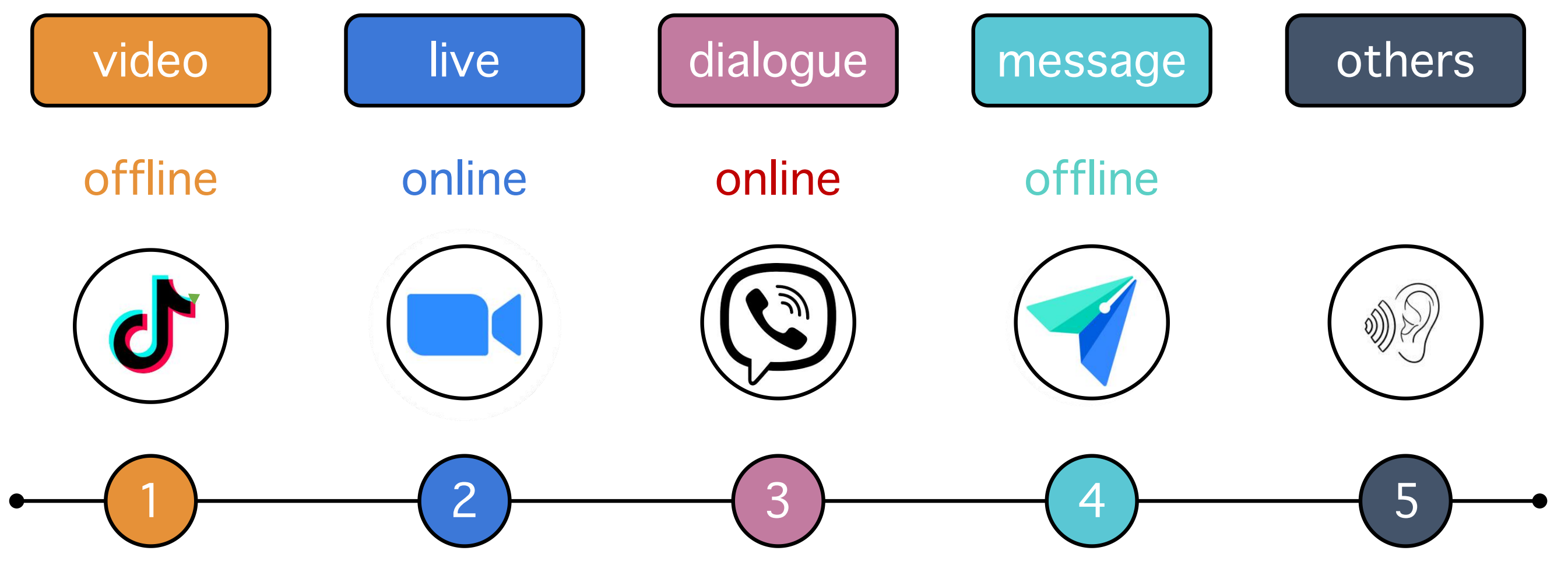
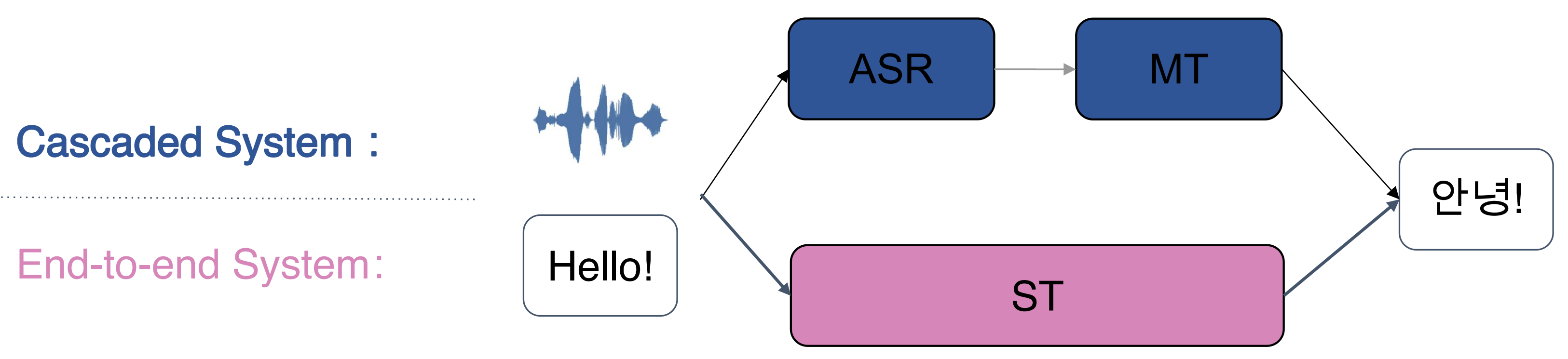
中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

火山翻译

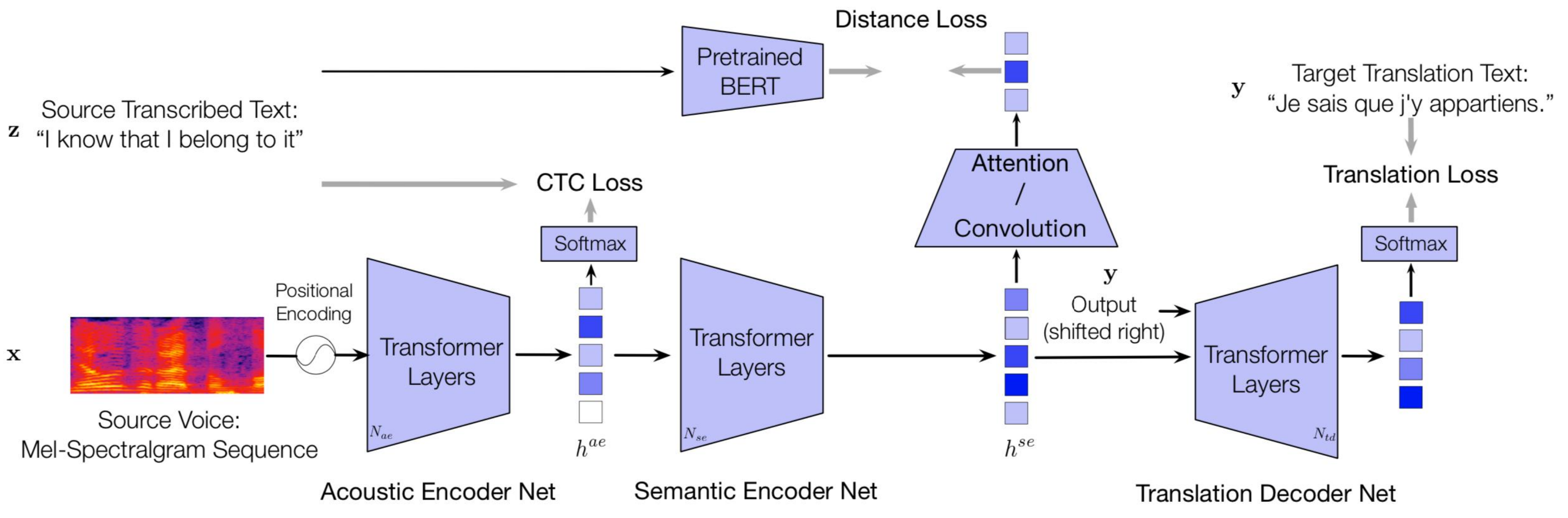
国际人工智能会议
AAAI 2021 论文北京预讲会

1 Background

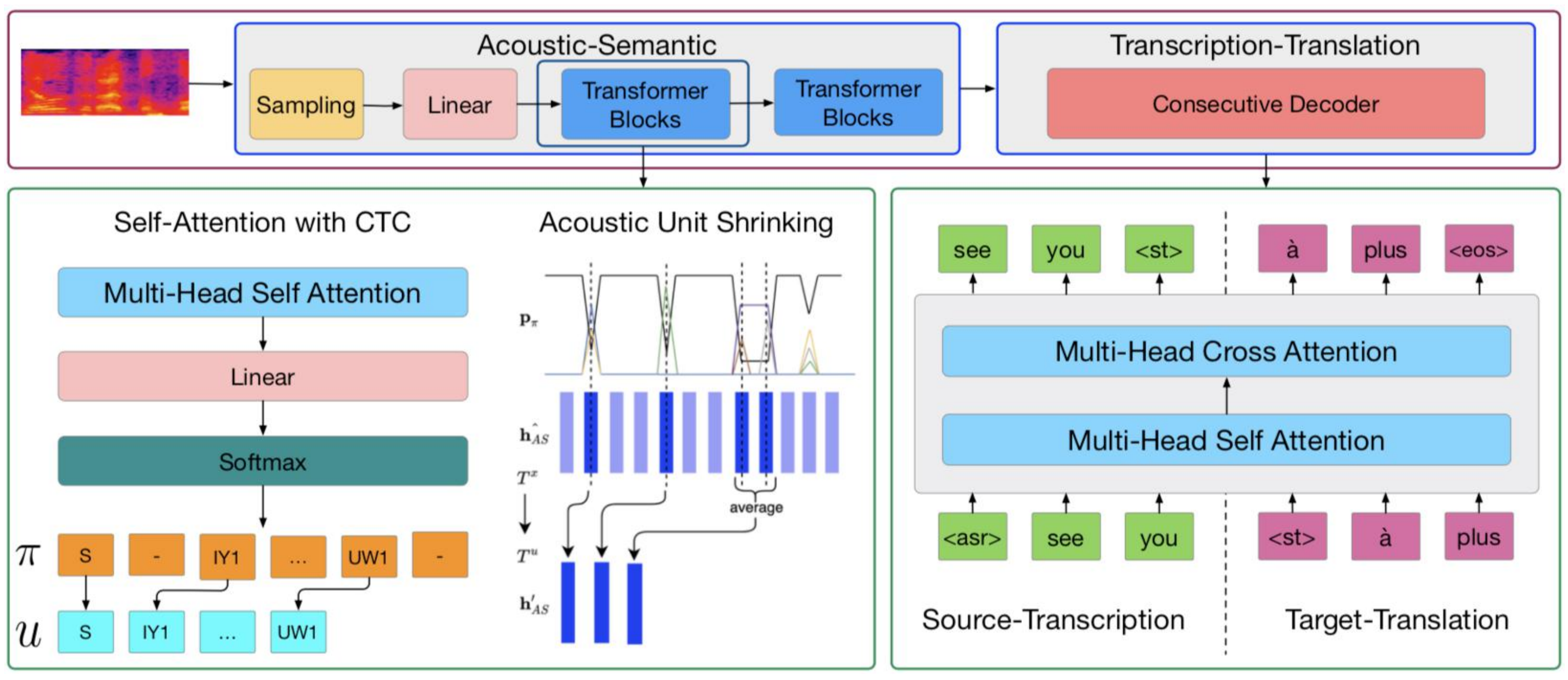




2 Listen, Understand and Translate



3 Consecutive Decoding



2.1 Methodology

Our proposed LUT consists of three modules, including an acoustic encoder, a semantic encoder and a translation decoder:

- ❑ An acoustic encoder network that encodes the audio input sequence into hidden features corresponding to the source text;
- ❑ A semantic encoder network that extracts hidden semantic representation for translation, which behaves like a normal machine translation encoder;
- ❑ A translation decoder network that emits sentence tokens in the target language.

2.2 Experiments

❑ Results on Librispeech English-French

Method	Enc Pre-train (speech data)	Dec Pre-train (text data)	greedy	beam
MT system				
Transformer MT	-	-	20.98	21.51
Base ST setting				
LSTM ST (Bérard et al. 2018)	✗	✗	12.30	12.90
+pre-train+multitask (Bérard et al. 2018)	✓	✓	12.60	13.40
LSTM ST+pre-train (Inaguma et al. 2020)	✓	✓	-	16.68
Transformer+pre-train (Liu et al. 2019a)	✓	✓	13.89	14.30
+knowledge distillation (Liu et al. 2019a)	✓	✓	14.96	17.02
TCEN-LSTM (Wang et al. 2019)	✓	✓	-	17.05
Transformer+ASR pre-train (Wang et al. 2020)	✓	✗	-	15.97
Transformer+curriculum pre-train (Wang et al. 2020)	✓	✗	-	17.66
LUT	✗	✗	**16.70**	**17.75**
Expanded ST setting				
LSTM+pre-train+SpecAugment (Bahar et al. 2019)	✓(236h)	✓	-	17.00
Multilingual ST+PT (Inaguma et al. 2019)	✓(472h)	✗	-	17.60
Transformer+ASR pre-train (Wang et al. 2020)	✓(960h)	✗	-	16.90
Transformer+curriculum pre-train (Wang et al. 2020)	✓(960h)	✗	-	18.01
LUT	✓(207h)	✗	**17.55**	**18.34**

❑ Results on TED English-Chinese

Method	Enc Pre-train (speech data)	Dec Pre-train (text data)	BLEU
MT system			
Transformer MT (Liu et al. 2019a)	-	-	27.08
Base setting			
Transformer+pre-train (Liu et al. 2019a)	✓	✓	16.80
+knowledge distillation (Liu et al. 2019a)	✓	✓	19.55
Multi-task+pre-train* (Inaguma et al. 2019)(re-implemented)	✓	✗	20.45
LUT	✗	✗	**20.84**

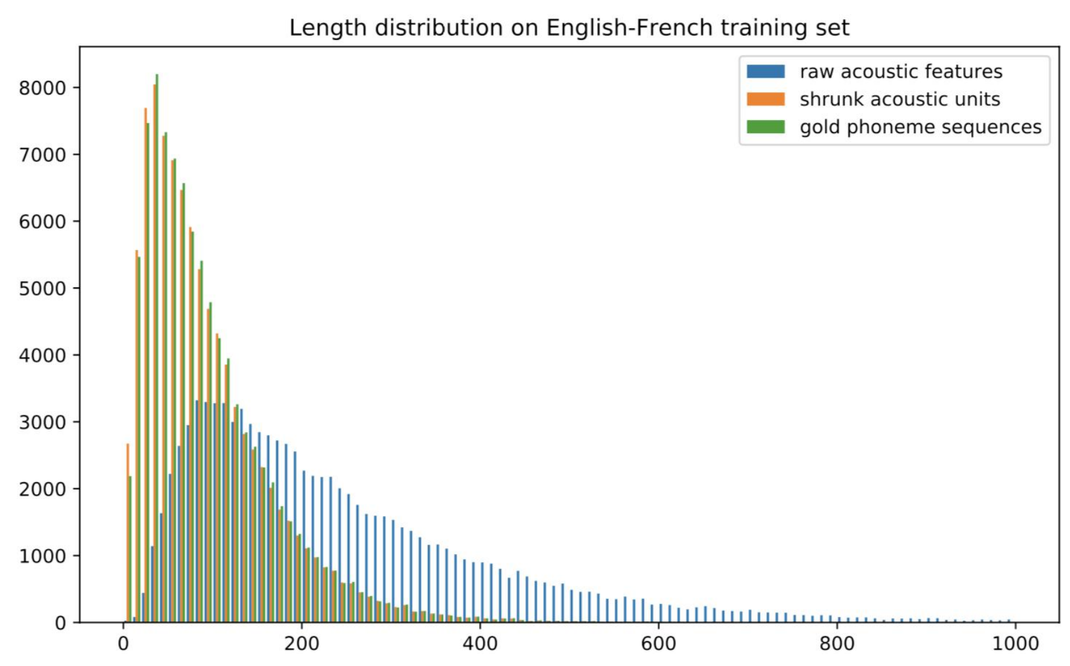
3.1 Methodology

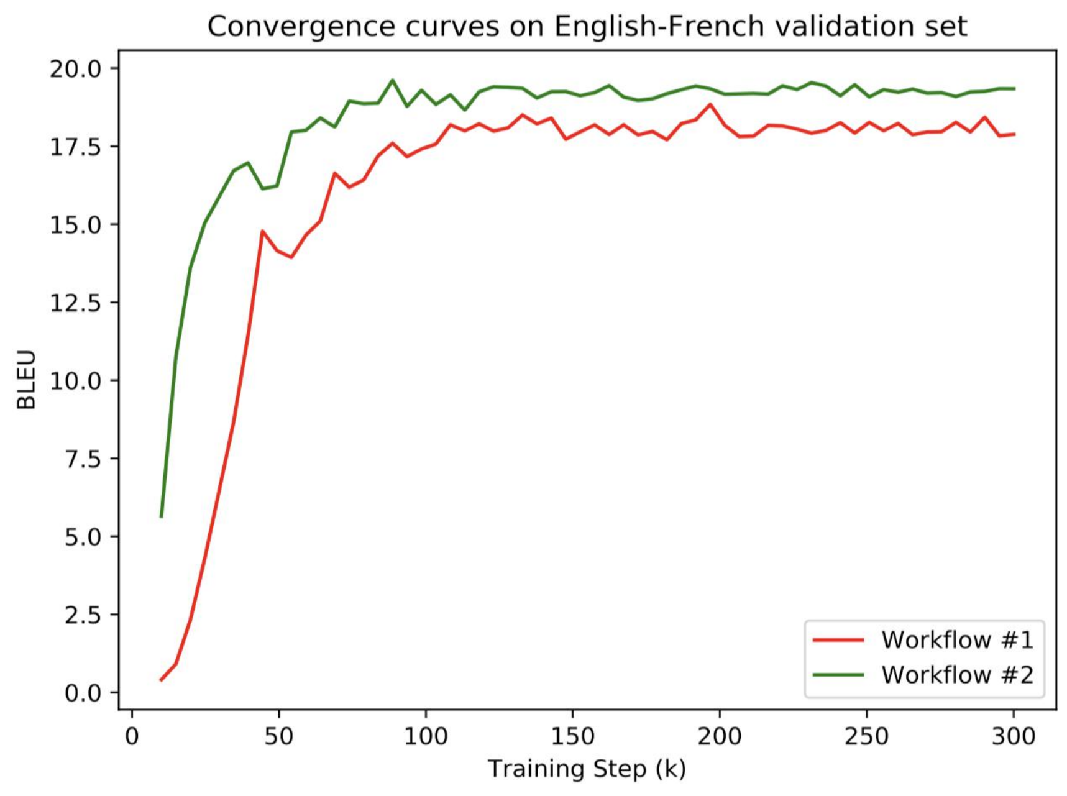
We divide our method COSTT into two phases, including the acoustic-semantic modeling phase (AS) and the transcription-translation modeling phase (TT).

- ❑ The AS phase accepts the speech features, outputs the acoustic representation, and encodes the shrunk acoustic representation into semantic representation.
- ❑ The TT phase accepts the AS's representation and consecutively outputs source transcription and target translation text sequences with a single shared decoder.

3.2 Experiments

❑ Effects of Shrinking Mechanism & Pre-training





❑ Results on IWSLT2018 English-German

Method	Enc Pre-train (speech data)	Dec Pre-train (text data)	tst2013
MT system			
Transformer MT	-	-	27.87
Base setting			
ESPnet (Inaguma et al. 2020)	✗	✗	12.50
+enc pre-train	✓	✗	13.12
+enc dec pre-train	✓	✓	13.54
Transformer+ASR pre-train (Wang et al. 2020)	✓	✗	15.35
+curriculum pre-train (Wang et al. 2020)	✓	✗	16.27
COSTT	✗	✗	**16.30**
Expanded setting			
Multi-task+pre-train (Inaguma et al. 2019)	✓(472h)	✗	14.60
CL-fast* (Kano, Sakti, and Nakamura 2018)	✓(479h)	✗	14.33
TCEN-LSTM (Wang et al. 2019)	✓(479h)	✓(40M)	17.67
Transformer+curriculum pre-train (Wang et al. 2020)	✓(479h)	✓(4M)	18.15
COSTT	✓(272h)	✓(1M)	**18.63**

主办方：中国中文信息学会青年工作委员会
承办方：智源社区