Revisiting End-to-End Speech-to-Text Translation From Scratch

Biao Zhang [1]   **Barry Haddow** [1]   **Rico Sennrich** [2 1]

Abstract

End-to-end (E2E) speech-to-text translation (ST) often depends on pretraining its encoder and/or decoder using *source transcripts* via speech recognition or text translation tasks, without which translation performance drops substantially. However, transcripts are not always available, and how significant such pretraining is for E2E ST has rarely been studied in the literature. In this paper, we revisit this question and explore the extent to which the quality of E2E ST trained on speech-translation pairs alone can be improved. We reexamine several techniques proven beneficial to ST previously, and offer a set of best practices that biases a Transformer-based E2E ST system toward training from scratch. Besides, we propose parameterized distance penalty to facilitate the modeling of locality in the self-attention model for speech. On four benchmarks covering 23 languages, our experiments show that, without using any transcripts or pretraining, the proposed system reaches and even outperforms previous studies adopting pretraining, although the gap remains in (extremely) low-resource settings. Finally, we discuss neural acoustic feature modeling, where a neural model is designed to extract acoustic features from raw speech signals directly, with the goal to simplify inductive biases and add freedom to the model in describing speech. For the first time, we demonstrate its feasibility and show encouraging results on ST tasks.[1]

1. Introduction

End-to-end (E2E) speech-to-text translation (ST) is the task of translating a source-language audio directly to a foreign text without any intermediate outputs (Duong et al., 2016; Bérard et al., 2016), which has gained increasing popularity and obtained great success recently (Sung et al., 2019; Salesky et al., 2019; Zhang et al., 2020; Chen et al., 2020; Han et al., 2021; Zheng et al., 2021; Anastasopoulos et al., 2021). Different from the traditional cascading method which decomposes ST into two sub-tasks – automatic speech recognition (ASR) for transcription and machine translation (MT) for translation, E2E ST jointly handles them in a single, large neural network. This endows E2E ST with special advantages on reducing translation latency and bypassing transcription mistakes made by ASR models, making it theoretically attractive.

However, directly modeling speech-to-text mapping is nontrivial. The translation alignment between speech and text is no longer subject to the monotonic assumption. Also, the high variation of speech increases the modeling difficulty. Therefore, rather than training E2E ST models from scratch, researchers often resort to pipeline-based training with auxiliary tasks utilizing source transcripts, which first pretrains the speech encoder on ASR data and/or the text decoder on MT data followed by a finetuning on ST data. Such pretraining was reported to greatly improve translation quality (Di Gangi et al., 2019; Wang et al., 2019a; Zhang et al., 2020; Xu et al., 2021), and has become the *de-facto* standard in recent ST studies and toolkits (Inaguma et al., 2020; Wang et al., 2020a; Zhao et al., 2021; Zheng et al., 2021). Despite these successes, nevertheless, how significant the pretraining is for E2E ST and how far we can go using speech-translation pairs alone are still open questions.

In this paper, we aim at exploring the extent to which the quality of ST models trained from scratch can be improved, whether the performance gap against pretraining-based ST can be narrowed, and also when the pretraining really matters.[2][3] We argue that the inferior performance of ST from scratch is mainly a result of the dominance of pretrain-

---

[1]School of Informatics, University of Edinburgh [2]Department of Computational Linguistics, University of Zurich. Correspondence to: Biao Zhang <b.zhang@ed.ac.uk>.

[1]Source code is available at https://github.com/bzhangGo/zero.

---

[2]Note there are two types of pretraining for E2E ST in general: 1) pretraining with **triplet** (ASR/MT) data alone, and 2) pretraining with **external** unlabeled or ASR/MT data. In this study, we refer *pretraining* mainly to the former case, although we also compare our work to systems pretrained on unlabeled data.

[3]By *ST from scratch*, we refer to the setup where ST models are trained on speech-translation pairs alone without using transcripts or any type of pretraining.

ing, and consequent lack of focus on optimizing E2E ST models trained from scratch. To test this hypothesis, we investigate methods to bias a Transformer-based E2E ST model (Vaswani et al., 2017) towards training from scratch. We summarize a set of best practices for our setup by revisiting several existing techniques that have been proven useful to ST previously. We further introduce two proposals to add freedom to Transformer to model speech with the hope of gaining translation quality: 1) a parameterized distance penalty that facilitates self-attention to capture local dependencies of speech; and 2) neural acoustic feature modeling providing a trainable alternative to the heuristic rule-based acoustic feature extraction.

To examine the generality of our methods, we conducted (bilingual) experiments on four speech translation benchmarks, including MuST-C, Covost2, LibriSpeech, and Kosp2e, which cover 23 languages of different families with varying training data sizes. Experimental results show that the significance of pretraining has been over-estimated in prior work, and integrating techniques to improve E2E ST from scratch is feasible and promising. Our main findings:

- With proper adaptation, E2E ST trained from scratch only on speech-translation pairs can match or even surpass previous studies using ASR/MT pretraining on source transcripts.

- Pretraining still matters, mainly in (extremely) low-resource regimes and when large-scale external ASR or MT corpora are available.

- We present a set of best practices for E2E ST from scratch, including smaller vocabulary size, wider feed-forward layer, deep speech encoder with the post-LN (layer normalization) structure, Connectionist Temporal Classification (CTC)-based regularization using translation as the target, and a novel parameterized distance penalty.

- We demonstrate that dropping heuristic rule-based acoustic features is feasible, and that neural acoustic features can be learned in an end-to-end ST framework.

2. Why Revisiting ST *From Scratch*?

In our view, there are several reasons making E2E ST from scratch intriguing.

First of all, our study does not preclude pretraining (or more generally, multi-task learning) for ST. We believe that leveraging knowledge from auxiliary tasks via pretraining to improve ST is a remarkable research direction. But rather, our study contributes to a better understanding of the genuine role of pretraining in E2E ST. Re-assessing the importance of pretraining is a useful signal to inform future research projects and practical deployments of ST.
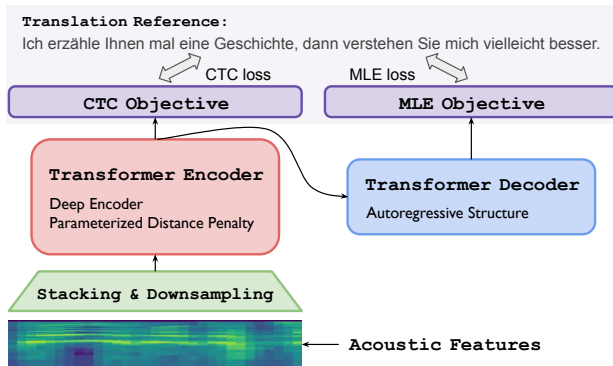

Figure 1: Overview of the proposed ST system. The example is for En-De translation. During inference, the CTC layer is dropped and only the autoregressive decoder is used.

Secondly, focusing on ST from scratch has an even higher relevance in settings where ASR/MT data is scarce. By only requiring speech-translation training pairs, ST from scratch reduces data requirements and the associated costs. This is especially important for the estimated 3000 languages in the world that have no written form at all, for which it would be impractical to collect large amounts of phonetically transcribed data.

Thirdly, removing pretraining eases model analysis and simplifies the training pipeline, which also offers a testbed to identify inductive biases that support ST with better data efficiency. Pretraining often takes extra training time and computing resources. As pretraining itself affects the final results, it becomes more difficult to figure out the source of the improved performance when new algorithms or architectures are incorporated. In contrast, ST from scratch simplifies model development, and lets us efficiently re-examine recently proposed techniques for ST, and explore novel techniques. This allows us to build strong models for future research to build on or compare to.

3. Methods for ST From Scratch

We argue that the inferior performance of ST from scratch as reported in the literature is due to a lack of system adaptation with respect to training and modeling. In this section, starting with a brief overview of our baseline system, we discuss several potential directions that could strengthen E2E ST trained from scratch. The overall framework of the proposed system is shown in Figure 1.

3.1. Baseline

Our baseline follows the encoder-decoder paradigm (Bahdanau et al., 2015) and uses Transformer (Vaswani et al., 2017) as its backbone. Except for (speech, translation) pairs denoted as (X, Y), respectively, we assume that there is no

access to other data at training for ST from scratch.

The encoder stacks N_{enc} identical layers, each of which has a multi-head self-attention sublayer and a feed-forward sublayer. To enhance its short-range dependence modeling, we apply the logarithmic distance penalty (Di Gangi et al., 2019) to each head of its self-attention:

$$\text{Head}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_{head}}} - \pi(\mathbf{D})\right)\mathbf{V}, \quad (1)$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{|X| \times d_{head}}$ are the query, key and value inputs, respectively. d_{head} is the attention head dimension. $|\cdot|$ denotes sequence length. $\mathbf{D} \in \mathbb{R}^{|X| \times |X|}$ stores the position distance, i.e. $D_{i,j} = |i - j| + 1$, and $\pi(\cdot) = \log(\cdot)$.

Analogous to the encoder, the decoder stacks N_{dec} identical layers. We reuse the standard Transformer decoder for our baseline, and optimize all model parameters using the traditional maximum likelihood objective (MLE), or $\mathcal{L}^{\text{MLE}}$.

3.2. Hyperparameter Tuning

Hyperparameters often highly affect ST from scratch, but exhaustively searching for optimal settings is impractical. Instead, we take inspiration from past studies and re-examine several configurations that have been proven beneficial to ST with pretraining. We hypothesize that such configurations also have a high chance to generalize to ST from scratch. For example, since ST is generally a low-resource task, using smaller vocabulary (Inaguma et al., 2020), larger dropout rate (Sennrich & Zhang, 2019), reduced attention heads and model dimension (Inaguma et al., 2020; Zhao et al., 2021) might help to avoid overfitting. We also test different settings for acoustic feature extraction, deep encoder (Zhang et al., 2019) and wide feed-forward layer (Inaguma et al., 2020), apart from tuning the length penalty at inference (Wu et al., 2016).

3.3. CTC-based Regularization

CTC, or Connectionist Temporal Classification, is a latent alignment objective that models probabilistic distribution by marginalizing over all valid mappings between the input and output sequence (Graves et al., 2006). Under a strong conditional independence assumption, it can be computed efficiently and tractably via dynamic programming. We refer readers to Graves et al. (2006) for more algorithmic details. CTC with source transcripts as output labels has been found to be an effective auxiliary task for E2E ST (Bahar et al., 2019). Another application of CTC is to use translations as the output labels, thus modelling the translation task. So far, this idea has been applied to non-autoregressive MT and ST successfully (Libovický & Helcl, 2018; Chuang et al., 2021).

In this paper, we regard CTC with translations as output labels as a regularizer and stack it onto the encoder for *ST modeling* as shown in Figure 1. The overall training objective becomes as below:

$$\mathcal{L}(X, Y) = (1 - \lambda)\mathcal{L}^{\text{MLE}}(Y|X) + \lambda\mathcal{L}^{\text{CTC}}(Y|X), \quad (2)$$

where λ is a hyperparameter controlling the degree of the regularization. Chuang et al. (2021) showed that CTC improves the reordering tendency of the self-attention in non-autoregressive ST, although it assumes monotonicity. We expect that such reordering could reduce the learning difficulty of ST and ease the decoder's job, delivering better translation quality. One problem of applying CTC to ST is that the input speech sequence might be shorter than its translation sequence, which violates CTC's presumption. We simply ignore these samples during training. Note that the CTC layer will be abandoned after training.

3.4. Parameterized Distance Penalty

The distance penalty in Eq. 1 penalizes attention logits logarithmically with distance based on a *hard-coded* function, reaching a certain degree of balance in modeling local and global dependencies. However, such a function lacks flexibility and inevitably suffers from insufficient capacity when characterizing data-specific locality. To solve this problem, we propose parameterized distance penalty (PDP) which includes a learnable parameter for each distance. PDP is inspired by the relative position representation (Shaw et al., 2018; Raffel et al., 2020) and is formulated as below:

$$\pi^{\text{PDP}}(\mathbf{D}) = \log(\mathbf{D})f(\mathbf{D}), \quad (3)$$

$$f(D_{i,j}) = \begin{cases} \mathbf{w}_{D_{i,j}}, & \text{if } D_{i,j} < R \\ \mathbf{w}_R, & \text{otherwise} \end{cases} \quad (4)$$

where $\mathbf{w} \in \mathbb{R}^R$ is a trainable vector, R is a hyperparameter, and $\mathbf{w}_i$ denotes its i-th element. PDP is easily parallelizable, adding little computational overhead. We initialize each $\mathbf{w}_i$ to 1 so that PDP starts from $\pi(\cdot)$ and then gradually adjusts itself during training. Besides, $\mathbf{w}$ is attention head-specific, i.e. each head has its own parameterization. By doing so, we enable different heads capturing varying degree of locality, which further increases modeling freedom.

4. Experimental Setup

Dataset We work on four benchmarks covering different domains and 23 languages from diverse language families.

MuST-C MuST-C is extracted from TED talks (Di Gangi et al., 2019), offering translations from English (En) to 8 languages: German (De), Spanish (Es), French (Fr), Italian (It), Dutch (Nl), Portuguese (Pt), Romanian (Ro) and Russian (Ru). The training sets of each language

Table 1: Ablation results on MuST-C En-De test set. *#Params*: number of model parameters. *BLEU*: higher is better, SacreBLEU. Numbers in bold denote top scores.

ID	System	#Params	BLEU↑
1	Baseline	51M	18.1
Tune beam search, dropout and batch size			
2	1 + adjust length penalty at inference	51M	18.8
3	1 + higher dropout (0.2→0.4)	51M	17.4
4	1 + apply dropout to raw waveform signals (rate 0.1)	51M	14.6
5	1 + reduce batch size by half	51M	17.6
Tune model dimension and depth			
6	2 + reduce model dimension and attention heads ($H : 8 \rightarrow 4, d_{model} : 512 \rightarrow 256$)	20M	19.0
7	6 + enlarge feed-forward layer ($d_{ff} : 2048 \rightarrow 4096$)	33M	19.3
8	6 + enlarge encoder depth with DS-Init ($N_{enc} : 6 \rightarrow 12$)	28M	20.4
9	8 + enlarge feed-forward layer ($N_{enc} = 12, d_{ff} : 2048 \rightarrow 4096$)	47M	21.1
10	2 + enlarge encoder depth with DS-Init ($N_{enc} : 6 \rightarrow 12$)	70M	20.3
Add parameterized distance penalty (PDP)			
11	2 + PDP ($R = 512$)	51M	19.5
12	11 + initialize **w** in PDP randomly	51M	18.3
13	11 + use 80-dimensional log mel-scale filterbank ($F : 40 \rightarrow 80$)	51M	19.3
14	11 + remove delta and delta-delta features ($d_{speech} : 120 \rightarrow 40$)	50M	18.8
Tune vocabulary size and LN			
15	9 + PDP	47M	21.8
16	15 + small BPE vocabulary ($V : 16K \rightarrow 8K$)	46M	21.8
17	16 + change post-LN to pre-LN	46M	20.6
Final system: add CTC			
18	16 + CTC regularization ($\lambda = 0.3$) (also, the proposed system)	48M	**22.7**
Compare to ST with ASR pretraining			
19	for comparison: 16 + ASR pretraining	46M	**22.9**
20	for comparison: 1 + ASR pretraining	51M	20.7

are at a similar scale, roughly 452 hours with 252K utterances on average.

LibriSpeech En-Fr The Augmented LibriSpeech dataset is collected by aligning e-books in French with English utterances of LibriSpeech (Kocabiyikoglu et al., 2018). We only use the 100 hours clean training set and its augmented references offered by Google Translate for training, totalling 94K utterances.

Kosp2e Ko-En Kosp2e is constructed from a mix of four domains (textbook, news, AI agent and diary) for Korean-to-English (Ko-En) speech translation (Cho et al., 2021). The training set has about 190 hours with 106K utterances.

CoVoST CoVoST (version 2) is a large-scale multilingual ST corpus collected from Common Voice (Ardila et al., 2020), providing translations from En to 15 languages – Arabic (Ar), Catalan (Ca), Welsh (Cy), De, Estonian (Et), Persian (Fa), Indonesian (Id), Japanese (Ja), Latvian (Lv), Mongolian (Mn), Slovenian (Sl), Swedish (Sv), Tamil (Ta), Turkish (Tr), Chinese (Zh) – and from 21 languages to En, including the 15 target languages

as well as Es, Fr, It, Nl, Pt and Ru (Wang et al., 2020b). The training set for En→Xx translation is of similar scale, roughly 427 hours with 289K utterances. In contrast, the training data size for Xx→En translation varies greatly, from about 1.2 hours/1.2K utterances (Id) to 263 hours/207K utterances (Fr). We mainly work on Fr, De, Es, Ca, It, Ru, and Zh for Xx→En.

For each benchmark, we use the official train/dev/test split for experiments. We convert all audios to a sampling rate of 16KHz and truncate segments to 3000 frames. We extract 40-dimensional log mel-scale filterbank features ($F = 40$) with a step size of 10ms and window size of 25ms, which are then expanded with their delta and delta-delta features followed by mean subtraction and variance normalization, resulting in the final 120-dimensional acoustic features ($d_{speech} = 120$). We tokenize and truecase all texts via Moses (Zh and Ja excluded) (Koehn et al., 2007), and handle infrequent words via subwords (Sennrich et al., 2016; Kudo & Richardson, 2018) with a vocabulary size of 16K ($V = 16K$).

Model Setting On top of the acoustic input, we concatenate three consecutive frames without overlapping as a way

of downsampling (Zhang et al., 2020), as in Figure 1. We then add a linear layer to get the encoder input of dimension d_{model}. We use the sinusoidal encoding to distinguish different positions, and employ the post-LN (layer normalization) structure for Transformer (Vaswani et al., 2017).

Regarding Baseline, we set $d_{model} = 512$, $d_{head} = 64$, the number of attention head $H = 8$, the feed-forward layer size $d_{ff} = 2048$ and $N_{enc} = N_{dec} = 6$. Note $d_{model} = H \cdot d_{head}$. By default, we set $R = 512$ and $\lambda = 0.3$.

We employ Adam (Kingma & Ba, 2015, $\beta_1 = 0.9, \beta_2 = 0.98$) for parameter update using adaptive learning rate schedule as in (Vaswani et al., 2017) with a warmup step of 4K and label smoothing of 0.1. Dropout of rate 0.2 is applied to residual connections and ReLU activations. We organize training samples of around 20K target subwords into one batch, and train models up to 50K steps.

Evaluation We average the best 10 checkpoints according to dev set performance for evaluation. For decoding, we adopt beam search, where we set the beam size and length penalty to 8 and 0.6, respectively. We will examine the impact of the length penalty on translation later. Unless otherwise stated, we measure translation quality with detokenized case-sensitive BLEU (Papineni et al., 2002) offered by SacreBLEU (Post, 2018).[4] Note that we did not perform any filtering to the test set at evaluation time.

5. Results and Analysis

We test different hyperparameters and our proposals mainly on MuST-C En-De. Table 1 summarizes the results.

Apart from architecture, length penalty in beam search also matters. Length penalty is used to bias beam search generating longer or shorter outputs, which often largely affects translation quality as shown in Figure 2.[5] Tuning this setting alone results in +0.7 BLEU gains ($1 \rightarrow 2$).

Applying more dropout and smaller batch size helps little. Dropout is a popular regularizer to avoid overfitting. We tried using larger dropout rate and adding dropout to raw waveforms, but ended up with significantly slower convergence and worse performance ($1 \rightarrow 3, 4$). Also, reducing training batch deteriorates ST ($1 \rightarrow 5$).

Deepening speech encoder, widening feed-forward layer, and reducing model dimension benefit ST. Halving model dimension greatly reduces the number of model parameters but still retains translation quality ($2 \rightarrow 6$). Enlarging the encoder depth (from 6 to 12) and the feed-forward

---

[4]Signature: *BLEU+c.md+#ref.1+s.exp+tok.13a+v.1.4.14*

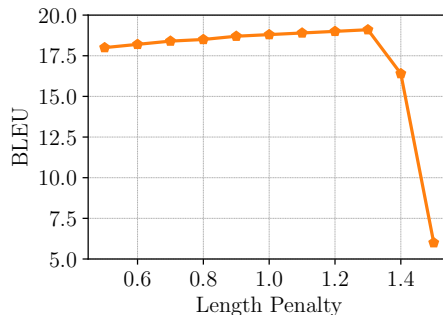[5]Note that its impact is dataset-dependent. On CoVoST, BLEU changes little when varying it.

Figure 2: Dev SacreBLEU scores as a function of length penalty ($0.5 \rightarrow 1.5$) for Baseline on MuST-C En-De. Trade-off exists.
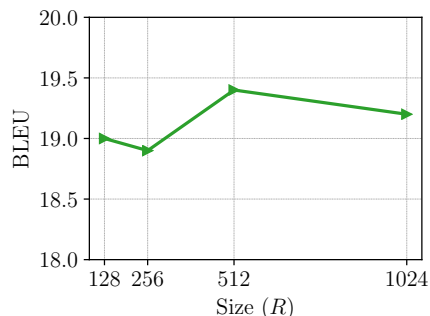

Figure 3: Dev SacreBLEU on MuST-C En-De when changing R in PDP for system 11. Setting $R = 512$ yields the best result.

dimension (from 2048 to 4096) leads to substantial quality improvement, +2.1 BLEU ($6 \rightarrow 9$). After varying dimensions, we could achieve a BLEU score of **21.1**. Note, we employed the depth-scaled initialization to smooth model gradients for deep Transformer (Zhang et al., 2019, DS-Init) and set $\alpha = 0.5$. Besides, deep speech encoder improves ST from scratch with the Baseline dimensions ($2 \rightarrow 10$).

The proposed parameterized distance penalty improves ST. The hyperparameter R in Eq. 3 affects the flexibility of PDP in modeling local context. Figure 3 shows its impact on ST. In general, setting $R = 512$ achieves good performance. Note, its optimal setting might (and is likely to) be dataset-dependent.

Applying PDP to ST gains BLEU ($2 \rightarrow 11$) and is complementary to model dimension manipulation ($9 \rightarrow 15$), reaching a test BLEU score of 21.8. We also tested the effectiveness of initializing all $\mathbf{w}_i$ to 1. Using the vanilla random initialization instead delivers inferior quality, -1.2 BLEU ($11 \rightarrow 12$).

Inadequate acoustic feature extraction hurts ST. In previous ST systems (Inaguma et al., 2020; Zhao et al., 2021), acoustic feature extraction often uses 80-dimensional filterbanks without delta and delta-delta features. We checked
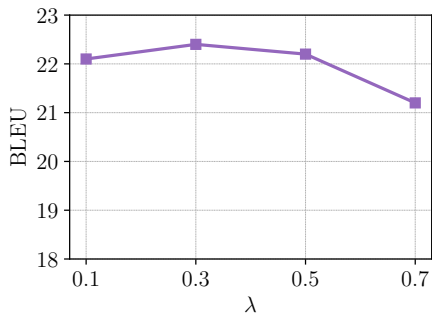
Figure 4: Dev SacreBLEU as a function of λ on MuST-C En-De for system 18. We set $\lambda = 0.3$ in our experiments.
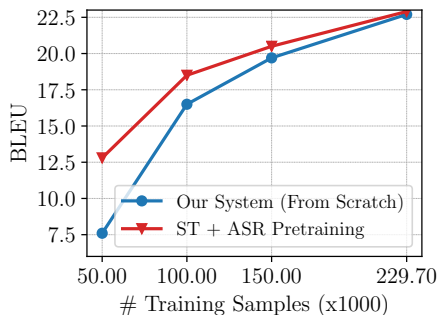

Figure 5: Impact of the amount of training data on MuST-C En-De translation. Results are for test SacreBLEU.

this in our setup. Using more filterbanks does not help much ($11 \rightarrow 13$), and delta features benefit ST a lot ($11 \rightarrow 14$).

Reducing vocabulary size affects En-De translation little. Previous studies also suggest to use smaller vocabularies in low-resource settings (Karita et al., 2019; Sennrich & Zhang, 2019). Reducing vocabulary size by half yields little impact on En-De translation ($15 \rightarrow 16$). We adopt smaller vocabularies due to three reasons: 1) it reduces the number of parameters; 2) we observed that it has much greater influence on other languages; and 3) CTC with smaller vocabulary is more computationally efficient.

Post-LN vs. Pre-LN Another way to train deep Transformer is to use the pre-LN structure (Wang et al., 2019b). It has been shown that the post-LN, once successfully optimized, often outperforms its pre-LN counterpart (Zhang et al., 2019). We reconfirmed this observation, and found that the post-LN ST with DS-Init shows clear superiority in performance, +1.2 BLEU ($17 \rightarrow 16$).

CTC greatly improves ST from scratch. Finally, we integrate the CTC regularization into our best system. The hyperparameter λ in Eq. 2 controls the trade-off between two different objectives. Figure 4 shows that λ directly

affects ST, and setting $\lambda = 0.3$ achieves the best result. Under this setting, CTC benefits ST with another significant quality gain, +0.9 BLEU, reaching a test BLEU of **22.7** (18).

Large combined effect, with and without pretraining From Baseline to system 18, we improve ST by 4.6 BLEU. Note that this system also outperforms the baseline with ASR pretraining (20). Comparing systems 19 and 20, we can also see that our proposals benefit models with pretraining, although the improvement (2.2 BLEU) is smaller than for models trained from scratch. Consequently, the gap between our best system trained from scratch and its pretrained counterpart has become very narrow (18 vs. 19).[6]

For all follow-up experiments, we focus on models trained from scratch, and use system 18 as our proposed system.

Pretraining matters in low-resource regime. Pretraining might not be crucial when rich training data is given, but it matters as the amount of training data decreases. Figure 5 demonstrates this. ASR pretraining helps low-resource ST.

Results On Other Languages Putting all together, we obtain a set of best practices, involving $N_{enc} = 12, N_{dec} = 6, d_{model} = 256, H = 4, d_{ff} = 4096, V = 8K$, using PDP with $R = 512$ and applying CTC with $\lambda = 0.3$. We then keep this configuration and train models for other language pairs. Tables 2-5 list the results.

Our revisiting of ST from scratch shows that its performance gap to ST with pretraining has generally been overestimated in the literature. This gap can be largely reduced and even fully closed after biasing E2E ST towards training from scratch. Our system achieves an average BLEU of 25.1 and 17.3 on MuST-C and CoVoST En→Xx, respectively, which surpasses many popular neural systems, such as the ones supported by Fairseq (Wang et al., 2020a) and NeurST (Zhao et al., 2021). Similarly, our system achieves very promising performance on LibriSpeech En-Fr and Kosp2e Ko-En, delivering 18.9 and 5.8 BLEU, respectively. Note Cho et al. (2021) employed extra large-scale ASR data for pretraining, which is merely 0.1 BLEU higher than ours. While this is beyond the scope of our work, our results suggest that it is worthwhile to revisit large-scale pretraining based on our stronger baseline, which will lead to either new state-of-the-art results or a re-evaluation of the effectiveness of large-scale pretraining.

Our results also show that pretraining matters mainly in two aspects: 1) low-resource scenarios, where our system

---

[6]Note for ASR pretraining, we also adopt the CTC loss based on transcripts to regularize the encoder apart from the decoder-side transcript-based MLE loss, while ST finetuning is based on the translation-based MLE loss alone, following (Inaguma et al., 2020; Zhang et al., 2020).

Table 2: Results of different systems on MuST-C tst-COMMON. *Avg*: average score over different languages. [†]: systems that might perform filtering to the test set, so comparison could be unfair. [‡]: systems using large-scale **external** ASR and/or MT data.

System	Aux. Data		De	Es	Fr	It	Nl	Pt	Ro	Ru	Avg
	ASR	MT									
Adapted Transformer (Di Gangi et al., 2019)	✓		17.3	20.8	26.9	16.8	18.8	20.1	16.5	10.5	18.5
ESPnet-ST (Inaguma et al., 2020)[†]	✓	✓	22.9	28.0	32.8	23.8	27.4	28.0	21.9	15.8	**25.1**
AFS (Zhang et al., 2020)	✓		22.4	26.9	31.6	23.0	24.9	26.3	21.0	14.7	23.9
Contextual Modeling (Zhang et al., 2021)	✓		22.9	27.3	32.5	23.1	26.0	27.1	23.6	15.8	24.8
Fairseq-ST (Wang et al., 2020a)[†]	✓		22.7	27.2	32.9	22.7	27.3	28.1	21.9	15.3	24.8
NeurST (Zhao et al., 2021)	✓		22.8	27.4	33.3	22.9	27.2	28.7	22.2	15.1	24.9
E2E-ST-JT (Du et al., 2021)[†]	✓		23.1	27.5	32.8	23.6	27.8	28.7	22.1	14.9	**25.1**
Chimera (Han et al., 2021)[‡]	✓	✓	27.1	30.6	35.6	25.0	29.2	30.2	24.0	17.4	**27.4**
our system			22.7	28.1	33.4	23.2	26.9	28.3	22.6	15.4	**25.1**
our system + neural acoustic feature modeling			23.0	28.0	33.5	23.5	27.1	28.2	23.0	15.6	**25.2**

Table 3: Results of different systems for En→Xx and Xx→En on CoVoST. We report character-level BLEU for Chinese and Japanese following Wang et al. (2020b). Languages underlined have training data fewer than 100K samples.

System	Aux. Data		Xx→En							
	ASR	MT	Fr	De	Es	Ca	It	Ru	Zh	Avg
ST from scratch (Wang et al., 2020b)			24.3	8.4	12.0	14.4	0.2	1.2	1.4	8.8
ST + ASR Pretraining (Wang et al., 2020b)	✓		26.3	17.1	23.0	18.8	11.3	14.8	5.8	**16.7**
our system			26.9	14.1	15.7	17.2	2.4	3.6	2.0	11.7

						En→Xx										
Ar	Ca	Cy	De	Et	Fa	Id	Ja	Lv	Mn	Sl	Sv	Ta	Tr	Zh	Avg
8.7	20.2	22.2	13.6	11.1	11.5	18.9	26.9	11.5	6.6	11.5	20.1	9.9	8.9	20.6	14.8
12.1	21.8	23.9	16.3	13.2	13.1	20.4	29.6	13.0	9.2	16.0	21.8	10.9	10.0	25.4	17.1
12.3	22.9	24.5	17.5	13.6	12.7	21.4	28.8	13.6	9.9	15.2	22.9	10.8	10.3	23.3	**17.3**

Table 4: Results of different systems on LibriSpeech En-Fr test set. For comparison to previous work, we report both case-insensitive tokenized BLEU (tok) and SacreBLEU.

System	Aux. Data		BLEU	
	ASR	MT	tok	Sacre
ST + KD (Liu et al., 2019)		✓	17.02	
TCEN (Wang et al., 2019a)	✓	✓	17.05	
AFS (Zhang et al., 2020)	✓		18.56	
LUT (Dong et al., 2021)	✓	✓	18.34	
Chimera (Han et al., 2021)[‡]	✓	✓		*19.4*
our system			**18.90**	16.5

Table 5: Results of different systems on Kosp2e Ko-En test set.

System	Aux. Data		BLEU
	ASR	MT	
ST from scratch (Cho et al., 2021)			2.6
ST + pretraining (Cho et al., 2021)[‡]	✓		**5.9**
our system			5.8

are not considered here due to resource limitations. In addition, we did not aggressively optimize our system towards very low-resource scenarios, so there should still be room for quality improvement on CoVoSt Xx→En. Also note that comparison to ST models powered by ESPnet (Inaguma et al., 2020) and Fairseq (Wang et al., 2020a) might not be fair because both toolkits perform data filtering to the test set, although SacreBLEU is also used.

6. Neural Acoustic Feature Modeling

A general trend in deep learning is to replace handcrafted features with neural networks to let the model automati-

still lags far behind pretraining-enhanced ST, -5.0 BLEU on CoVoST Xx→En in Table 3; and 2) large-scale external ASR and/or MT data is available, where pretraining or joint modeling can largely improve ST, +2.3 BLEU on MuST-C in Table 2 yielded by Chimera (Han et al., 2021).

Notice that our system should be regarded as a lower-bound for ST from scratch, since many outstanding optimization techniques for E2E ST, e.g. SpecAugment (Park et al., 2019),

Table 6: Results of applying NAFM to ST on MuST-C En-De.

System	# Params	BLEU
our system	48M	22.7
our system + NAFM	54M	**23.0**
our system + two FFN blocks alone	54M	22.7

cally capture or learn the underlying pattern behind data. In E2E ST, one heuristic is the adoption of log mel-scale filter-banks for acoustic modeling. Despite its success, filterbank-based modeling prevents us from accessing full acoustic details and its transformation might suffer from information loss (Lam et al., 2021), making it sub-optimal for ST. Inspired by recent speech studies on modeling raw wave-forms (Lam et al., 2021), we propose neural acoustic feature modeling (NAFM) to remove such heuristic and increase the freedom of E2E ST in describing speech.

The extraction of filterbanks often involves a sequence of two specifically designed linear transformations. To simulate such structure, we employ two feed-forward neural blocks for NAFM as follows:

$$\mathbf{x}^{(1)} = \mathrm{LN}\left(\mathrm{FFN}\left(\mathbf{x}^{(0)}\right) + \mathbf{x}^{(0)}\right), \qquad (5)$$

$$\mathbf{x}^{(2)} = \mathrm{LN}\left(\mathrm{FFN}\left(\mathbf{x}^{(1)}\right) + \mathbf{x}^{(1)}\right), \qquad (6)$$

where $\mathbf{x}^{(0)} \in \mathbb{R}^{d_{speech}}$ is the raw speech frame, and FFN$(\cdot)$ is the feed-forward layer as in Transformer (Vaswani et al., 2017) with $d_{ff} = 4096$. We expect that, by adding trainable parameters tuned with translation losses, NAFM could induce ST-oriented acoustic features that improves ST.

However, directly using $\mathbf{x}^{(2)}$ as an alternative to the filter-bank features $\mathbf{x}^f$ results in poor convergence. We argue that filterbanks offer helpful inductive biases to ST, and propose to leverage such information to regularize NAFM. Formally, we add the following L_2 objective into training:

$$\mathcal{L}^{\mathrm{NAFM}}(X,Y) = \mathcal{L}(X,Y) + \gamma \frac{1}{|X|}(\|\mathbf{X}^{(2)} - \mathbf{X}^f\|^2), \quad (7)$$

where γ is a hyperparameter and set to 0.05 in experiments.

Results in Table 6 show that training E2E ST from scratch on raw waveforms is feasible. NAFM improves ST by 0.3 BLEU on MuST-C En-De, and such improvement is not a trivial result of simply adding parameters. The last row of Table 2 shows the effectiveness of NAFM on other languages. Overall, the performance of NAFM matches and even outperforms its filterbank-based counterpart across different languages. Although NAFM does not deliver significant gains, we believe that optimizing ST with raw waveforms has great potential and deserves more effort.

7. Related Work

Methods to improve E2E ST are many. Apart from developing novel model architectures (Di Gangi et al., 2019; Karita et al., 2019; Zhang et al., 2020), one promising way is to leverage knowledge transfer from auxiliary tasks. Multilingual or cross-lingual ST improves translation by adding translation supervisions from other languages (Inaguma et al., 2019; Bansal et al., 2019; Liu et al., 2019). Multi-task learning benefits ST by jointly modeling ASR and ST tasks within a single model (Anastasopoulos & Chiang, 2018; Zheng et al., 2021; Dong et al., 2021; Indurthi et al., 2021). Pretraining methods, including large-scale self-supervised pretraining (Schneider et al., 2019) and ASR/MT-based supervised pretraining, offer a warm-up initialization for E2E ST to improve its data efficiency (Le et al., 2021; Salesky et al., 2019; Xu et al., 2021). However, all these studies assume that (bilingual) ST from scratch is poor, while spending little effort on optimizing it. We challenge this assumption and demonstrate that ST from scratch can also yield decent performance after optimization.

We adopt the CTC objective as a regularizer to improve E2E ST. CTC was proposed for ASR tasks to handle the latent alignment between speech and transcript (Graves et al., 2006), which has been widely used to train ASR models. Based on source transcripts, CTC also improves autoregressive E2E ST via ASR pretraining (Zhao et al., 2021), encoder representation compression using the learned latent alignment (Liu et al., 2020; Gaido et al., 2021), and encoder regularization (Bahar et al., 2019). Particularly, Bahar et al. (2019) applied CTC in a similar way to ours but focused on a multi-task setup where source transcripts are used as CTC labels. Instead, we explore target translations as CTC labels. Besides, CTC contributes to non-autoregressive translation. Libovický & Helcl (2018) and Saharia et al. (2020) applied the CTC loss to non-autoregressive MT and obtained improved translation performance. Gu & Kong (2021) observed that CTC is essential to achieve fully or one-step non-autoregressive MT. In addition, Chuang et al. (2021) showed that CTC enhances the reordering behavior of non-autoregressive ST. Different from these studies, we apply CTC to improve autoregressive ST, although Haviv et al. (2021) showed that CTC helps autoregressive MT little.

There are several pioneering studies trying to relax the heuristics in acoustic features to improve speech representation. Sainath et al. (2013) and Seki et al. (2017) explored a neural filter bank layers as an alternative to the hand-engineered filterbanks. Hoshen et al. (2015) proposed a convolutional neural acoustic model that operates directly on raw waveforms, aiming at capturing the fine-grained time structure. Lam et al. (2021) further proposed a globally attentive locally recurrent network, gaining quality and robustness for ASR. These studies mainly focus on ASR. To

the best of our knowledge, applying NAFM to ST has never been investigated before, and we showed its feasibility.

8. Conclusion and Discussion

How much can we achieve for E2E ST from scratch without relying on transcripts or any pretraining? We answer this question by reexamining several techniques and devising two novel proposals, namely parameterized distance penalty (PDP) and neural acoustic feature modeling (NAFM). Via extensive experiments, we present a set of best practices for ST from scratch, including smaller vocabulary, deep post-LN encoder, wider feed-forward layer, ST-based CTC regularization and PDP. We show that ST models trained from scratch, when properly optimized, can match and even outperform previous work relying on pretraining.

Our study does not preclude pretraining (with source transcripts) for ST. Instead, we provide an improved understanding of its role on E2E ST. Our results show that pretraining matters mainly in two settings: (extremely) low-resource setup and scenarios where large-scale external ASR and MT data is available. The performance gap in such settings remains. From our perspective, how to leverage other types of data to improve pretraining for ST is a promising yet challenging research topic. We invite researchers to build upon our models to re-examine the importance of pretraining in various settings.

In addition, we examined and demonstrated the feasibility of performing E2E ST on raw waveforms through NAFM. Although we did not obtain consistent and substantial quality gains, NAFM still has the potential of fully leveraging all acoustic signals and yielding improved acoustic features for ST, achieving better results with more suitable architectures.

In the future, we are interested in exploring how the proposed techniques can advance the state-of-the-art when coupled with large-scale pretraining.

Acknowledgements

References

Anastasopoulos, A. and Chiang, D. Tied multitask learning for neural speech translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 82–91, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1008. URL https://www.aclweb.org/anthology/N18-1008.

Anastasopoulos, A., Bojar, O., Bremerman, J., Cattoni, R., Elbayad, M., Federico, M., Ma, X., Nakamura, S., Negri, M., Niehues, J., Pino, J., Salesky, E., Stüker, S., Sudoh, K., Turchi, M., Waibel, A., Wang, C., and Wiesner, M. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pp. 1–29, Bangkok, Thailand (online), August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.iwslt-1.1. URL https://aclanthology.org/2021.iwslt-1.1.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., and Weber, G. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.520.

Bahar, P., Bieschke, T., and Ney, H. A comparative study on end-to-end speech to text translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 792–799, 2019. doi: 10.1109/ASRU46091.2019.9003774.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

Bansal, S., Kamper, H., Livescu, K., Lopez, A., and Goldwater, S. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 58–68, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1006. URL https://www.aclweb.org/anthology/N19-1006.

Bérard, A., Pietquin, O., Servan, C., and Besacier, L. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on End-to-end Learning for Speech and Audio Processing*, Barcelona, Spain, 2016.

Chen, J., Ma, M., Zheng, R., and Huang, L. Mam: Masked acoustic modeling for end-to-end speech-to-text translation. *arXiv preprint arXiv:2010.11445*, 2020.

Cho, W. I., Kim, S. M., Cho, H., and Kim, N. S. kosp2e: Korean Speech to English Translation Corpus. In *Proc. Interspeech 2021*, pp. 3705–3709, 2021. doi: 10.21437/Interspeech.2021-1040.

Chuang, S.-P., Chuang, Y.-S., Chang, C.-C., and Lee, H.-y. Investigating the reordering capability in CTC-based non-autoregressive end-to-end speech translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1068–1077, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.92. URL https://aclanthology.org/2021.findings-acl.92.

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., and Turchi, M. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2012–2017, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1202. URL https://www.aclweb.org/anthology/N19-1202.

Di Gangi, M. A., Negri, M., and Turchi, M. Adapting Transformer to End-to-End Spoken Language Translation. In *Proc. Interspeech 2019*, pp. 1133–1137, 2019. doi: 10.21437/Interspeech.2019-3045. URL http://dx.doi.org/10.21437/Interspeech.2019-3045.

Dong, Q., Ye, R., Wang, M., Zhou, H., Xu, S., Xu, B., and Li, L. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12749–12759, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17509.

Du, Y., Zhang, Z., Wang, W., Chen, B., Xie, J., and Xu, T. Regularizing end-to-end speech translation with triangular decomposition agreement. *arXiv preprint arXiv:2112.10991*, 2021.

Duong, L., Anastasopoulos, A., Chiang, D., Bird, S., and Cohn, T. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 949–959, San Diego, California, June 2016. Association for Computational Lin-

guistics. doi: 10.18653/v1/N16-1109. URL https://www.aclweb.org/anthology/N16-1109.

Gaido, M., Cettolo, M., Negri, M., and Turchi, M. CTC-based compression for direct speech translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 690–696, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.57. URL https://aclanthology.org/2021.eacl-main.57.

Graves, A., Fernández, S., and Gomez, F. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *In Proceedings of the International Conference on Machine Learning, ICML 2006*, pp. 369–376, 2006.

Gu, J. and Kong, X. Fully non-autoregressive neural machine translation: Tricks of the trade. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 120–133, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.11. URL https://aclanthology.org/2021.findings-acl.11.

Han, C., Wang, M., Ji, H., and Li, L. Learning shared semantic space for speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2214–2225, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.195. URL https://aclanthology.org/2021.findings-acl.195.

Haviv, A., Vassertail, L., and Levy, O. Can latent alignments improve autoregressive machine translation? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2637–2641, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.209. URL https://aclanthology.org/2021.naacl-main.209.

Hoshen, Y., Weiss, R. J., and Wilson, K. W. Speech acoustic modeling from raw multichannel waveforms. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4624–4628, 2015. doi: 10.1109/ICASSP.2015.7178847.

Inaguma, H., Duh, K., Kawahara, T., and Watanabe, S. Multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 570–577. IEEE, 2019.

Inaguma, H., Kiyono, S., Duh, K., Karita, S., Yalta, N., Hayashi, T., and Watanabe, S. ESPnet-ST: All-in-one

speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 302–311, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.34. URL https://aclanthology.org/2020.acl-demos.34.

Indurthi, S., Zaidi, M. A., Kumar Lakumarapu, N., Lee, B., Han, H., Ahn, S., Kim, S., Kim, C., and Hwang, I. Task aware multi-task learning for speech to text tasks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7723–7727, 2021. doi: 10.1109/ICASSP39728.2021.9414703.

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N. E. Y., Yamamoto, R., Wang, X., Watanabe, S., Yoshimura, T., and Zhang, W. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 449–456, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Kocabiyikoglu, A. C., Besacier, L., and Kraif, O. Augmenting librispeech with French translations: A multimodal corpus for direct speech translation evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1001.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P07-2045.

Kudo, T. and Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

Lam, M. W., Wang, J., Weng, C., Su, D., and Yu, D. Raw Waveform Encoder with Multi-Scale Globally Attentive Locally Recurrent Networks for End-to-End Speech Recognition. In *Proc. Interspeech 2021*, pp. 316–320, 2021. doi: 10.21437/Interspeech.2021-2084.

Le, H., Pino, J., Wang, C., Gu, J., Schwab, D., and Besacier, L. Lightweight adapter tuning for multilingual speech translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 817–824, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.103. URL https://aclanthology.org/2021.acl-short.103.

Libovický, J. and Helcl, J. End-to-end non-autoregressive neural machine translation with connectionist temporal classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3016–3021, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1336. URL https://aclanthology.org/D18-1336.

Liu, Y., Xiong, H., Zhang, J., He, Z., Wu, H., Wang, H., and Zong, C. End-to-End Speech Translation with Knowledge Distillation. In *Proc. Interspeech 2019*, pp. 1128–1132, 2019. doi: 10.21437/Interspeech.2019-2582. URL http://dx.doi.org/10.21437/Interspeech.2019-2582.

Liu, Y., Zhu, J., Zhang, J., and Zong, C. Bridging the modality gap for speech-to-text translation. *ArXiv*, abs/2010.14920, 2020.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. Specaugment: A simple data augmentation method for automatic speech recognition. 2019.

Post, M. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6319.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Saharia, C., Chan, W., Saxena, S., and Norouzi, M. Non-autoregressive machine translation with latent alignments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1098–1108, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.83. URL https://aclanthology.org/2020.emnlp-main.83.

Sainath, T. N., Kingsbury, B., Mohamed, A.-r., and Ramabhadran, B. Learning filter banks within a deep neural network framework. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 297–302, 2013. doi: 10.1109/ASRU.2013.6707746.

Salesky, E., Sperber, M., and Black, A. W. Exploring phoneme-level speech representations for end-to-end speech translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1835–1841, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1179. URL https://www.aclweb.org/anthology/P19-1179.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. wav2vec: Unsupervised pre-training for speech recognition. Apr 2019. doi: http://doi.org/10.21437/Interspeech.2019-1873. URL https://arxiv.org/abs/1904.05862.

Seki, H., Yamamoto, K., and Nakagawa, S. A deep neural network integrated with filterbank learning for speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5480–5484, 2017. doi: 10.1109/ICASSP.2017.7953204.

Sennrich, R. and Zhang, B. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 211–221, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1021. URL https://aclanthology.org/P19-1021.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/

v1/P16-1162. URL https://www.aclweb.org/anthology/P16-1162.

Shaw, P., Uszkoreit, J., and Vaswani, A. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2074. URL https://aclanthology.org/N18-2074.

Sung, T.-W., Liu, J.-Y., Lee, H.-y., and Lee, L.-s. Towards end-to-end speech-to-text translation with two-pass decoding. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7175–7179. IEEE, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Wang, C., Wu, Y., Liu, S., Yang, Z., and Zhou, M. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation. *arXiv preprint arXiv:1909.07575*, 2019a.

Wang, C., Tang, Y., Ma, X., Wu, A., Okhonko, D., and Pino, J. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 33–39, Suzhou, China, December 2020a. Association for Computational Linguistics. URL https://aclanthology.org/2020.aacl-demo.6.

Wang, C., Wu, A., and Pino, J. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*, 2020b.

Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., and Chao, L. S. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1810–1822, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1176. URL https://aclanthology.org/P19-1176.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz

Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL http://arxiv.org/abs/1609.08144.

Xu, C., Hu, B., Li, Y., Zhang, Y., Huang, S., Ju, Q., Xiao, T., and Zhu, J. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2619–2630, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.204. URL https://aclanthology.org/2021.acl-long.204.

Zhang, B., Titov, I., and Sennrich, R. Improving deep transformer with depth-scaled initialization and merged attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 898–909, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1083. URL https://aclanthology.org/D19-1083.

Zhang, B., Titov, I., Haddow, B., and Sennrich, R. Adaptive feature selection for end-to-end speech translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2533–2544, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.230. URL https://aclanthology.org/2020.findings-emnlp.230.

Zhang, B., Titov, I., Haddow, B., and Sennrich, R. Beyond sentence-level end-to-end speech translation: Context helps. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2566–2578, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.200. URL https://aclanthology.org/2021.acl-long.200.

Zhao, C., Wang, M., Dong, Q., Ye, R., and Li, L. NeurST: Neural speech translation toolkit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 55–62, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021. acl-demo.7. URL https://aclanthology.org/2021.acl-demo.7.

Zheng, R., Chen, J., Ma, M., and Huang, L. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12736–12746. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zheng21a.html.