"Listen, Understand and Translate": Triple Supervision Decouples End-to-end Speech-to-text Translation

Qianqian Dong, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, Lei Li
Institute of Automation, Chinese Academy of Sciences, Beijing, China
ByteDance AI Lab, China

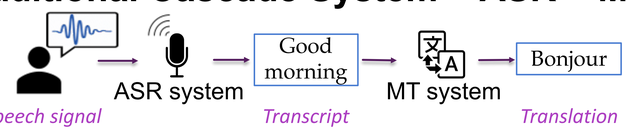
PAPER CODE

Background & Introduction

➢ **Task: Speech-to-text Translation (ST)**

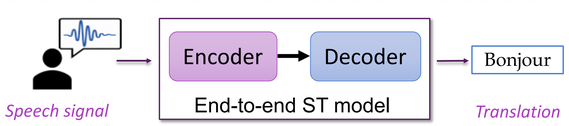
To read the ***audio signals of speech*** in one language, and translate to the ***text*** in another language. Speech translation has wide applications.

➢ **Traditional Cascade System = ASR + MT**



Challenges: **1.** Computationally inefficient
2. Error propagation issue

➢ **End-to-end model = encoder-decoder model**



Challenge: limited <*audio, transcript, translation*> Training data makes the enc-dec model hard to train

Predecessor's method: Pretraining
e.g. Pretrain the encoder using ASR task

Drawbacks:

1. A single encoder is hard to capture the representation of audio for the translation.
2. Limited in utilizing the information of *"transcription"* in the training.

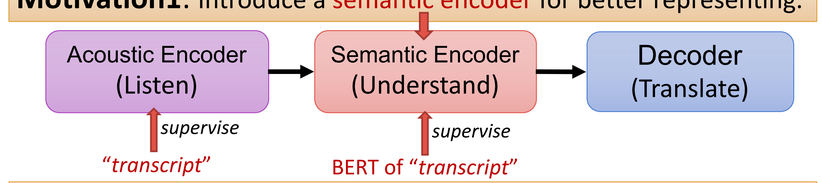
➢ **Q: How human translate?**

Our Contribution

We design "Listen-Understand-Translate"(LUT) based on ***human's behavior*** to fix ***two drawbacks*** of the end-to-end encoder-decoder ST model.

Motivations

Drawback1: A single encoder is not enough.
Motivation1: Introduce a semantic encoder for better representing.



Drawback2: Limit in using "transcript" info.
Motivation2: Utilizing the pretrained representation (e.g. BERT) of the "transcript" to capture the semantic feature.

Framework/Training of LUT

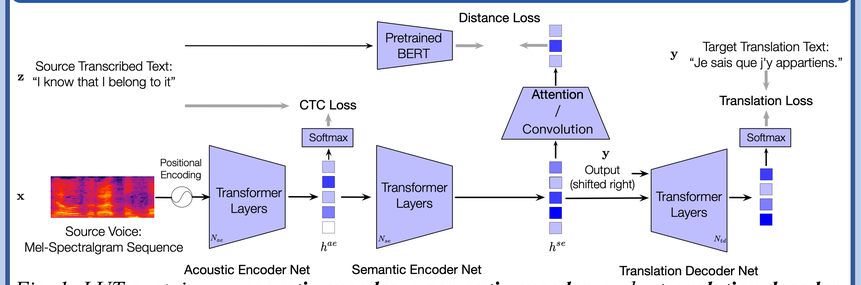


*Fig. 1: LUT contains an **acoustic encoder**, a **semantic encoder**, and a **translation decoder**.*

➢ **How to Listen? -- Acoustic encoder**

To recognize the transcript

CTC loss $L_{CTC} = -logP(z|x)$

$$P(z|x) = \sum_s P(s|x) = \sum_s softmax(h_{AE})$$

where *s* is from the possible "paths" given the sequence.

➢ **How to Understand? -- Semantic encoder**

To bridge the gap between the semantic hidden and pretrain representation of the transcript.

Distance loss $L_{distance} = MSE\big(v, h_{Bert}(z)\big)$,

where *v* is from the extra attention layer

➢ **How to Translate? -- Translation decoder**

To decode the translation.

Cross-entropy loss $L_{CE} = -\sum_i logp(y_i|y_{<i}, h_{SE})$

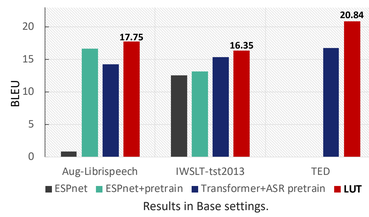
★ **Total Loss** = *CTC loss + Distance loss + CE loss*

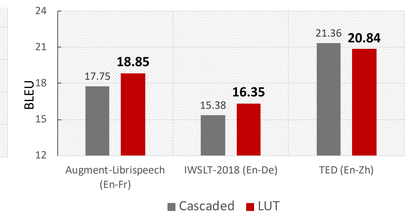
Experiments

We conduct experiments on three ST datasets

Datasets	Languages (From → To)	Training scales	Domains
Augmented Librispeech	En → Fr	236 hours	Audiobooks
IWSLT-2018	En → De	272 hours	Lecture
TED En-Zh	En → Zh	524 hours	Lecture

➢ **LUT achieves the SOTA** ***(left)***
➢ **LUT is better than the cascaded model** ***(right)***





➢ **Avoid Error Propagation!** LUT is Robust to the recognition errors, whereas the cascaded system fails.

		Case1	Case2
Transcription	Reference	it was mister jack maldon	chapter seventeen the abbes chamber
	CTC outputs	it was mister jack mal	chapter seventeen teen the abbes chamber
Output Translation		c'était monsieur Jack Maldon	chapitre xvii la chambre de l'abbé

➢ **Ablation Result**: CTC and Distance losses are useful, deep and balanced encoders are preferred. *(details see paper)*

➢ **We also do extra experiments and get more interesting findings!**

Finding1: Semantic encoder reveals context information!

	SpeakerVer	IntentIde
AT Output $\mathbf{h}^{ae}$	**97.6**	91.0
SE output $\mathbf{h}^{se}$	46.3	**93.1**

Tab.1: Classification accuracy on speaker and intent identification, using acoustic hidden and semantic hidden.

Finding2: Better ASR, Better ST!

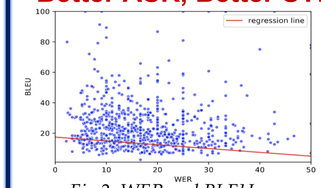


Fig.2: WER and BLEU are negatively correlated.

- ***More experiments results and details, please refer our paper.***
- *Any question, email:* dongqiangian2016@ia.ac.cn