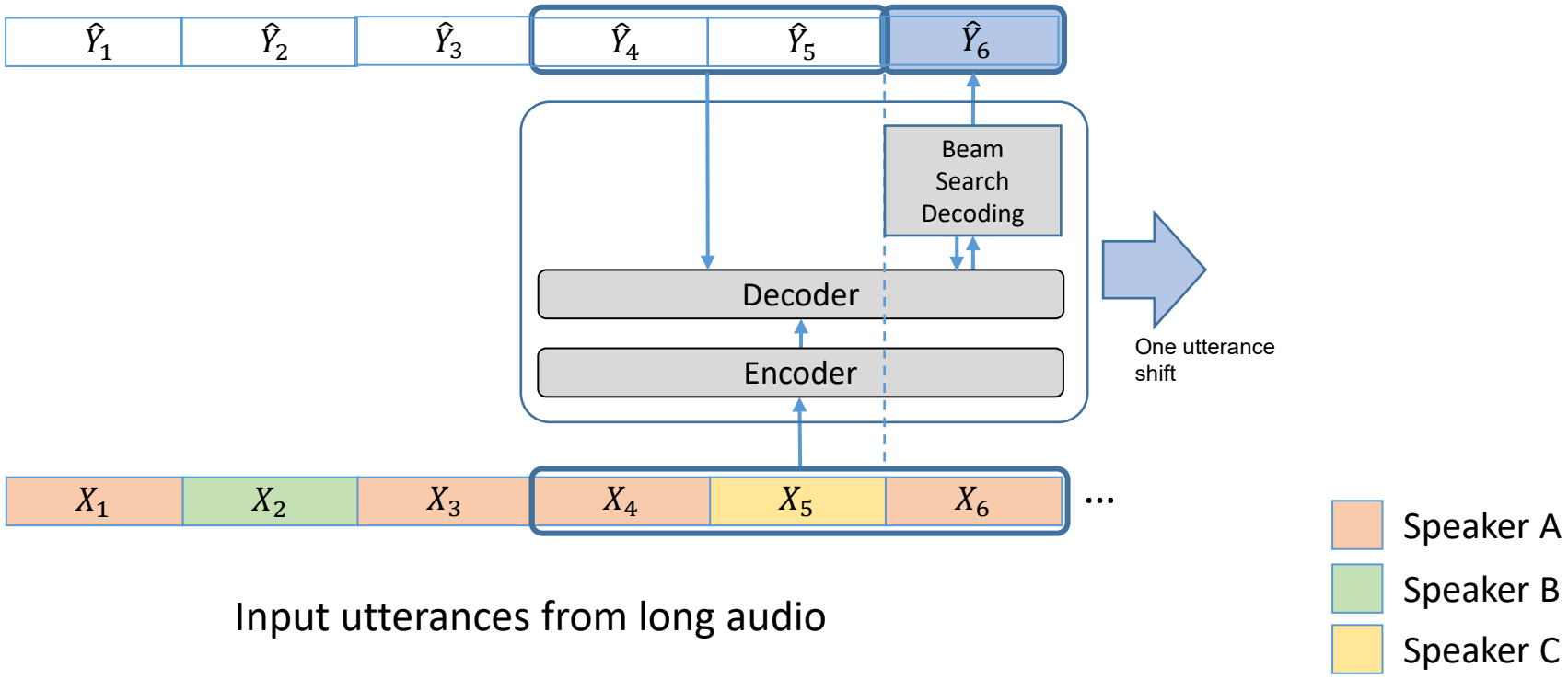
Speaker-dependent (SD) context for conversational speeches

- Arrange each speech segment to include only utterances from the same speaker



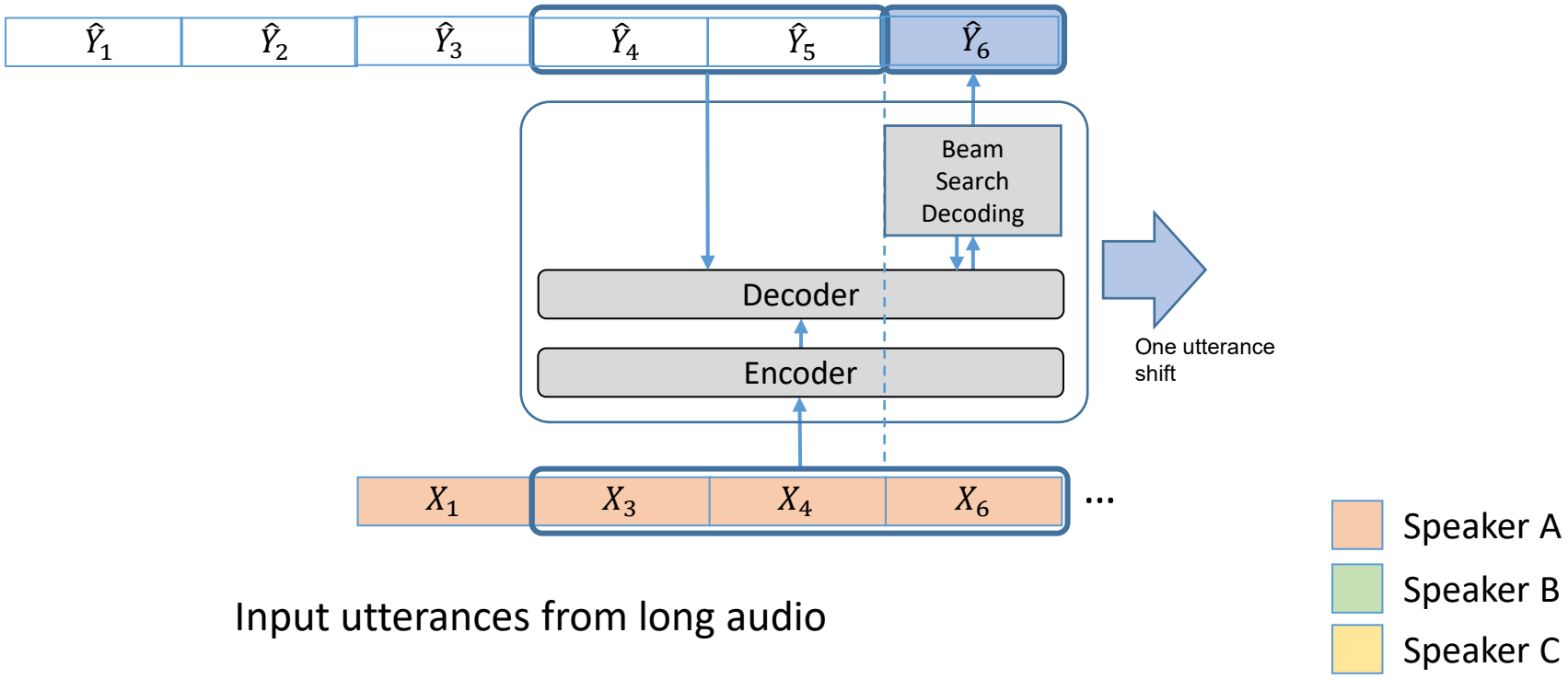
Speaker-dependent (SD) context for conversational speeches

- Arrange each speech segment to include only utterances from the same speaker



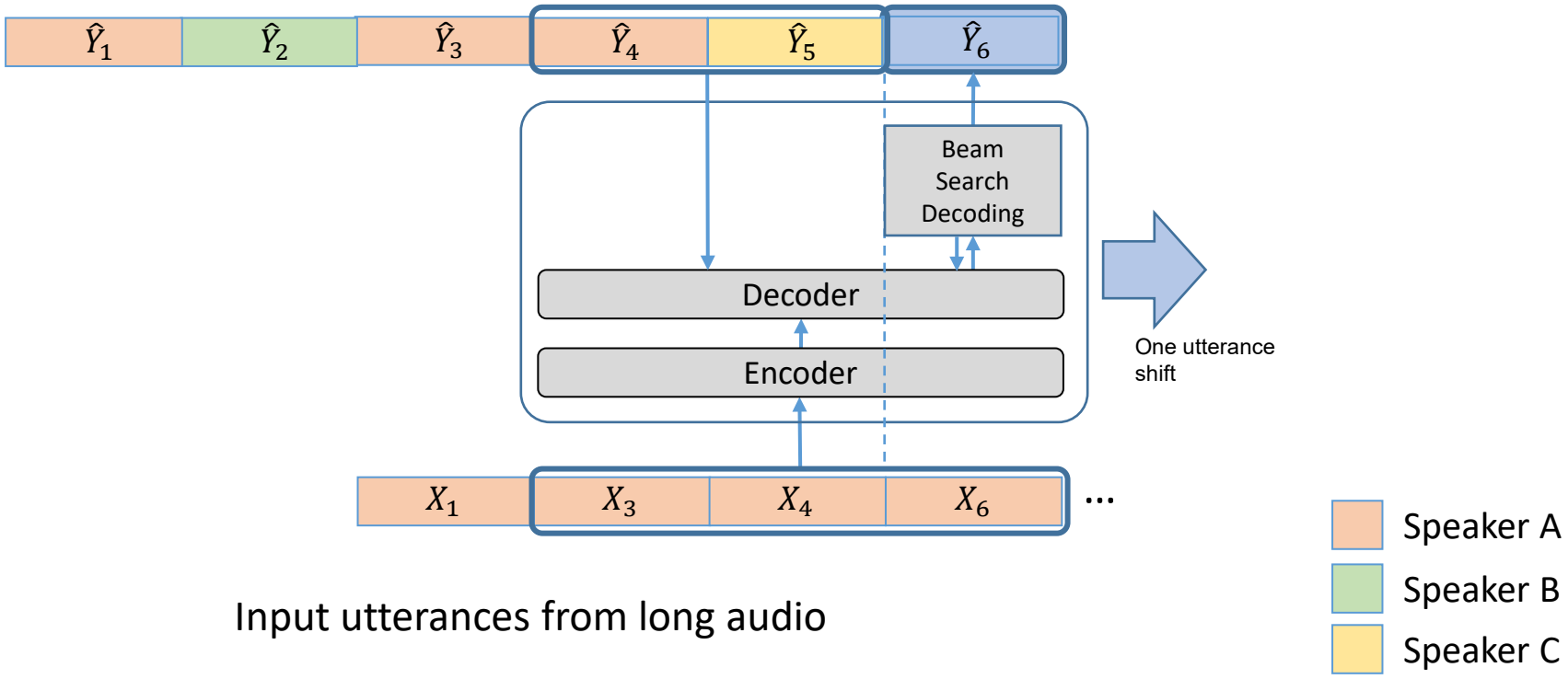
Speaker-dependent (SD) context for conversational speeches

- Arrange each speech segment to include only utterances from the same speaker



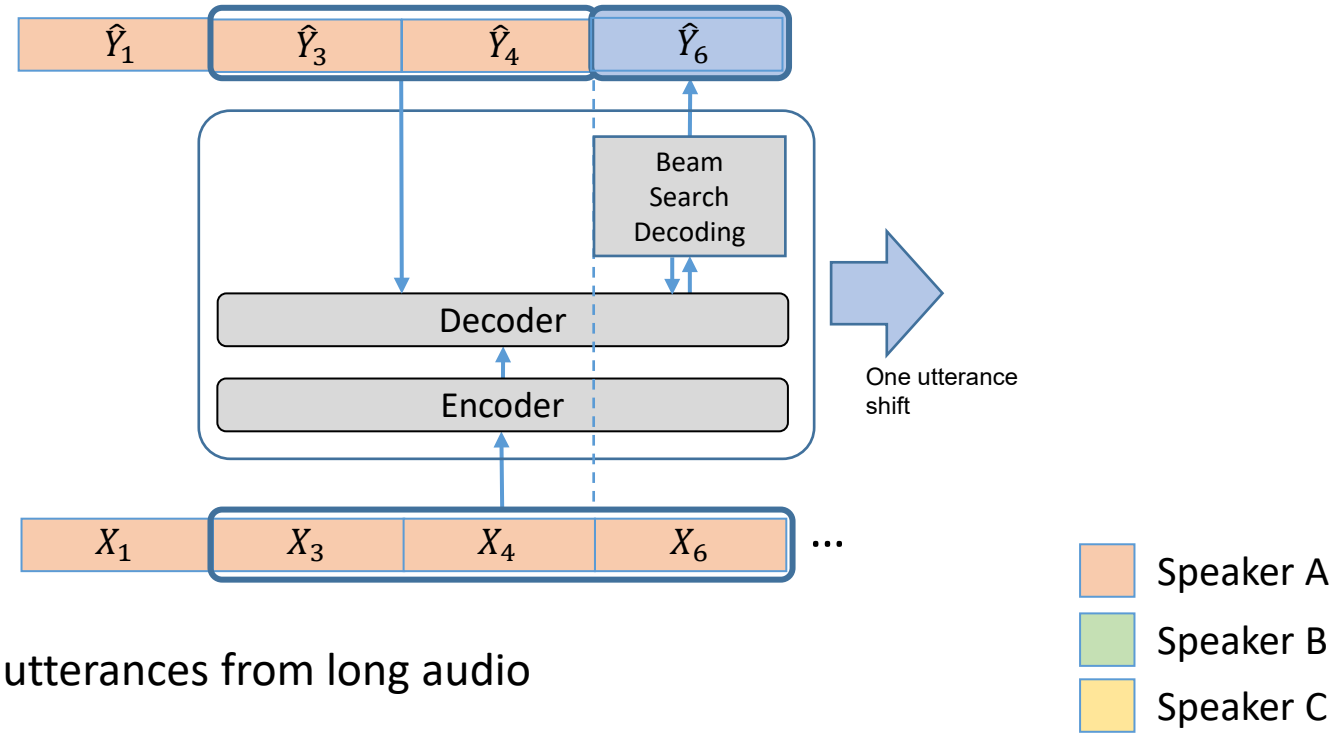
Speaker-dependent (SD) context for conversational speeches

- Also possible to arrange the text segment from the same speaker



Speaker-dependent (SD) context for conversational speeches

- Also possible to arrange the text segment from the same speaker



Relation to prior work

- Acoustic context
  - Online speaker adaptation using i-vectors [Dehak+'10, Audhkhasi+'17], where each speaker i-vector is appended to every input frame or hidden vector in the encoder
- Language context
  - Hierarchical Recurrent Encoder Decoder (HRED) is applied to RNN-based end-to-end ASR, where a sentence-level state vector is fed to the decoder network [Kim+'18, Masumura+'19]

→ Prior methods typically summarize the contextual information into a single vector, and the original information may not be preserved sufficiently

- Context-expanded Transformer (proposed)
  - Adapts each utterance using its previous utterances via self-attention mechanism without summarizing the utterances into a single vector
  - Expands the encoder and the decoder to exploit both acoustic and language contexts, with which the model parameters are optimized jointly
  - No additional parameters

Experiments

- Monologue and dialogue ASR benchmarks

dataset	language	type	hours	test sets
CSJ	ja	monologue	581	eval1 / eval2 / eval3
TED-LIUM3	en	monologue	452	dev / test
SWITCHBOARD	en	dialogue	260	(eval2000) callhm / swbd
HKUST	zh	dialogue	200	train_dev / dev

- Model architecture
  - 12-layer encoder / 6-layer decoder
  - Self-attention/source-attention had 256 dims. with 4-head attention
  - Feed-forward layer had 2,048 hidden units
  - Max Segment Length was 20 sec. by default
  - 2-layer LSTM language model with 1,000 cells (used in CSJ, TED-LIUM3, and HKUST)

Monologue ASR results

	Speed perturb.	CSJ CERs [%s]			TED-LIUM3 WERs [%]	
		eval1	eval2	eval3	dev	test
Baseline		6.0	4.2	4.7	11.9	8.7
+ i-vector		5.9	4.1	4.9	11.2	8.6
Proposed		**5.5**	**3.8**	**4.0**	**10.5**	**8.1**
ESPnet [Karita+'19 (2)]	✔	5.7	4.1	4.5	9.7	8.0
Proposed	✔	**5.3**	**3.6**	**3.8**	**9.2**	**7.5**

- Proposed method provides substantial error reductions for both CSJ and TED-LIUM3, while i-vector provides very limited reductions.
- Relative error rate reduction ranges from 5 to 15%
- Performance gain is preserved even with speed perturbation, achieving SOTA performance

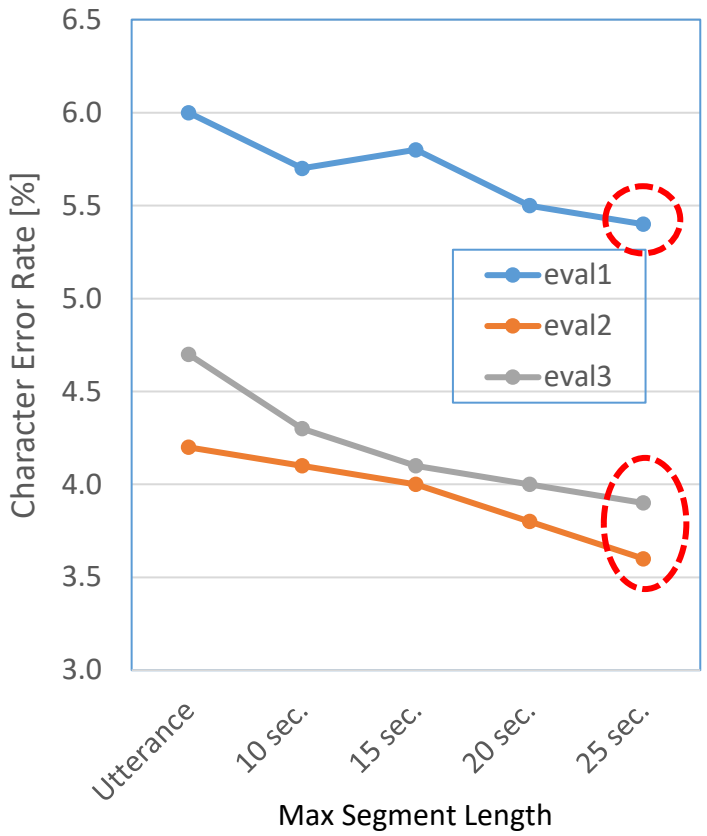
Dialogue ASR results

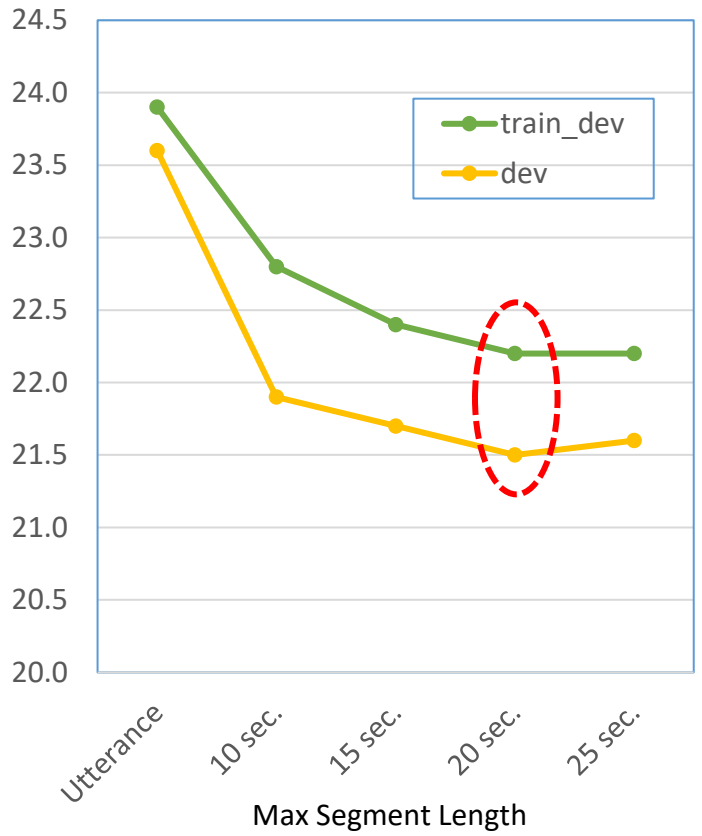
	context		SWITCHBOARD WERs [%]		HKUST CERs [%]	
	input	output	callhm	swbd	train_dev	dev
Baseline			17.7	8.9	23.9	23.6
+ i-vector	SD		17.8	8.8	-	-
ESPnet [Karita+'19 (2)]			18.1	9.0	-	23.5
Proposed	SI	SI	15.6	8.4	22.8	22.5
	SD	SI	15.4	**8.2**	22.5	22.1
	SD	SD	**15.3**	8.3	**22.2**	**21.5**

- The proposed method reduces recognition errors on SWITHBOARD and HKUST, where the SD context further reduces the errors especially for HKUST
- Relative error rate reduction ranges from 7 to 13.5 %

SI: Speaker-Independent
SD: Speaker-Dependent

Max Segment Length vs. CER in CSJ and HKUST





- The error rate gradually decreases as the segment length increases.
- The error rate saturates with 20 sec. in HKUST, but it decreases further with 25 sec. in CSJ.

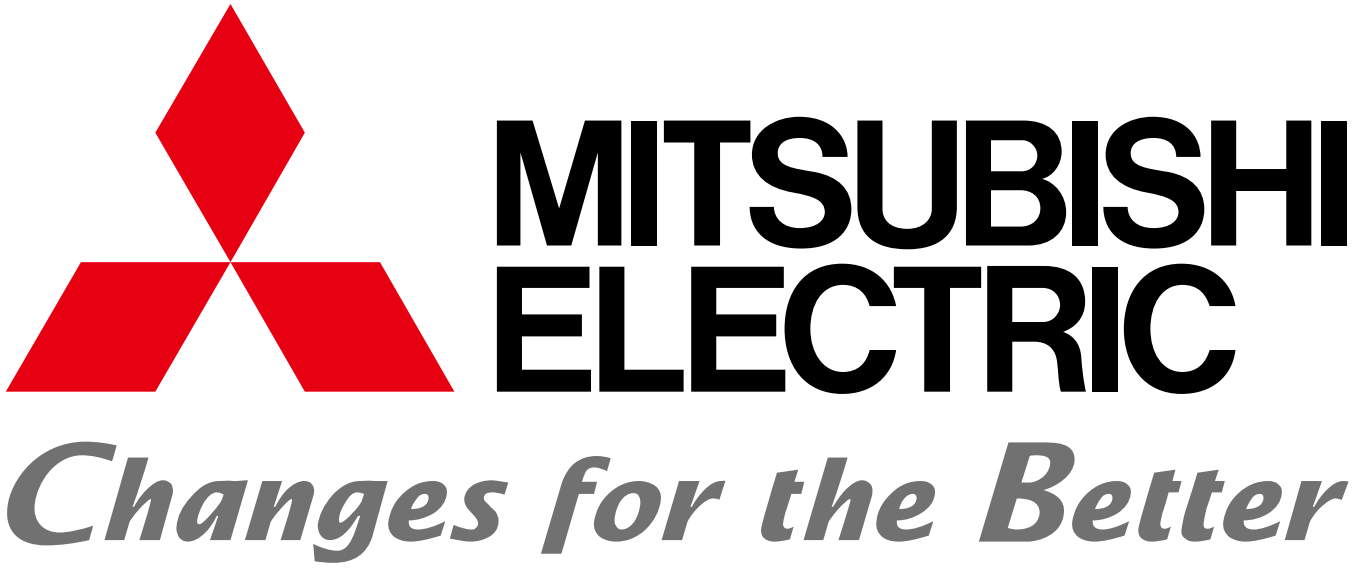
Contribution of input/output/LM contextual information

	Use context					CSJ eval1 CERs [%]	HKUST dev CERs [%]
	training		decoding				
	input	output	input	output	LM		
Baseline						6.0	23.6
Proposed	✓	✓	✓	✓	✓	**5.5** ↘	**21.5** ↘
	✓	✓	✓	✓		5.6 ↘	21.8 ↘
	✓	✓	✓			5.7 ↘	22.2 ↘
	✓	✓		✓	✓	8.0 ↗	25.1 ↗
	✓	✓				6.3 ↗	25.1 ↗
	✓		✓			5.7 ↘	22.5 ↘
		✓		✓	✓	5.8 ↘	23.0 ↘

- Both input and output contexts are important.
- Training/decoding mismatch on input context degrades performance.

Conclusions

- Proposed an approach to end-to-end ASR for long audio recordings such as lecture and conversational speeches.
- The proposed context-expanded Transformer
  - Accepts multiple consecutive utterances at the same time and predicts an output sequence for the last utterance, and
  - Repeats decoding in a sliding-window fashion with one-utterance shifts to recognize an entire recording.
- Demonstrated the effectiveness of the approach using monologue and dialogue benchmarks, achieving 5-15% relative error reduction from utterance-based ASR baselines.
- Future work will include reduction of computational complexity and memory usage to utilize longer contextual information.


MITSUBISHI
ELECTRIC
Changes for the Better