BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Aligning Pre-trained Models for Spoken Language Translation

Bc. Šimon Sedláček
Santosh Kesiraju Ph.D.

Speech translation

Speech translation (ST) is the task of mapping an input **speech utterance** in a certain **source language** to its corresponding **text translation** in a given **target language**.

$$\text{Cascade:} \quad \text{speech (EN)} \xrightarrow{\text{ASR}} \text{transcript (EN)} \xrightarrow{\text{MT}} \text{translation (PT)}$$

$$\text{End-to-end:} \quad \text{speech (EN)} \xrightarrow{\text{encoder}} z \xrightarrow{\text{decoder}} \text{translation (PT)}$$

Figure 1: Cascade and end-to-end ST system architecture diagram. Speech encoder output representations z can be used for auxiliary training objectives.

Cascade ST systems:
- Higher latency
- Potential error accumulation due to their sequential nature
+ Requrire no training
+ Reliability scales with model size and domain robustness

End-to-end ST systems:
- Require speech translation training data
- Reliability -> bigger model -> more parameters to tune
+ Less prone to error accumulation, differentiable end-to-end
+ Lower latency

Is it possible to leverage powerful **off-the-shelf pre-trained** source language **ASR** and source-to-target language **MT** models for building new end-to-end ST systems **without training the whole system**?

Alignment architectures

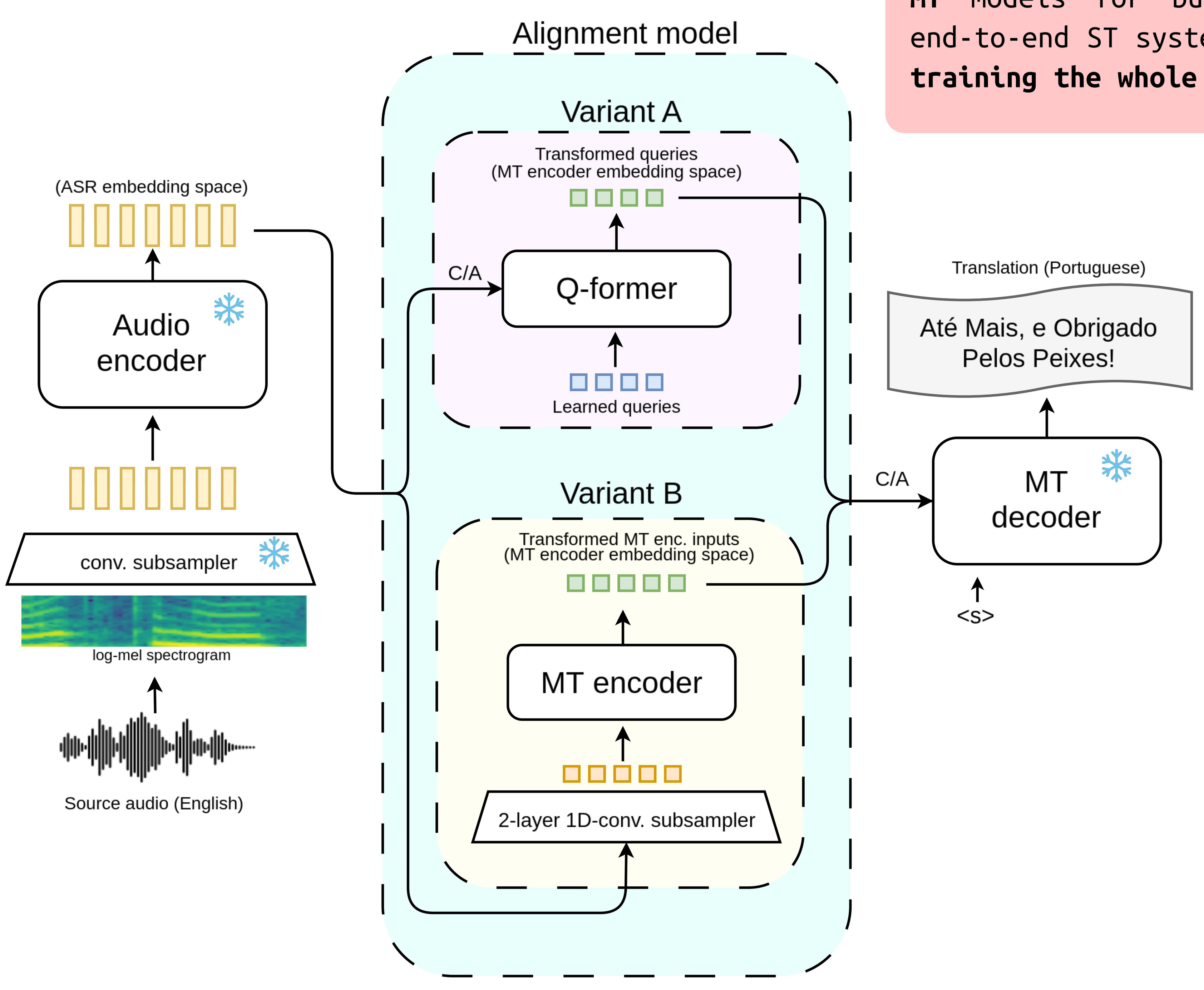

Figure 2: Architecture variants of the proposed aligned ST system.

Both the **ASR encoder** and the **MT decoder** are **frozen**. The original MT encoder is replaced with one of the following **alignment models**:

Variant A: **Q-former**
- Uses a set of 128 trainable **queries** as input
- Interacts with the speech embeddings via **cross-attention**
- Queries **extract information** from the ASR representations
- Variable-length to **fixed-length** sequence mapping

Variant B: **MT encoder**
- Audio representations are subsampled with a **conv. pooler**
- Initialized from the original **MT encoder**
- Trained to adapt to speech emgeddings
- Variable-length to **variable-length** mapping

Data and evaluation

All experiments were performed using the **How2 dataset**. The How2 dataset is a multi-modal corpus of English instructional videos, which contains a 300-hour speech subset. For this subset, there are also Portuguese translations in addition to English transcriptions. How2 is a standard speech-translation benchmark dataset.

Split	Videos	Hours	Clips/utterances	Per clip statistics
train	13,168	298.2	184,949	
val	150	3.2	2,022	5.8 seconds / 20 words
dev5	175	3.7	2,305	

Table 1: Statistics of the How2 dataset.

When evaluating the systems, the **BLEU metric** is used to measure the translation performance. BLEU measures the correspondece of generated machine translations to a set of given reference (preferably human) translations and ranges from 0 (worst) to 100 (virtually unattainable).

Base ASR and MT models

Model	In domain	How2 WER ↓		# params
		val	dev5	
CTC/attn. E-Branchformer base	Yes	12.6	12.2	38.5M
CTC/attn. E-Branchformer medium	No	12.1	11.7	174M

Table 2: Performance comparison of the ASR systems used in the alignment experiments on the How2 dataset. The *base* model was trained in-domain on the How2 corpus, the *medium* model is out-of-domain.

Model	In domain	How2 BLEU ↑		# params
		val	dev5	
MarianMT small	Yes	57.9	57.0	21.6M
T5-en-pt	No	40.0	38.8	223M

Table 3: Performance comparison of the MT systems used in the alignment experiments on the How2 dataset. The MarianMT model was trained in-domain on the How2 corpus, the T5 model is out-of-domain.

Results

Both the Q-former and the MT encoder alignment models are simple 6-layer transformer models with 4 attention heads and hidden size of 256.

Encoder	Connector	Decoder	How2 BLEU ↑		# trainable parameters
			val	dev5	
E-Branchformer base∗	Q-former	MarianMT∗	43.1	43.3	9.6M
	conv. + MT enc.		43.0	44.0	12.6M
E-Branchformer medium∗	Q-former		45.6	45.7	10.4M
	conv. + MT enc.		46.0	46.3	12.6M
E-Branchformer base∗	Q-Former	T5-en-pt∗	44.5	44.4	9.7M
E-Branchformer medium∗			**46.8**	**47.5**	10.5M
Baseline systems					
E-Branchformer base enc. (FT) + MarianMT dec. (FT)			45.6	45.2	38.5M
E-Branchformer base∗ + truecaser∗ + MarianMT∗			40.9	40.4	0

Table 4: Performance comparison of different ST systems trained with different alignment approaches. Modules annotated with '∗' are frozen.

The E-Branchformer/truecaser/MarianMT cascade system is outperformed by all of the aligned models, even though both the ASR and MT models are in-domain for the cascade system.

Aligned systems perform better with more capable speech encoders and language decoders, while the size of the alignment model stays constant.

The T5 decoder yields good results despite being out-of-domain on How2, which suggests that the alignment models can also serve as domain adapters.

Excel@FIT 2024