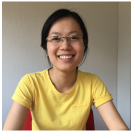
ICML 40 Years International Conference On Machine Learning

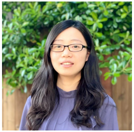
UGA Université Grenoble Alpes

Meta AI

Pre-training for Speech Translation: CTC Meets Optimal Transport



Phuong-Hang Le



Hongyu Gong

Changhan Wang

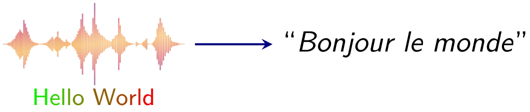
Juan Pino

Benjamin Lecouteux

Didier Schwab

Context and Motivation

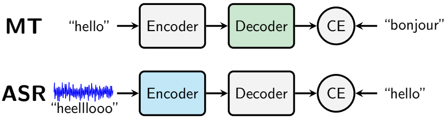
Task: **Speech-to-text translation (ST)**.



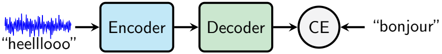
- Challenging, often requires two auxiliary tasks:
  - Automatic speech recognition **(ASR)**
  - Machine Translation **(MT)**.
- Two ways of using ASR and MT for ST: *pre-training* or *multi-task learning* (or both).

Two ways of using auxiliary tasks

Pre-training



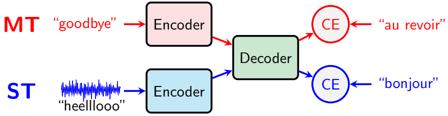
Then use the pre-trained components for **ST**



[Di Gangi et al., 2019, Wang et al., 2020b]

- ✓ Pre-train once, use many times
- ✗ Loss of pre-trained alignment information: MT encoder & ASR decoder discarded
- ✗ High modality gap

Multi-task learning



A simplified example of multi-task learning

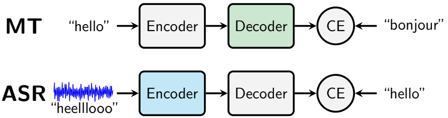
[Tang et al., 2021]

(CE: cross-entropy)

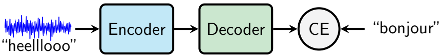
- ✓ Strong performance
- ✗ High ST training complexity.

Two ways of using auxiliary tasks

Pre-training

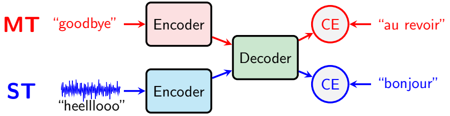


Then use the pre-trained components for **ST**



[Di Gangi et al., 2019, Wang et al., 2020b]

Multi-task learning



A simplified example of multi-task learning

[Tang et al., 2021]

(CE: cross-entropy)

- ✓ Pre-train once, use many times
- ✗ Loss of pre-trained alignment information: MT encoder & ASR decoder discarded
- ✗ High modality gap

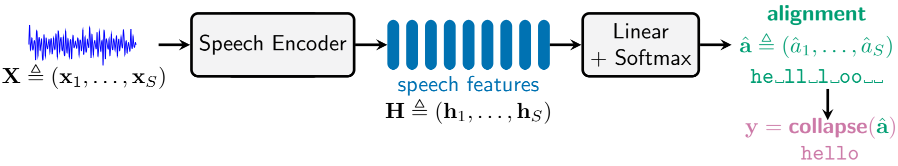
- ✓ Strong performance
- ✗ High ST training complexity.

Contributions

Siamese pre-training with CTC and Optimal Transport

- ✓ Pre-train once, use many times
- ✓ Low modality gap
- ✓ Strong performance
- ✓ Low ST training complexity.

Review of CTC

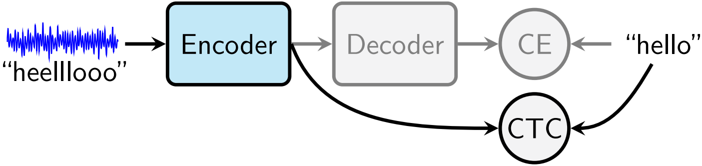


- CTC predicts a token $\hat{a}_t \in \mathcal{V}$ for each time step t:

$$p(a_t \mid \mathbf{X}) = \text{softmax}(\mathbf{W}\mathbf{h}_t + \mathbf{b})[a_t] \ \forall a_t \in \mathcal{V},$$
$$\hat{a}_t = \underset{a_t \in \mathcal{V}}{\text{argmax}}\, p(a_t \mid \mathbf{X}).$$

- For details (collapsing, Viterbi decoding, etc.), see [Graves et al., 2006].

CTC can reduce modality gap in pre-training



ASR pre-training with CTC. *CE is optional.*

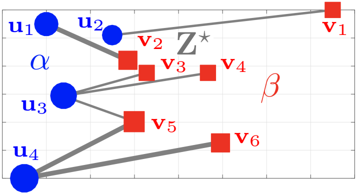
✓ ASR encoder trained with CTC already learns to align speech input to text output without a decoder.
→ **Pre-trained alignment information is preserved in encoder.**

✗ Solves “ASR decoder discarded” issue but not “MT encoder discarded”.

Review of discrete optimal transport

Problem: Transporting all masses of distribution α to distribution β.

- $\mathbf{a} \in \mathbb{R}_+^m, \mathbf{b} \in \mathbb{R}_+^n$: *masses* of α and β ($\mathbf{1}^\top \mathbf{a} = \mathbf{1}^\top \mathbf{b} = 1$).
- $\mathbf{u}_1, \ldots, \mathbf{u}_m \in \mathbb{R}^d$, $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^d$: *locations* of the masses $\mathbf{a}$ and $\mathbf{b}$.
- $c(\mathbf{u}_i, \mathbf{v}_j)$: *cost* of transporting a unit of mass from $\mathbf{u}_i$ to $\mathbf{v}_j$.



- $Z_{ij} \geq 0$: *quantity* of mass to be transported from $\mathbf{u}_i$ to $\mathbf{v}_j$.

OT finds ***transportation plan*** $\mathbf{Z}^*$ having minimum total cost:

$$\min_{\mathbf{Z} \in \mathbb{R}_+^{m \times n}} \quad \sum_{i=1}^{m} \sum_{j=1}^{n} Z_{ij} c(\mathbf{u}_i, \mathbf{v}_j),$$

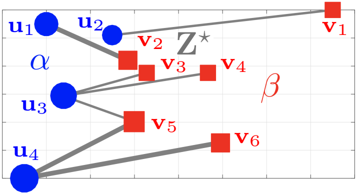
$$\text{s.t.} \quad \sum_{j=1}^{n} Z_{ij} = a_i \ \forall i, \qquad \sum_{i=1}^{m} Z_{ij} = b_j \ \forall j$$

(sum of row i is a_i, sum of column j is b_j)

Review of discrete optimal transport

Problem: Transporting all masses of distribution α to distribution β.

- $\mathbf{a} \in \mathbb{R}_+^m, \mathbf{b} \in \mathbb{R}_+^n$: *masses* of α and β $(\mathbf{1}^\top \mathbf{a} = \mathbf{1}^\top \mathbf{b} = 1)$.
- $\mathbf{u}_1, \ldots, \mathbf{u}_m \in \mathbb{R}^d$, $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^d$: *locations* of the masses $\mathbf{a}$ and $\mathbf{b}$.
- $c(\mathbf{u}_i, \mathbf{v}_j)$: *cost* of transporting a unit of mass from $\mathbf{u}_i$ to $\mathbf{v}_j$.



- $Z_{ij} \geq 0$: *quantity* of mass to be transported from $\mathbf{u}_i$ to $\mathbf{v}_j$.

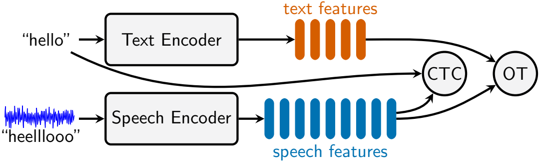
OT finds ***transportation plan*** $\mathbf{Z}^*$ having minimum total cost:

$$\min_{\mathbf{Z} \in \mathbb{R}_+^{m \times n}} \quad \sum_{i=1}^{m} \sum_{j=1}^{n} Z_{ij} c(\mathbf{u}_i, \mathbf{v}_j),$$

$$\text{s.t.} \quad \sum_{j=1}^{n} Z_{ij} = a_i \ \forall i, \quad \sum_{i=1}^{m} Z_{ij} = b_j \ \forall j$$

(sum of row i is a_i, sum of column j is b_j)

Learning to align speech and text features with OT



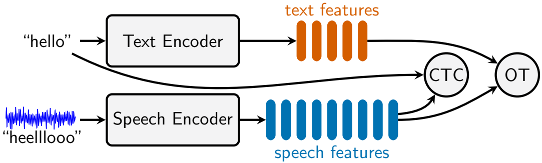
Siamese network for speech-text alignment

- Speech features $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$, text features $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$.
- Define *uniform* distributions α, β with masses located at $\mathbf{U}, \mathbf{V}$.
- The OT (or Wasserstein) loss is the minimum transportation cost:

$$\mathbf{OT}(\mathbf{U}, \mathbf{V}) = \min_{\mathbf{Z} \in \mathbb{R}_+^{m \times n}} \sum_{i=1}^{m} \sum_{j=1}^{n} Z_{ij} c(\mathbf{u}_i, \mathbf{v}_j) \quad \text{s.t.} \sum_{j=1}^{n} Z_{ij} = \frac{1}{m}, \sum_{i=1}^{m} Z_{ij} = \frac{1}{n}.$$

- OT pulls speech and text features *closer in Wasserstein space*.
- $\mathbf{Z}^*$ can be seen as an *alignment map* between the two sequences.

Learning to align speech and text features with OT



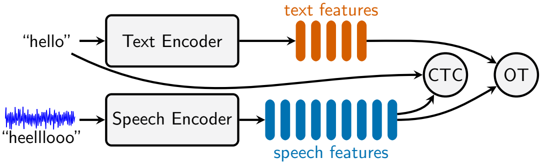
Siamese network for speech-text alignment

- Speech features $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$, text features $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$.
- Define *uniform* distributions α, β with masses located at $\mathbf{U}, \mathbf{V}$.
- The OT (or Wasserstein) loss is the minimum transportation cost:

$$\mathbf{OT}(\mathbf{U}, \mathbf{V}) = \min_{\mathbf{Z} \in \mathbb{R}_+^{m \times n}} \sum_{i=1}^{m} \sum_{j=1}^{n} Z_{ij} c(\mathbf{u}_i, \mathbf{v}_j) \quad \text{s.t.} \sum_{j=1}^{n} Z_{ij} = \frac{1}{m}, \sum_{i=1}^{m} Z_{ij} = \frac{1}{n}.$$

- OT pulls speech and text features *closer in Wasserstein space*.
- $\mathbf{Z}^*$ can be seen as an *alignment map* between the two sequences.

Learning to align speech and text features with OT



Siamese network for speech-text alignment

- Speech features $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_m)$, text features $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_n)$.
- Define *uniform* distributions α, β with masses located at $\mathbf{U}, \mathbf{V}$.
- The OT (or Wasserstein) loss is the minimum transportation cost:

$$\mathbf{OT}(\mathbf{U}, \mathbf{V}) = \min_{\mathbf{Z} \in \mathbb{R}_+^{m \times n}} \sum_{i=1}^{m} \sum_{j=1}^{n} Z_{ij} c(\mathbf{u}_i, \mathbf{v}_j) \quad \text{s.t.} \sum_{j=1}^{n} Z_{ij} = \frac{1}{m}, \ \sum_{i=1}^{m} Z_{ij} = \frac{1}{n}.$$

- OT pulls speech and text features *closer in Wasserstein space*.
- $\mathbf{Z}^*$ can be seen as an *alignment map* between the two sequences.

Positional encoding for optimal transport

- **Motivation:** OT loss ignores sequence orders, while speech/text inputs are *monotonically* aligned.
- **Idea:** Integrating normalized positions $s_i = \frac{i-1}{m-1}$ and $t_j = \frac{j-1}{n-1}$ into cost function:

$$c(\mathbf{u}_i, \mathbf{v}_j) = \left(\|\mathbf{u}_i - \mathbf{v}_j\|_p^p + \gamma^p \left| s_i - t_j \right|^p \right)^{1/p}.$$

- **Intuition:** *Mismatches in position will be penalized due to high cost.*
  - This favors aligning $\mathbf{u}_1 \to \mathbf{v}_1$ instead of $\mathbf{u}_1 \to \mathbf{v}_n$, for example.

Positional encoding for optimal transport

- **Motivation:** OT loss ignores sequence orders, while speech/text inputs are *monotonically* aligned.
- **Idea:** Integrating normalized positions $s_i = \frac{i-1}{m-1}$ and $t_j = \frac{j-1}{n-1}$ into cost function:

$$c(\mathbf{u}_i, \mathbf{v}_j) = \left(\|\mathbf{u}_i - \mathbf{v}_j\|_p^p + \gamma^p \left|s_i - t_j\right|^p\right)^{1/p}.$$

- **Intuition:** *Mismatches in position will be penalized due to high cost.*
  - This favors aligning $\mathbf{u}_1 \to \mathbf{v}_1$ instead of $\mathbf{u}_1 \to \mathbf{v}_n$, for example.

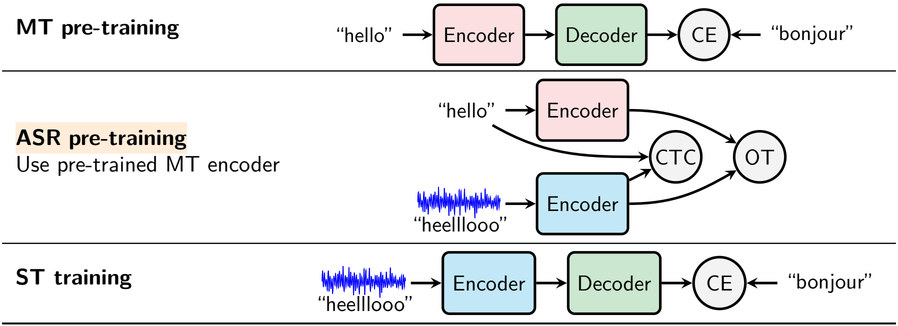
Positional encoding for optimal transport

- **Motivation:** OT loss ignores sequence orders, while speech/text inputs are *monotonically* aligned.
- **Idea:** Integrating normalized positions $s_i = \frac{i-1}{m-1}$ and $t_j = \frac{j-1}{n-1}$ into cost function:

$$c(\mathbf{u}_i, \mathbf{v}_j) = \left(\|\mathbf{u}_i - \mathbf{v}_j\|_p^p + \gamma^p \left| s_i - t_j \right|^p \right)^{1/p}.$$

- **Intuition:** *Mismatches in position will be penalized due to high cost*.
  - This favors aligning $\mathbf{u}_1 \to \mathbf{v}_1$ instead of $\mathbf{u}_1 \to \mathbf{v}_n$, for example.

Proposed recipe for speech translation

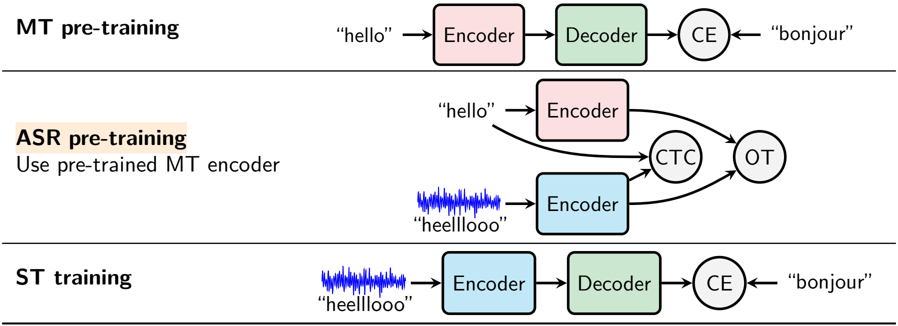


Proposed ASR & MT pre-training recipe

✓ Using all pre-trained components → preserving learned alignment information.

✓ OT reduces modality gap by aligning speech and text features.

Proposed recipe for speech translation



Proposed ASR & MT pre-training recipe

✓ Using all pre-trained components → preserving learned alignment information.

✓ OT reduces modality gap by aligning speech and text features.

Summary of main experimental results

Main results on standard benchmarks **MuST-C** [Di Gangi et al., 2019] and **CoVoST-2** [Wang et al., 2020b]:

- Siamese pre-training (**Siamese-PT**) can use other differentiable distances (e.g., Euclidean distance, KL-divergence), but OT achieves best results.
- Siamese-PT outperforms pre-training with CE, or CTC, or CTC+CE.
- With only *vanilla encoder-decoder* and even *without external data*, our method is competitive with recent SoTA methods.
- Siamese-PT can be applied on top of strong multi-task learning systems [Tang et al., 2021], leading to further improvements.

Comparison to state-of-the-art results

Method		Multi	External Data		BLEU								
			Unlabeled	*Labeled*	**de**	**es**	**fr**	**it**	**nl**	**pt**	**ro**	**ru**	**avg**
FAIRSEQ S2T [Wang et al., 2020a]		✓	-	-	24.5	28.2	34.9	24.6	28.6	31.1	23.8	16.0	26.5
ESPnet-ST [Inaguma et al., 2020]		✓	-	-	22.9	28.0	32.7	23.8	27.4	28.0	21.9	15.8	25.1
Dual-decoder [Le et al., 2020]		✓	-	-	23.6	28.1	33.5	24.2	27.6	30.0	22.9	15.2	25.6
Adapters [Le et al., 2021]		✓	-	-	24.7	28.7	35.0	25.0	28.8	31.1	23.8	16.4	26.6
BiKD [Inaguma et al., 2021]		-	-	-	25.3	-	35.3	-	-	-	-	-	-
JointSpeechText [Tang et al., 2021]		-	-	✓	26.8	31.0	37.4	-	-	-	-	-	-
TaskAware [Indurthi et al., 2021]		-	-	✓	**28.9**	-	-	-	-	-	-	-	-
ConST [Ye et al., 2022]		-	✓	✓	28.3	32.0	38.3	27.2	**31.7**	33.1	25.6	**18.9**	29.4
STPT [Tang et al., 2022]		-	✓	✓	-	**33.1**	**39.7**	-	-	-	-	-	-
CE pre-training	MEDIUM	✓	-	-	24.6	28.7	34.9	24.6	28.4	30.7	23.7	15.9	26.4
CTC pre-training	MEDIUM	✓	-	-	25.9	29.7	36.6	25.6	29.6	32.0	24.6	16.7	27.6
CTC+CE pre-training	MEDIUM	✓	-	-	25.6	29.5	36.4	25.2	29.5	31.6	24.5	16.5	27.4
Siamese-PT (this work)	MEDIUM	✓	-	-	26.2	29.8	36.9	25.9	29.8	32.1	24.8	16.8	27.8
CE pre-training	LARGE	✓	-	-	26.9	30.8	37.7	26.7	30.8	33.3	26.2	17.9	28.8
CTC pre-training	LARGE	✓	-	-	27.6	31.4	38.2	27.2	31.1	33.6	26.4	18.4	29.2
CTC+CE pre-training	LARGE	✓	-	-	27.2	31.2	38.0	27.0	31.5	33.7	26.2	18.3	29.1
Siamese-PT (this work)	LARGE	✓	-	-	27.9	31.8	39.2	**27.7**	**31.7**	**34.2**	**27.0**	18.5	**29.8**

BLEU on test sets of MuST-C

- By simply increasing model size, our method applied to *vanilla encoder-decoder architecture without external data* performs on par with strong multi-task learning systems trained with external data.

Main Takeaways

- *Encoder trained with CTC is stronger* than the one trained with encoder-decoder-CE.
- Siamese pre-training with CTC and optimal transport helps *reduce modality gap without any changes in the ST model.*
- Optimal transport is very effective for *learning to align sequences of features from different modalities*.

Thank you for your attention!

Please read our paper for more details.

cha Code and pre-trained models:

`https://github.com/formiel/fairseq`.