Non-Parametric Domain Adaptation for End-to-End Speech Translation

Anonymous ACL submission

Abstract

The end-to-end speech translation (E2E-ST) has received increasing attention due to the potential of its less error propagation, lower latency, and fewer parameters. However, the effectiveness of neural-based approaches to this task is severely limited by the available training corpus, especially for domain adaptation where in-domain triplet training data is scarce or nonexistent. In this paper, we propose a novel non-parametric method that leverages domain-specific text translation corpus to achieve domain adaptation for E2E-ST system. To this end, we first incorporate an additional encoder into the pre-trained E2E-ST model to realize text translation modelling and then unify the decoder's output representation for text and speech translation tasks by reducing the correspondent representation mismatch in available triplet training data. During domain adaptation, a k-nearest-neighbor (kNN) classifier is introduced to produce the final translation distribution using the external datastore built by the domain-specific text translation corpus, while the universal output representation is adopted to perform a similarity search. Experiments on the Europarl-ST benchmark demonstrate that when in-domain text translation data is used only, our proposed approach significantly improves the baseline by 12.82 BLEU on average in all translation directions.

1 Introduction

Speech translation (ST), the task of automatically translating a speech signal in a given language into a text in another language, is a widely studied topic thanks to the increasing demand for international communications. Traditional ST systems cascade automatic speech recognition (ASR) and machine translation (MT) (Ney, 1999; Sperber et al., 2017; Zhang et al., 2019; Iranzo-Sánchez et al., 2020a; Machácek et al., 2021). Recently, various large-scale speech-translation datasets have been proposed, e.g., Libri-Trans (Kocabiyikoglu et al., 2018), MuST-C (Gangi et al., 2019), and CoVoST (Wang et al., 2020a). With these large-scale annotations, building an end-to-end speech translation (E2E-ST) system (Vila et al., 2018; Liu et al., 2019; Li et al., 2021; Han et al., 2021; Dong et al., 2021) has become popular, since it has lower latency and less error propagation compared with previous ST methods. Recent researches have also shown that there is no significant difference between end-to-end models and cascaded systems in translation performance (Bentivogli et al., 2021).

In many practical application scenarios, such as political negotiations, business meetings, etc., there is no available domain-specific speech-translation data to conduct the end-to-end training, which essentially limits the promotion of E2E-ST systems. The general practice is that the E2E-ST model learns knowledge well enough in the general domain, and then we directly use it to translate speech input in the target domain. Unfortunately, due to the domain shift issue (Gretton et al., 2006; Ramponi and Plank, 2020), the generalization capabilities of current end-to-end models are weak across different scenarios. Instead of speech-translation annotations, the bilingual text in the target domain is usually abundant and easy to collect. Thus, it is essential to explore the capability of E2E-ST system in this scenario, in which a large amount of in-domain bilingual text is utilized.

In this work, we focus on this domain adaptation setting and aim to replace the domain-specific parameter updating in the neural-based E2E-ST model with a non-parametric search to make it adaptable and achieve domain adaptation without any speech-translation annotations. Actually, the non-parametric approach kNN-MT, recently proposed by Khandelwal et al. (2021), is a promising alternative to achieve this goal. The kNN-MT equips a pre-trained MT model with a k-nearest-neighbor (kNN) classifier over an external datastore to improve translation accuracy without retraining. However, it requires the in-domain speech-translation corpus to construct an effective datastore when we apply this method in the speech translation setting. To tackle this problem, we propose a novel **N**on-**P**arametric **D**omain **A**daptation framework based on kNN-MT for E2E-ST, named as NPDA-kNN-ST. The key core of this method is to directly leverage in-domain text translation corpus to generate the corresponding datastore, and encourage it to play a similar role with the real in-domain speech-translation data, through the carefully designed architecture and loss function.

Specifically, we first incorporate an additional trainable encoder for text modelling into the pre-trained E2E-ST model. Based on that, we further make the decoder's output representation for text and speech translation tasks close, by reducing the representation inconsistency of these two tasks in training triplet data and keeping the parameters of the original pre-trained E2E-ST model fixed. In this way, the additional encoder module learns the semantic mapping in feature space from the source language text to the speech signal, which enables the construction of an effective in-domain datastore when only text translation data is used. Then, we introduce a kNN classifier to produce the final translation distribution based on the in-domain datastore built by the correspondent text translation data. Meanwhile, the universal output representation is used to perform a similarity search and guide the translation process.

We evaluate the effectiveness of our method on the Europarl-ST benchmark, and demonstrate that our approach significantly improves the baseline by 12.82 BLEU on average in all translation directions when only using large-scale in-domain text translation data. Besides, additional experiments on Europarl-ST and MuST-C datasets show that the in-domain text translation datastore generated by our method can play a similar role with the real in-domain speech-translation data, thanks to the universal output representation.

2 Background

In this section, we first give a formal definition of the E2E-ST task and then briefly introduce the nearest neighbor machine translation (kNN-MT).

2.1 End-to-End Speech Translation

The E2E-ST receives speech signals in a source language and directly generates the text in a target language without an intermediate transcription process. Concretely, the E2E-ST corpus consists of a set of triplet data $\mathcal{D}_{ST} = \left\{(\mathbf{x}^{(n)}, \mathbf{z}^{(n)}, \mathbf{y}^{(n)})\right\}_{n=1}^{N}$, where $\mathbf{x}^{(n)} = (x_1^{(n)}, x_2^{(n)}, ..., x_{|\mathbf{x}^{(n)}|}^{(n)})$ is the input sequence of the speech wave (in most cases, acoustic features are used), $\mathbf{z}^{(n)} = (z_1^{(n)}, z_2^{(n)}, ..., z_{|\mathbf{z}^{(n)}|}^{(n)})$ represents the transcription sequence from the source language and $\mathbf{y}^{(\mathbf{n})} = (y_1^{(n)}, y_2^{(n)}, ..., y_{|\mathbf{y}^{(n)}|}^{(n)})$ denotes the translation sequence of target language. The goal of E2E-ST model is to seek an optimal translation sequence $\mathbf{y}$ without generating an intermediate transcription $\mathbf{z}$, and the standard training objective is to optimize the maximum likelihood estimation (MLE) loss of the training data:

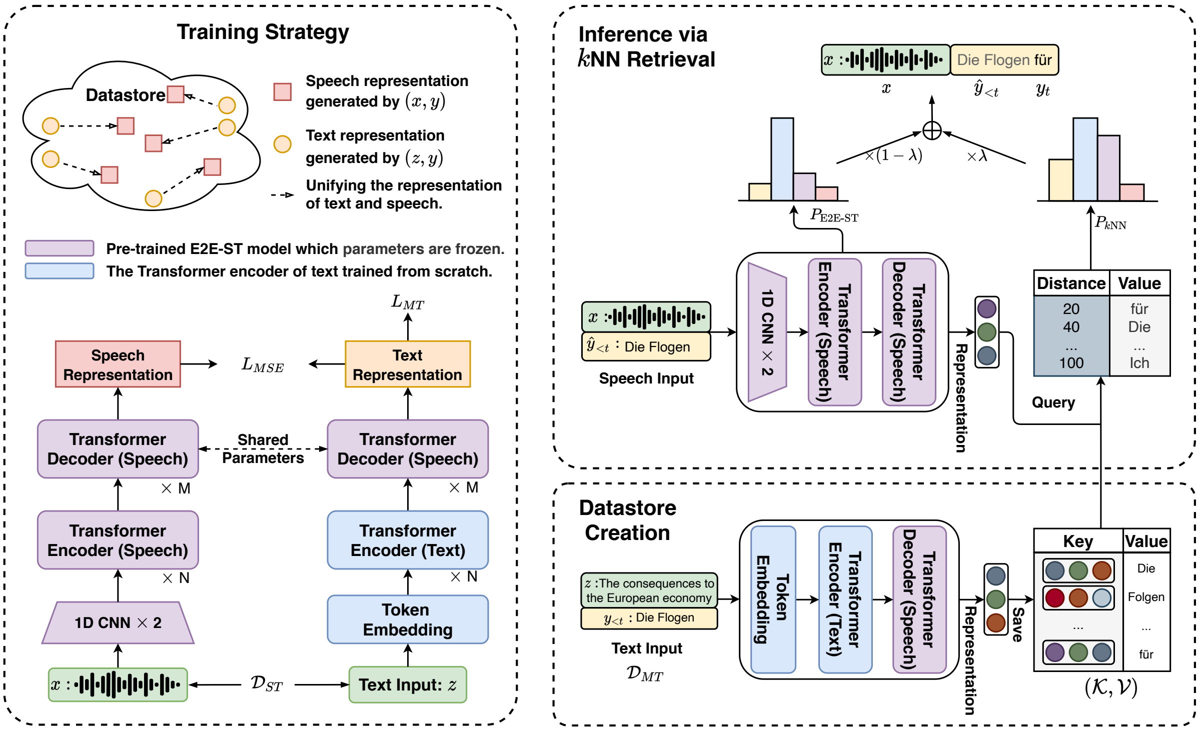
$$\mathcal{L}_{ST}(\theta) = \frac{1}{N} \sum_{n=1}^{N} \log P\left(\mathbf{y}^{(\mathbf{n})} \mid \mathbf{x}^{(\mathbf{n})}; \theta\right), \quad (1)$$

where we adopt a single encoder-decoder structure to fit the conditional distribution $P(\mathbf{y}^{(\mathbf{n})}|\mathbf{x}^{(\mathbf{n})})$ and θ is the model parameter. In order to develop the high-quality E2E-ST system, the ASR and MT tasks $((\mathbf{x}^{(n)}, \mathbf{z}^{(n)})$ and $(\mathbf{z}^{(n)}, \mathbf{y}^{(n)}))$ are typically used to pre-train the encoder and decoder, respectively (Bansal et al., 2019; Wang et al., 2020c). However, in practice, it is not realistic to obtain a large amount of high-quality speech-translation data in every domain that we are interested in, while in-domain text translation corpus is usually cheaper and easier to collect. Thus, it is essential to investigate the capability of the E2E-ST model that only uses large-scale in-domain text translation corpus to achieve domain adaptation, making the E2E-ST system more practical.

2.2 Nearest Neighbor Machine Translation

Recently, Khandelwal et al. (2021) proposed a non-parametric method kNN-MT, which has shown the promising capability of directly augmenting the pre-trained neural machine translation (NMT) model with domain-specific token-level kNN retrieval to improve the translation performance without retraining. The kNN-MT mainly involves two steps: datastore creation and token inference with cached datastore.

Datastore Creation. The datastore of kNN-MT is the cache of a set of key-value pairs. Given a parallel sentence pair $(\mathbf{z}, \mathbf{y}) \in (\mathcal{Z}, \mathcal{Y})$, the pre-trained NMT model generates the context representation $f_\theta(\mathbf{z}, y_{<t})$ at each timestep t. Then the whole datastore $(\mathcal{K}, \mathcal{V})$ is constructed by taking the output


Training Strategy
Datastore
Speech representation generated by (x, y)
Text representation generated by (z, y)
Unifying the representation of text and speech.
Pre-trained E2E-ST model which parameters are frozen.
The Transformer encoder of text trained from scratch.
L_{MT}
Speech Representation
L_{MSE}
Text Representation
Transformer Decoder (Speech)
Shared Parameters
Transformer Decoder (Speech)
× M
× M
Transformer Encoder (Speech)
Transformer Encoder (Text)
× N
× N
1D CNN × 2
Token Embedding
x :
$\mathcal{D}_{ST}$
Text Input: z
Inference via kNN Retrieval
x :
Die Flogen für
x
$\hat{y}_{<t}$
y_t
$\times(1-\lambda)$
$\times\lambda$
$P_{\text{E2E-ST}}$
$P_{k\text{NN}}$
x :
$\hat{y}_{<t}$: Die Flogen
Speech Input
1D CNN × 2
Transformer Encoder (Speech)
Transformer Decoder (Speech)
Representation
Distance
Value
20
40
...
100
für
Die
...
Ich
Query
Datastore Creation
z :The consequences to the European economy
$y_{<t}$: Die Flogen
Text Input
$\mathcal{D}_{MT}$
Token Embedding
Transformer Encoder (Text)
Transformer Decoder (Speech)
Representation
Save
Key
Value
Die
Folgen
...
für
...
$(\mathcal{K}, \mathcal{V})$

(a) Unifying text and speech representation.

(b) Performing domain adaptation via kNN retrieval.

Figure 1: The overview of our non-parametric domain adaptation framework for E2E-ST (NPDA-kNN-ST).

hidden states $f_\theta(\mathbf{z}, y_{<t})$ as key and y_t as value:

$$(\mathcal{K}, \mathcal{V}) = \bigcup_{(\mathbf{z},\mathbf{y})\in(\mathcal{Z},\mathcal{Y})} \{(f_\theta(\mathbf{z}, y_{<t}), y_t), \forall y_t \in \mathbf{y}\}. \tag{2}$$

Inference via kNN Retrieval. In the inference stage, the kNN-MT model predicts the probability distribution of t-th target token $\hat{y}_t$ with the context representation $f_\theta(\mathbf{z}, \hat{y}_{<t})$. Specifically, kNN-MT utilizes the context representation to query the cached datastore $(\mathcal{K}, \mathcal{V})$ and retrieves k nearest neighbor key-value pairs w.r.t Euclidean distance. Then the probability distribution of $\hat{y}_t$ generated by kNN retrieval is calculated as follow:

$$\begin{aligned} p_{k\text{NN}}(\hat{y}_t|\mathbf{z}, \hat{y}_{<t}) \propto \\ \sum_{(h_i,v_i)\in\mathcal{R}} \mathbb{1}_{y_t=v_i} \exp\left(\frac{-d(h_i, f_\theta(\mathbf{z}, \hat{y}_{<t}))}{T}\right), \end{aligned} \tag{3}$$

where $\mathcal{R} = \{(h_i, v_i), i \in \{1, 2, ..., k\}\}$ is the set of k nearest neighbors, $d(\cdot, \cdot)$ represents the Euclidean distance, and T is the temperature to control the sharpness of the softmax function. The final output distribution is an interpolation between distributions from the NMT model and the kNN retrieved neighbors with a tuned parameter $\lambda \in [0, 1]$:

$$\begin{aligned} p(\hat{y}_t|\mathbf{z}, \hat{y}_{<t}) = \lambda\, p_{k\text{NN}}(\hat{y}_t|\mathbf{z}, \hat{y}_{<t}) \\ + (1-\lambda)\, p_{\text{NMT}}(\hat{y}_t|\mathbf{z}, \hat{y}_{<t}). \end{aligned} \tag{4}$$

3 Method

When we apply kNN-MT in the speech translation task, it needs the real speech-translation corpus to build an effective datastore for kNN retrieval. However, this requirement could not be satisfied in the domain adaptation scenario mentioned before, as there is no available domain-specific speech-translation corpus. In this paper, we focus on this domain adaptation setting and attempt to replace the domain-specific parameter updating with a non-parametric search to achieve domain adaptation. To this end, we design a novel **N**on-**P**arametric **D**omain **A**daptation framework based on kNN-MT for E2E-ST, named as NPDA-kNN-ST. The overview framework of NPDA-kNN-ST is illustrated in Figure 1, mainly divided into two parts: a) unifying text and speech representation to enable datastore creation; b) performing domain adaptation through kNN retrieval. Next, we will introduce the model architecture, training objective, and inference process of our approach in detail.

3.1 Unifying Text and Speech Representation

The NPDA-kNN-ST aims to directly build an in-domain effective datastore with only text translation corpus, and make it play a similar role with the real in-domain speech-translation data. It means that whether word tokens or speech signals are taken as input, we should construct the universal output representation for them in a unified model. As shown in Figure 1(a), we introduce an additional transformer encoder and reuse the original transformer decoder of the pre-trained E2E-ST model for source text modelling, by which we only increase a few parameters for our method.

Based on this model structure, we try to make the decoder's output representation for text and speech translation tasks close, by which the text translation data can be used to build an effective in-domain datastore. We achieve this by leveraging out-of-domain triplet data $\mathcal{D}_{ST}$. More specifically, given a triplet data point in the corpus $(\mathbf{x}, \mathbf{z}, \mathbf{y}) \in \mathcal{D}_{ST}$, the original E2E-ST model takes speech-translation pair $(\mathbf{x}, \mathbf{y})$ as input and generates output representation $f_{(\theta_e,\theta_d)}(\mathbf{x};\ y_{<t})$ for each target token y_t. Meanwhile, with corresponding text translation pair $(\mathbf{z}, \mathbf{y})$, the model with an additional transformer encoder produces another representation for y_t, which can be denoted as $f_{(\theta'_e,\theta_d)}(\mathbf{z};\ y_{<t})$. Then, we take the end-to-end paradigm, and directly update the introduced transformer encoder by minimizing the squared euclidean distance of the two sets of decoder representations and optimizing MLE loss of text translation pair:

$$\begin{aligned}\mathcal{L}_{MSE}(\theta'_e) = & \frac{1}{|\mathcal{D}_{ST}|} \sum_{(\mathbf{x},\mathbf{z},\mathbf{y}) \in \mathcal{D}_{ST}} \\ & \sum_t ||f_{(\theta_e,\theta_d)}(\mathbf{x}; y_{<t}) - f_{(\theta'_e,\theta_d)}(\mathbf{z}; y_{<t})||^2, \\ \mathcal{L}_{MT}(\theta'_e) = & \frac{1}{N} \sum_{n=1}^{N} \log P\left(\mathbf{y^{(n)}} \mid \mathbf{z^{(n)}}; \theta'_e, \theta_d\right), \\ \mathcal{L}(\theta'_e) = & \mathcal{L}_{MT}(\theta'_e) - \mathcal{L}_{MSE}(\theta'_e),\end{aligned} \tag{5}$$

where θ_e and θ_d are parameters of encoder and decoder in the pre-trained E2E-ST model respectively, θ'_e represents the parameter of the new transformer encoder and token embedding, and we keep θ_e and θ_d fixed during training to avoid the E2E-ST performance degradation in the inference stage. The out-of-domain validation set and its correspondent loss are adopted to select the best model in our experiments.

3.2 Domain Adaptation via kNN Retrieval

We consider the domain adaptation scenario of E2E-ST that only domain-specific text translation corpus $\mathcal{D}_{MT} = \left\{(\mathbf{z}^{(n)}, \mathbf{y}^{(n)})\right\}_{m=1}^{M}$ is available. During domain adaptation, the entire inference process is illustrated in Figure 1(b). Once we gain the well-trained model with Equation 5, the new transformer encoder and original transformer decoder of the E2E-ST model are utilized to forward pass the text translation corpus to create an in-domain datastore $(\mathcal{K}, \mathcal{V})$. This construction process is the same as the kNN-MT. Due to the universal decoder's representation, this datastore could directly be used for the in-domain kNN retrieval when translating speech input $\mathbf{x}$. The final probability of NPDA-kNN-ST to predict the next token $\hat{y}_t$ is an interpolation of two distributions with a hyper-parameter λ:

$$\begin{aligned}p'(\hat{y}_t|\mathbf{x}, \hat{y}_{<t}) = & \lambda\, p_{k\text{NN}}(\hat{y}_t|\mathbf{x}, \hat{y}_{<t}) \\ & + (1-\lambda)\, p_{\text{E2E-ST}}(\hat{y}_t|\mathbf{x}, \hat{y}_{<t}),\end{aligned} \tag{6}$$

where $p_{\text{E2E-ST}}$ indicates the general domain E2E-ST prediction and $p_{k\text{NN}}$ represents the in-domain retrieval based on Equation 3. Actually, this prediction way can also be replaced with other kNN variants (Zheng et al., 2021; He et al., 2021).

4 Experiments

4.1 Setup

We conduct experiments to evaluate our proposed method in two aspects: a) domain adaptation on Europarl-ST benchmark with the pre-trained E2E-ST model on MuST-C dataset; b) the performance comparisons on MuST-C benchmark when speech-translation and text-translation data are leveraged to build datastore, respectively.

MuST-C Datasets. MuST-C (Gangi et al., 2019) is a publicly available large-scale multilingual ST corpus, consisting of triplet data sources: source speech, source transcription, and target translation. The speech sources of MuST-C are from English TED Talks, which are aligned at the sentence level with their manual transcriptions and translations. MuST-C contains translations from English (EN) to 8 languages: Dutch (NL), French (FR), German (DE), Italian (IT), Portuguese (PT), Romanian (RO), Russian (RU), and Spanish (ES). The statistics of different language pairs are illustrated in Appendix A.1.

Europarl-ST Datasets. Europarl-ST (Iranzo-Sánchez et al., 2020b) collects from the official transcriptions and translations of European Parliament debate. In order to evaluate the domain adaptation, we select seven languages (DE, FR, IT, RO, NL, PT, and ES) that intersect with MuST-C dataset, in which the training size of Europarl-ST is one ninth of MuST-C dataset. For our method, we only leverage the bilingual text in the entire speech-translation data to achieve domain adaptation. The statistics of different language pairs are shown in Appendix A.1.

Europarl Datasets. To verify the effectiveness of our proposed method with the large-scale text translation data, we introduce the easily accessible in-domain parallel corpus – Europarl[1] (Koehn, 2005). In our experiments, we randomly select 2M sentence pairs for each translation direction, except for the EN-RO. We adopt the entire Europarl for EN-RO, which consists of almost 400k samples.

Data Pre-processing. We follow the FAIRSEQ S2T (Wang et al., 2020b) recipes to perform data pre-processing. For the speech data in Europarl-ST and MuST-C, we extract an 80-dimensional log-Mel filter bank as the acoustic feature. The acoustic features are normalized by global channel mean and variance. The SpecAugment approach (Park et al., 2019) is used in all experiments and we remove samples consisting of more than 3k frames. For the external text translation data, we delete the bilingual data in Europarl that intersects with the validation/test sets of the Europarl-ST dataset. We adopt unigram sentencepiece[2] to build 5K and 8K sub-word vocabularies for the transcriptions and the translations, respectively. For the multilingual model, both vocabulary sizes are set to 10K.

Baseline. We compare our proposed approach (NPDA-kNN-ST) with several baseline methods:

- **Cascaded ST System:** Iranzo-Sánchez et al. (2020b) provides the version of cascaded ST system on Europarl-ST, in which the large-scale external MT data is used to build the MT system.
- **E2E-ST-Base:** we leverage the MuST-C dataset to train the E2E-ST model following the training process in FAIRSEQ S2T (Wang et al., 2020b). This model is also used as the pre-trained model for NPDA-kNN-ST.
- **E2E-ST-SP:** we build a domain-specific E2E-ST model on Europarl-ST, and its training process is consistent with E2E-ST-Base.
- **E2E-ST-FT:** we fine-tune E2E-ST-Base with the speech training corpus of Europarl-ST.
- **LNA-D:** Li et al. (2021) integrate Wave2vec 2.0 (Baevski et al., 2020) and mBART (Chipman et al., 2021) into a multilingual E2E-ST model, in which layernorm and attention layers in the decoder are fine-tuned with the MuST-C dataset.
- **kNN-MT:** we directly apply kNN-MT (Khandelwal et al., 2021) for E2E-ST-Base and construct the cached datastore with the in-domain speech-translation data.

Implementation Details. All experiments are implemented based on the FAIRSEQ[3] (Ott et al., 2019) toolkit. For the model structure of all baselines, it consists of two one-dimensional convolutional layers with a downsampling factor of 4, 12 transformer encoder layers, and 6 transformer decoder layers. The additional encoder in our approach includes 12 transformer encoder layers and token embedding, which parameters are initialized randomly. The input embedding size of the transformer layer is 256, the FFN layer dimension is 1024, and the number of self-attention heads is 4. For the multilingual model, the above parameters are set to 512, 2048, and 8, respectively. During training, we deploy the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 2e-3 and 10K warm-up updates to optimize model parameters. Both label smoothing coefficient and dropout ratios are set to 0.1. The batch size is set to 20K tokens, and we accumulate the gradient for every 4 batches. We set patience to 5 to select the best checkpoint on the validation set. The Faiss[4] (Johnson et al., 2021) is used to build the in-domain datastore to carry out fast nearest neighbor search. We utilize the Faiss to learn 8192 cluster centroids for each translation direction, and search 64 clusters for each target token in decoding. During inference, the beam size is set to 5 for all methods. The hyper-parameters (k, λ and T) for kNN retrieval are tuned on the in-domain validation set. More details can be found in Appendix A.2. In our experiments, we report the case-sensitive BLEU score (Papineni et al., 2002) using sacreBLEU[5].

[1] https://www.statmt.org/europarl/

[2] https://github.com/google/sentencepiece

[3] https://github.com/pytorch/fairseq

[4] https://github.com/facebookresearch/faiss

[5] https://github.com/mjpost/sacrebleu, with a configuration of 13a tokenizer, case-sensitiveness, and full punctuation

Model	In-domain		Params. (M)	Target Language							
	ST	MT	Tuned/Total	DE	FR	ES	NL	IT	RO	PT	Avg.
Cascaded ST System	✓	✓	-	22.40	23.40	28.00	/	/	/	/	/
E2E-ST-Base	×	×	0.0/31.1	15.71	16.45	23.49	16.06	14.25	16.95	18.28	17.31
LNA-D	×	×	384.8/793.0	22.50	30.00	32.23	/	21.50	/	28.40	/
E2E-ST-SP	✓	×	31.1/31.1	16.20	24.52	26.00	19.50	18.35	20.62	21.34	20.93
E2E-ST-FT	✓	×	31.1/31.1	21.84	30.97	32.25	23.77	23.36	25.47	26.30	26.28
kNN-MT	✓	×	0.0/31.1	18.29	27.69	28.93	20.70	20.45	22.37	23.08	23.07
NPDA-kNN-ST	×	✓	17.1/48.1	18.76	27.73	29.01	20.79	20.54	23.54	23.54	23.42
NPDA-kNN-ST$^+$	×	✓	17.1/48.1	**23.23**	**35.26**	**33.71**	**27.71**	**33.76**	**28.29**	**28.96**	**30.13**

Table 1: BLEU score [%] of different methods on the Europarl-ST dataset. "Tuned Params." refers to the number of fine-tuned parameters. "NPDA-kNN-ST$^+$" directly uses large-scale Europarl data to build the in-domain datastore, while "NPDA-kNN-ST" leverages the text translation part in the Europarl-ST training data. "In-domain" indicates whether the method uses in-domain ST/MT data.

Model	Extra.	Params. (M)	Target Language								
		Tuned/Total	DE	FR	ES	NL	IT	RO	PT	RU	Avg.
						Bilingual Results					
E2E-ST-Base	×	31.1/31.1	22.57	32.61	27.08	27.46	22.74	21.80	28.07	15.45	24.72
AFS	×	-	22.40	31.60	26.90	24.90	23.00	21.00	26.30	14.70	23.85
kNN-MT	×	0.0/31.1	22.97	33.00	27.99	27.93	**23.55**	22.16	28.80	15.73	25.27
NPDA-kNN-ST	×	17.1/48.1	**23.08**	**33.24**	**28.03**	**28.11**	23.44	**22.22**	**28.83**	**15.82**	**25.35**
						Multilingual Results					
E2E-ST-Base	×	76.3/76.3	24.18	34.98	28.28	28.80	24.62	23.22	31.13	15.88	26.39
LNA-D	✓	53.5/76.3	24.16	34.52	28.30	28.35	24.46	23.29	30.51	15.84	26.18
Adapter Tuning	✓	38.4/76.3	24.63	34.75	28.73	28.80	24.96	23.70	30.96	15.89	26.61
kNN-MT	×	0.0/76.3	25.15	**35.67**	**30.22**	**30.36**	25.83	23.66	**31.67**	17.16	27.47
NPDA-kNN-ST	×	23.7/100.0	**25.21**	35.56	30.05	30.31	**25.91**	**23.90**	31.66	**17.23**	**27.48**

Table 2: BLEU score [%] of different E2E-ST methods on the MuST-C dataset. "AFS" and "Adapter Tuning" represent the methods proposed by Zhang et al. (2020) and Le et al. (2021), respectively. Besides, Le et al. (2021) reproduce the translation performance of "LNA-D" on the MuST-C dataset for fair comparison. "Extra." indicates whether the method uses additional data.

4.2 Main Results

Domain Adaptation on Europarl-ST. We verify the effectiveness of NPDA-kNN-ST for domain adaptation on Europarl-ST. As illustrated in Table 1, we can observe that NPDA-kNN-ST significantly outperforms E2E-ST-Base in all language pairs. When the large-scale Europarl data is used, NPDA-kNN-ST$^+$ even achieves 12.82 BLEU improvements over E2E-ST-Base on average, and gains the best performance in all models. These results demonstrate that our proposed method can make full use of in-domain parallel text to achieve domain adaptation when in-domain speech translation data is inaccessible. Besides, NPDA-kNN-ST obtains comparable translation performance with kNN-MT that leverages the truly in-domain speech-translation data to construct a datastore. It further indicates that our method could generate an effective in-domain datastore with text translation data, which is equivalent to the real speech-translation data. We also compare our method with LNA-D that builds the large multilingual E2E-ST model based on Wave2vec and mBART. In spite of adopting a huge model scale and pre-training techniques, this approach still fails to outperform NPDA-kNN-ST$^+$ due to the domain shift problem. This result shows the necessity of domain adaptation when applying large-scale general E2E-ST models in a certain domain. It also brings an interesting research direction that incorporates our method with LNA-D, and we leave it as future work.

E2E-ST Performance on MuST-C. We further evaluate the effect of unifying text and speech representation with an additional encoder on MuST-C.

Model	**DE**	**FR**	**ES**	**NL**	**IT**	**RO**	**PT**	**Avg.**
BLEU Score(↑)								
NPDA-kNN-ST	**18.76**	**27.73**	**29.01**	**20.79**	**20.54**	**23.54**	**23.54**	**23.42**
- w/o MSE Loss	18.44	26.66	28.10	19.93	19.89	22.20	22.45	22.52
- Optimize Embedding Only	18.50	27.42	28.64	20.44	20.15	22.92	23.09	23.02
Cosine Similarity (↑)								
NPDA-kNN-ST	**0.865**	**0.874**	**0.858**	**0.860**	**0.867**	**0.861**	**0.850**	**0.862**
- w/o MSE Loss	0.827	0.836	0.811	0.817	0.825	0.828	0.809	0.822
- Optimize Embedding Only	0.844	0.857	0.839	0.844	0.849	0.845	0.832	0.844
Squared Euclidean Distance (↓)								
NPDA-kNN-ST	**5.387**	**4.723**	**5.050**	**5.637**	**5.098**	**4.996**	**5.707**	**5.228**
- w/o MSE Loss	6.260	5.566	6.070	6.650	6.040	5.938	6.690	6.173
- Optimize Embedding Only	5.610	4.863	5.400	5.950	5.434	5.266	6.043	5.509

Table 3: BLEU score [%], cosine similarity and squared euclidean distance of our method's variants on the Europarl-ST dataset. All datastores are constructed by Europarl-ST training set."w/o MSE Loss" means that the MSE loss function is removed. "Optimize Embedding Only" means that only the token embedding is introduced to the pre-trained E2E-ST model and fine-tuned.

In this experiment, we compare the translation performance when speech and text translation data are leveraged to construct the datastore respectively, and verify the improvement of combining kNN retrieval with the traditional E2E-ST model at the same time. As illustrated in Table 2, we consider both bilingual and multilingual settings, and compare our method with other baselines, including AFS (Zhang et al., 2020), LNA-D and Adapter Tuning (Le et al., 2021). We can see that, when directly incorporating kNN retrieval into the E2E-ST-Base model, NPDA-kNN-ST yields 0.63 and 1.09 BLEU improvements on average in bilingual and multilingual settings, respectively. These results indicate the benefit of introducing kNN retrieval, even when the E2E-ST models are trained on the same data. In addition, NPDA-kNN-ST achieves similar performance with kNN-MT in both bilingual and multilingual settings, which proves the effectiveness of our proposed method on unifying text and speech representation again.

5 Analysis

Ablation Study. To analyze different modules in our method, we carry out an ablation study on the Europarl-ST dataset, including removing the MSE loss function and only introducing token embedding for unifying text and speech representation. In addition to the BLEU score, we measure the cosine similarity and squared euclidean distances between the synthetic representations generated by our method and ideals generated using ground-truth speech-translation data. As shown in Table 3, even without in-domain speech-translation data, NPDA-kNN-ST can generate the representations that are close enough to the ideals (0.86 on cosine similarity and 5.2 on squared euclidean distances), leading to the efficient in-domain retrieval. Besides, two training losses contribute significantly to the excellent performance of our model. Among that, the MT loss is more important, as optimizing model with MSE loss only could not achieve effective domain adaptation in our experiments. Another observation is that our model can be smaller by only introducing the token embedding and reusing the transformer encoder of the pre-trained E2E-ST model, causing small performance degradation.

The Impact of Datastore Size. As illustrated in Table 1, the datastore constructed by the bigger domain-specific text translation corpus seems to obtain better translation performance when using NPDA-kNN-ST. We evaluate the performance differences caused by different datastore sizes on Europarl-ST. For each translation direction of Europarl-ST, we adopt a ratio range of (0.2, 0.4, 0.6, 0.8, 1.0) to randomly sample from the Europarl corpus to build the datastore of different scales for quick experiments. The detailed results are shown in Figure 2. In general, the translation performance in all directions is positively correlated with the datastore size. More specifically, in the direction of IT and FR, performance is increasing rapidly with the expansion of the datastore, and both exceed 10 BLEU scores. The performance improvement

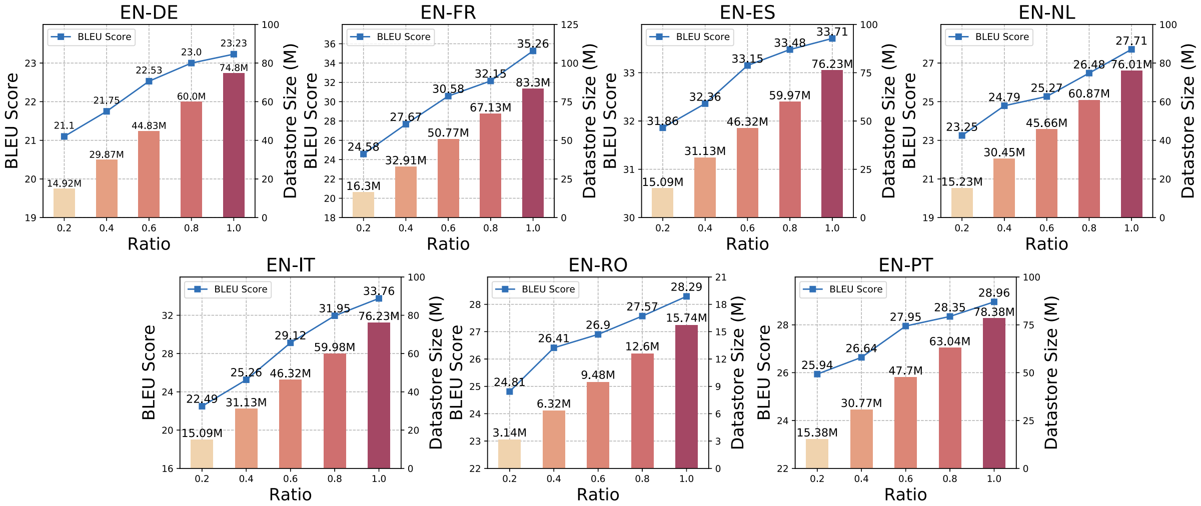


Figure 2: BLEU score [%] of NPDA-kNN-ST with different datastore sizes on the Europarl-ST dataset.

in the DE, ES, NL, and PT directions is relatively smooth. In addition, since the overall datastore size of RO is small, it still shows a reliable performance improvement. Thus, an enormous domain-specific text translation corpus could further improve E2E-ST performance with NPDA-kNN-ST, but brings a larger datastore, which is the trade-off in practice.

6 Related Works

Speech Translation. Early ST methods (Ney, 1999; Sperber et al., 2017; Zhang et al., 2019; Iranzo-Sánchez et al., 2020a; Machácek et al., 2021) cascade the ASR and MT tasks. With the rapid development of deep learning, the neural networks widely used in ASR and MT have been adapted to construct a new end-to-end speech-to-text translation paradigm. However, due to the scarcity of triplet training data, developing an E2E-ST model is still very challenging. Various techniques have been proposed to ease the training process by using source transcriptions, including pre-training (Bansal et al., 2019; Wang et al., 2020c), multi-task learning (Weiss et al., 2017; Anastasopoulos and Chiang, 2018; Sperber et al., 2019), meta-learning (Indurthi et al., 2020), interactive decoding (Liu et al., 2020), consecutive decoding (Dong et al., 2021), and adapter tuning (Le et al., 2021). Different from previous methods, we first investigate the domain adaptation for E2E-ST and propose a non-parametric domain adaptation method to make the E2E-ST system more practical.

Domain Adaptation. The domain adaptation approaches in MT are mainly divided into two categories: 1) model-centric, which focuses on modifying the MT model architecture to learn domain-related information (Wang et al., 2017; Wuebker et al., 2018; Bapna et al., 2019; Guo et al., 2021); 2) data-centric, focusing on utilization of the monolingual corpus (Zhang and Zong, 2016; Zhang et al., 2018), synthetic corpus (Hoang et al., 2018; Hu et al., 2019; Wei et al., 2020), or parallel corpus (Chu et al., 2017) in the specific domain for fine-tuning strategies to improve performance. Recently, non-parametric methods provide a new paradigm for domain adaptation by retrieving the datastore of similar instances (Gu et al., 2018; Bapna and Firat, 2019; Khandelwal et al., 2021; Zheng et al., 2021). We follow this research line and extend this non-parametric method in the domain adaptation scenario for E2E-ST.

7 Conclusion

In this paper, we present a novel non-parametric method that leverages domain-specific bilingual text to achieve domain adaptation for the E2E-ST system. This approach builds the universal output representation for text and speech translation tasks by a carefully designed architecture and loss function. Based on that, a kNN classifier is introduced to improve translation performance with an external datastore constructed by the in-domain text translation data. Experimental results on Europarl-ST demonstrate that our proposed method obtains significant improvement over the pre-trained E2E-ST model when using large-scale in-domain bilingual text corpus. In the future, we would like to explore the combination of our method and large-scale E2E-ST model, such as LNA-D.

References

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. In *NAACL*.

Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*.

Sameer Bansal, H. Kamper, Karen Livescu, Adam Lopez, and S. Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *NAACL*.

Ankur Bapna, N. Arivazhagan, and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *EMNLP*.

Ankur Bapna and Orhan Firat. 2019. Non-parametric adaptation for neural machine translation. In *NAACL*.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *ACL/IJCNLP*.

Hugh A. Chipman, Edward I. George, Robert E. McCulloch, and Thomas S. Shively. 2021. mbart: Multidimensional monotone bart. *Bayesian Analysis*.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *ACL*.

Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021. Consecutive decoding for speech-to-text translation. In *AAAI*.

Mattia Antonino Di Gangi, R. Cattoni, L. Bentivogli, Matteo Negri, and M. Turchi. 2019. Must-c: a multilingual speech translation corpus. In *NAACL*.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alex Smola. 2006. A kernel method for the two-sample-problem. In *NIPS*.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. 2018. Search engine guided neural machine translation. In *AAAI*.

Demi Guo, Alexander M. Rush, and Yoon Kim. 2021. Parameter-efficient transfer learning with diff pruning. In *ACL/IJCNLP*.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *FINDINGS*.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021. Efficient nearest neighbor language models. *ArXiv*, abs/2109.04212.

Cong Duy Vu Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *NMT@ACL*.

Junjie Hu, M. Xia, Graham Neubig, and Jaime G. Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *ACL*.

S. Indurthi, HJ Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. 2020. End-end speech-to-text translation with modality agnostic meta-learning. In *ICASSP*, pages 7904–7908.

Javier Iranzo-Sánchez, Adrià Giménez-Pastor, J. Silvestre-Cerdà, Pau Baquero-Arnal, Jorge Civera Saiz, and Alfons Juan-Císcar. 2020a. Direct segmentation models for streaming speech translation. In *EMNLP*.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020b. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7:535–547.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. *ArXiv*, abs/2010.00710.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting librispeech with French translations: A multimodal corpus for direct speech translation evaluation. In *LREC*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MTSUMMIT*.

Hang Le, J. Pino, Changhan Wang, Jiatao Gu, D. Schwab, and L. Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *ACL/IJCNLP*.

Xian Li, Changhan Wang, Yun Tang, C. Tran, Yuqing Tang, J. Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Multilingual speech translation from efficient finetuning of pretrained models. In *ACL/IJCNLP*.

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation. In *INTERSPEECH*.

Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2020. Synchronous speech recognition and speech-to-text translation with interactive decoding. In *AAAI*.

Dominik Machácek, Matús Zilinec, and Ondrej Bojar. 2021. Lost in interpreting: Speech translation from source or interpreter? *ArXiv*, abs/2106.09343.

H. Ney. 1999. Speech translation: coupling of recognition and translation. *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, 1:517–520 vol.1.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, S. Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*.

Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Daniel S. Park, William Chan, Y. Zhang, C. Chiu, Barret Zoph, E. D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *INTERSPEECH*.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey. *ArXiv*, abs/2006.00632.

Matthias Sperber, Graham Neubig, J. Niehues, and A. Waibel. 2017. Neural lattice-to-sequence models for uncertain inputs. In *EMNLP*.

Matthias Sperber, Graham Neubig, J. Niehues, and A. Waibel. 2019. Attention-passing models for robust and data-efficient end-to-end speech translation. *Transactions of the Association for Computational Linguistics*, 7:313–325.

Laura Cross Vila, Carlos Escolano, José A. R. Fonollosa, and Marta Ruiz Costa-jussà. 2018. End-to-end speech translation with the transformer. In *IberSPEECH*.

Changhan Wang, J. Pino, Anne Wu, and Jiatao Gu. 2020a. Covost: A diverse multilingual speech-to-text translation corpus. In *LREC*.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and J. Pino. 2020b. Fairseq s2t: Fast speech-to-text modeling with fairseq. In *AACL*.

Chengyi Wang, Yu Wu, Shujie Liu, M. Zhou, and Zhenglu Yang. 2020c. Curriculum pre-training for end-to-end speech translation. In *ACL*.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *EMNLP*.

Hao-Ran Wei, Zhirui Zhang, Boxing Chen, and Weihua Luo. 2020. Iterative domain-repaired back-translation. *ArXiv*, abs/2010.02473.

Ron J. Weiss, J. Chorowski, Navdeep Jaitly, Yonghui Wu, and Z. Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *INTERSPEECH*.

Joern Wuebker, Patrick Simianer, and John DeNero. 2018. Compact personalized models for neural machine translation. In *EMNLP*.

Biao Zhang, Ivan Titov, B. Haddow, and Rico Sennrich. 2020. Adaptive feature selection for end-to-end speech translation. In *Findings of EMNLP*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *EMNLP*.

Peidong Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2019. Lattice transformer for speech translation. In *ACL*.

Zhirui Zhang, Shujie Liu, Mu Li, M. Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. *ArXiv*, abs/1803.00353.

Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. 2021. Adaptive nearest neighbor machine translation. In *ACL/IJCNLP*.

A Appendix

A.1 Dataset Statistics

The statistics of datasets used in our experiments are shown in Table 4 and 5.

	Speech Duration	Train Pairs	Dev Pairs	Test Pairs
DE	408 hrs	225,278	1,419	2,588
FR	492 hrs	269,256	1,409	2,579
ES	504 hrs	260,050	1,313	2,450
IT	465 hrs	248,155	1,305	2,521
NL	442 hrs	243,516	1,419	2,563
PT	385 hrs	201,462	1,365	2,449
RO	432 hrs	231,471	1,366	2,503
RU	489 hrs	259,531	1,313	2,460

Table 4: The statistics of all EN-X translation directions in the MuST-C dataset.

	Speech Duration	Train Pairs	Dev Pairs	Test Pairs
DE	83 hrs	32,629	1,321	1,254
FR	81 hrs	31,778	1,282	1,215
ES	81 hrs	31,608	1,273	1,268
IT	80 hrs	29,553	1,123	1,131
NL	80 hrs	31,402	1,270	1,236
PT	81 hrs	31,751	1,295	1,263
RO	72 hrs	28,599	1,071	1,096

Table 5: The statistics of all EN-X translation directions in the Europarl-ST dataset.

A.2 Hyper-Parameter Tuning for kNN-MT and NPDA-kNN-ST

The performance of kNN-MT and NPDA-kNN-ST is highly related to the choice of hyper-parameters. We adopt grid search of $k \in \{4, 8, 16, 32\}$, $\lambda \in \{0.1, 0.2, ..., 0.9\}$ and $T \in \{1, 10, 20, 50, 100, 200\}$ for each translation direction on Europarl-ST validation set when using kNN-MT and NPDA-kNN-ST. The optimal choice is shown in Table 6 and Table 7.

	DE	FR	ES	NL	IT	RO	PT	RU
kNN-MT								
$\boldsymbol{k}$	8	16	8	16	16	16	16	8
$\boldsymbol{\lambda}$	0.2	0.3	0.2	0.2	0.3	0.1	0.2	0.3
$\boldsymbol{T}$	10	20	20	50	10	50	10	20
NPDA-kNN-ST								
$\boldsymbol{k}$	32	16	8	16	8	32	16	8
$\boldsymbol{\lambda}$	0.4	0.3	0.3	0.3	0.3	0.3	0.2	0.3
$\boldsymbol{T}$	10	20	20	50	10	10	20	20

Table 6: Optimal choice of hyper-parameters for each translation direction on MuST-C validation set for E2E-ST experiments.

	DE	FR	ES	NL	IT	RO	PT
kNN-MT							
$\boldsymbol{k}$	16	16	16	16	16	8	8
$\boldsymbol{\lambda}$	0.5	0.7	0.6	0.6	0.7	0.6	0.6
$\boldsymbol{T}$	10	20	10	20	20	50	50
NPDA-kNN-ST							
$\boldsymbol{k}$	16	32	16	16	32	16	32
$\boldsymbol{\lambda}$	0.5	0.7	0.6	0.7	0.7	0.7	0.7
$\boldsymbol{T}$	10	10	10	20	10	10	10
NPDA-kNN-ST$^+$							
$\boldsymbol{k}$	32	4	8	8	4	8	8
$\boldsymbol{\lambda}$	0.8	0.8	0.8	0.8	0.8	0.8	0.8
$\boldsymbol{T}$	10	10	10	10	10	10	10

Table 7: Optimal choice of hyper-parameters for each translation direction on Europarl-ST validation set for domain adaptation experiments.