Expert Systems With Applications 248 (2024) 123317

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Speaker voice normalization for end-to-end speech translation

Zhengshan Xue [a], Tingxun Shi [b], Xiaolei Zhang [b], Deyi Xiong [a,*]

[a] College of Intelligence and Computing, Tianjin University, Tianjin, China
[b] Not affiliated to an organization, Beijing, China

ARTICLE INFO

Keywords:
Machine translation
Speech translation
Speaker normalization

ABSTRACT

Speaker voices exhibit acoustic variation. Our preliminary experiments reveal that normalized voice can significantly improve end-to-end speech translation. To mitigate the negative impact of acoustic voice variation across speakers on speech translation, we propose SVN-ST, a Speaker-Voice-Normalized end-to-end Speech Translation framework. In SVN-ST, we use synthetic speech inputs generated from a Text-to-Speech system to complement raw speech inputs. In order to explore synthetic speech inputs, we introduce two essential components for SVN-ST: an alignment adapter at the encoder side and a normalized speech knowledge distillation module at the decoder side. The former forces the representations of raw speech inputs to be close to those of synthetic (normalized) speech inputs while the latter attempts to guide the translations of raw speech inputs with those yielded from synthetic speech inputs. Two additional losses are also defined to equip with the two components. Experimental results on the MuST-C benchmark dataset demonstrate that SVN-ST outperforms previous state-of-the-art end-to-end non-normalized speech translation systems by 0.4 BLEU and cascaded speech translation systems by 2.3 BLEU. On the Covost 2 testset, SVN-ST also outperforms other normalized speech methods on robustness. Further analyses suggest that our model effectively aligns speech representations from different speakers, enhances robustness, and significantly improves sentence-level translation quality.

1. Introduction

Recent years have witnessed a shift of research focus from cascaded speech translation (ST) to end-to-end (E2E) speech translation. E2E ST directly translate the spoken source language into the target language, reducing translation latency and error propagation in comparison to cascade-based ST (Duong, Anastasopoulos, Chiang, Bird, & Cohn, 2016; Liu, Xiong, Zhang et al., 2019; Sperber & Paulik, 2020; Vila, Escolano, Fonollosa, & Costa-jussà, 2018; Zhang, Haddow, & Sennrich, 2022). A wide variety of approaches have been explored for E2E ST, particularly for the mitigation of the modality gap issue (Tang, Pino, Li, Wang, & Genzel, 2021; Wang, Wu, Liu, Yang and Zhou, 2020; Ye, Wang, & Li, 2021, 2022). Substantial improvements have been achieved by these methods, some of which even outperform traditional cascaded ST.

Despite the remarkable progress made by E2E ST in terms of translation quality, most models take raw acoustic signals as input, extract acoustic features via speech processing methods like Fbank or pre-trained speech models such as wav2vec 2.0 (Baevski, Zhou, Mohamed, & Auli, 2020), and align extracted features with texts. However, an important issue that has not been adequately addressed is the presence of acoustic voice variation across speakers. Different individuals speaking the same text may produce speeches with distinct voice characteristics, resulting in diversified speech representations (Lee et al., 2021). Such acoustic voice variation across speakers poses a negative impact on model robustness and challenges on speech modeling. Our preliminary experiments conducted on the MuST-C (Di Gangi, Cattoni, Bentivogli, Negri and Turchi, 2019) English–German (En-De) benchmark dataset have revealed that replacing raw speech signals with speeches generated by a Text-to-Speech (TTS) system is able to gain improvements ranging from 2.9 to 3.9 BLEU points on the test sets, suggesting that acoustic voice variation has a detrimental effect on E2E ST systems.

Building upon this finding from our preliminary experiments, we propose **SVN-ST**, a **S**peaker-**V**oice-**N**ormalized end-to-end **S**peech **T**ranslation framework, which explores synthetic data as a collection of unified voice "anchor points" to mitigate voice variation across different speakers. Specifically, we introduce two modules for the synthetic speech data exploration. First, we minimize the Mean Squared Error (MSE) loss between the representations of raw speech inputs and their corresponding synthetic counterparts at the encoder side. This module, referred to as alignment adapter, aims to reduce the negative impact of acoustic voice characteristics, such as timbre, on speech encoding. Second, we propose normalized speech Knowledge Distillation (KD) at the decoder side of SVN-ST. Unlike previous KD (Liu, Xiong, He et al.,

[*] Corresponding author.
E-mail addresses: xuezhengshan@tju.edu.cn (Z. Xue), tingxun.shi@gmail.com (T. Shi), zxiaolei977@outlook.com (X. Zhang), dyxiong@tju.edu.cn (D. Xiong).

https://doi.org/10.1016/j.eswa.2024.123317
Received 30 October 2023; Received in revised form 5 January 2024; Accepted 21 January 2024
Available online 22 January 2024

2019; Tang et al., 2021; Xu et al., 2021) approaches used in E2E ST, which transfer knowledge from text translation to speech translation, we transfer knowledge online from synthetic speech translation to raw speech translation. The alignment adapter forces the representations of raw speech inputs to be close to those of synthetic (normalized) speech inputs while the normalized speech KD attempts to guide the translations of raw speech inputs with those yielded from synthetic speech inputs.

Our contributions can be summarized as follows:

- We conduct preliminary experiments on the MuST-C English–German dataset, which demonstrate that normalizing acoustic features leads to a remarkable improvement in the performance of ST models.
- We propose a novel ST framework, SVN-ST, to address the negative impact of acoustic voice variation on speech modeling and translation, which includes an alignment adapter at the encoder side and a normalized speech KD module at the decoder side.
- We conduct comprehensive experiments using both the MuST-C and CoVoST 2 (Wang, Wu, Gu, & Pino, 2021) datasets. Experiment results show that our proposed framework achieves substantial improvements over state-of-the-art E2E ST approaches and speech normalization method. Furthermore, our analyses indicate that our approach is capable of enhancing model robustness by learning more compact representations for the same content spoken by different speakers.

2. Related work

E2E ST has recently gained increasing attention due to its excellent performance and low latency. However, the limited availability of parallel speech translation data poses a serious challenge to it. A wide range of methods have been proposed to alleviate this challenge. Just to name a few, Bansal, Kamper, Livescu, Lopez, and Goldwater (2018) and Xu et al. (2021), utilize additional large-scale ASR or MT data to pre-train sub-modules of E2E ST. Recognizing the similarities between MT and ST, Liu, Xiong, Zhang et al. (2019) and Tang et al. (2021) explore knowledge distillation techniques to transfer knowledge from MT to ST. Additionally, Lam, Schamoni, and Riezler (2022) employ audio alignments to synthesize speech translation data for data augmentation.

The modality gap between speech and text poses another challenge to E2E ST. Multi-task learning approaches, such as those proposed by Anastasopoulos and Chiang (2018), Dong et al. (2021), Du et al. (2022), Wang, Wu et al. (2020), Ye et al. (2021) and Zhao, Luo, Chen, and Gilman (2021), aim to leverage shared parameters among multiple tasks to learn how to align different modalities. Advanced architectures have also been developed to integrate speech-specific characteristics into the encoder, including locality modeling for self-attention (Di Gangi, Negri and Turchi, 2019) and adaptive speech representation grouping (Liu, Zhu, Zhang, & Zong, 2020; Salesky, Sperber, & Black, 2019; Zhang, Titov, Haddow and Sennrich, 2020). Han, Wang, Ji, and Li (2021) employ vector quantization techniques to learn a shared semantic space for speech and text, while Ye et al. (2022) use contrastive learning to minimize the distance between speech and text representations. Tang et al. (2021) use a novel attention-based regularization technique to pull the representations from different modalities closer. Moreover, Fang, Ye, Li, Feng, and Wang (2022) propose a method that combines speech and text representations to bridge the modality gap.

Despite the successes of the aforementioned methods, few of them explicitly consider the impact of speech attributes, such as timbre. Chen, Ma, Zheng, and Huang (2021) introduce Spectrogram Reconstruction (SpecRec), a technique that improves speech representation by recovering missing speech frames, providing an alternative solution for enhancing E2E ST. Furthermore, Lee et al. (2021) propose a speech normalization technique for improving Speech-to-Speech Translation, which leverages self-supervised discrete units obtained from HuBERT and CTC loss to remove variations in speech from multiple speakers without altering the lexical content. Huang et al. (2022) employ pre-trained HuBERT as the teacher model to generate normalized pseudo text labels as golden labels, then fine-tune a student HuBERT model (which accepts disturbed speech as input) with CTC loss to get a more certain self-supervised representation which is agnostic to acoustic variation, thus improving translation performance. Qian et al. (2022) employ a competent unsupervised voice conversion system to convert all the utterances to those spoken by a single speaker, then use HuBERT to generate a set of speech representations and finally quantize speech representations into discrete unit labels. These labels are also treated as golden labels to train another predictor whose input is masked speech representation in a self-supervised way. Gat et al. (2023) utilize HuBERT to extract discrete units from original speech. They employ CTC loss to train a quantizer between the output of the augmented signal and the deduplicated output of the original signal. Significantly different from Chen et al. (2021), we reduce acoustic voice variation by learning representations of raw speech inputs close to those of synthetic speech inputs that do not have acoustic voice variation across speakers, while Chen et al. (2021) attempt to learn better speech representations via recovering the missing speech frames. Our work is also significantly different from other related works (Gat et al., 2023; Huang et al., 2022; Lee et al., 2021; Qian et al., 2022) in that, we reduce acoustic voice variation via alignment adapter which bridges the gap between the representation of a raw speech and synthetic speech input in continuous space, while those mentioned related works use self-supervised discrete units of a reference speaker speech and perform CTC finetuning with a pre-trained speech encoder.

3. Preliminary experiments

We conducted preliminary experiments to investigate the impact of acoustic voice variation on speech translation. Specifically, we utilized TTS to synthesize audio signals from transcripts for a given dataset. We then trained two models, one using the raw speech inputs and the other using the synthetic speech inputs, and subsequently compared their performance. As multi-task learning (MTL) has been widely used in E2E ST systems, we initially trained the models for the ST task alone, followed by the gradual addition of ASR and MT tasks.

3.1. Datasets

We took the widely-used MuST-C dataset as our dataset for the preliminary experiments. Particularly, the English-to-German (En-De) task, which is the most commonly utilized task in this dataset and comprises 234,000 sentences of text and 408 h of audio, was selected for analysis. *dev* was used as the validation dataset and *tst-common* was used as the test dataset. We employed a publicly available TTS service ESPnet2-TTS[1] (Hayashi et al., 2021) to generate synthetic audio signals from transcripts. For consistency, we produced synthetic speeches for the training, validation and test datasets.

3.2. Settings

Preprocessing The input speech wave was restricted to 16-bit 16 kHz mono-channel audio, with those exceeding a duration of 30 s being removed. Regarding the text data, we trained a 10K merge SentencePiece model (Kudo & Richardson, 2018) jointly and shared the vocabulary across the source and target languages.

[1] https://huggingface.co/spaces/akhaliq/ESPnet2-TTS/blob/main/app.py.

Table 1
Preliminary experiment results on the MuST-C En-De dataset, evaluated on the test data (*tst-common*). "Raw" represents the model trained on the raw speech data, while "Synthetic" represents the model trained on the synthetic speech data.

Task(s)	Raw	Synthetic	Δ
ST alone	24.1	**28.0**	+3.9
+ASR	25.4	**28.3**	+2.9
+ASR +MT	25.5	**28.6**	+3.1

Implementation We implemented our model based on fairseq toolkit[2] (Ott et al., 2019). We used Adam optimizer (β_1 = 0.9, β_2 = 0.98) with the `inverse_sqrt` scheduler, of which the warmup updates was set to 25K, learning rate was set to 10^{-4}. The maximal number of tokens per batch was set to 500,000. Dropout and label smoothing were both set to 0.1, while the maximal number of updates was not specified. We applied the early-stop strategy, training would terminate if the best BLEU score has not been updated for 15 consecutive validation runs. We used the base architecture of wav2vec 2.0[3] as the speech encoder, which was followed by two stacked CNN layers. We set hidden size to 512, stride size to 2, and the kernel size to 5 for the CNN layers. The output of the CNN layers was then input to a standard Transformer-Base encoder–decoder model, which consists of 6 layers, 8 attention heads, a hidden size of 512, and an FFN hidden size of 2048. To comprehensively evaluate the impact of acoustic voice variation, we sequentially incorporate ASR and MT into the widely-adopted multi-task learning framework for E2E ST. The transcript and machine translation texts utilized in the evaluation were obtained solely from the MuST-C En-De task, with no external data being utilized.

During inference, we averaged the 10 consecutive checkpoints starting from the best checkpoint inclusive. Beam size was set to 10 and length penalty was set to 0.7. The scores we reported are based on the case-sensitive detokenized BLEU evaluated by SacreBLEU (Post, 2018).

3.3. Results

Table 1 presents the results obtained from the preliminary experiments. The models trained using the synthetic speech data consistently outperform their counterparts trained with the raw speech data across all tasks. Particularly, the model exclusively trained on synthetic data exhibits superior performance compared to the model trained on raw speech with multi-task learning by a margin of 2.5 BLEU (28.0 vs. 25.5). When MTL training was applied to both raw and synthetic speech data, the performance gap is further widened to 3.1 BLEU. This observation suggests that speech is influenced by various factors, including pitch, timbre, and duration, with timbre being a fundamental factor that distinguishes different utterances (Bonnici, Benning, & Saitis, 2022). The implicit normalization of diverse timbres in synthetic speech facilitates token-level attention alignment between speech frames and translations, thereby improves alignment accuracy and ultimately leads to enhanced translation quality.

Fig. 1 presents two examples of the encoder–decoder cross-attention heatmap of models on different speech sources (raw speech and synthetic speech). In the first example (the upper part of Fig. 1), the input ("And so it will be quite a different shape".) is roughly monotonically aligned to the golden reference ("So erhält es eine ganz andere Form".) and MTL with the synthetic speech input is able to learn the same pattern. MTL with the raw speech input, on the other side, fails to distribute enough attention weights for most input frames. We have similar findings in the second example.

4. SVN-ST

Inspired by the preliminary experiment results shown in Section 3, we propose an end-to-end speech translation framework, SVN-ST, which normalizes speaker voice for ST by incorporating synthetic speech data into E2E ST. Considering the computation efficiency, we share the same decoder for translating both raw speech and synthetic speech inputs. At the encoder side, we stack an extra alignment adapter over the shared encoder and calculate an alignment loss, which attempts to pull the representation of a raw speech input closer to the representation of the corresponding synthetic speech input. At the decoder side, we utilize normalized speech knowledge distillation (KD) to transfer knowledge online from synthetic speech translation to raw speech translation. The overall architecture of our proposed model is shown in Fig. 2, and will be detailed in Section 4.2.

4.1. Problem formulation and notation

In speech translation, the input to the model can be formally denoted as a triplet, $\mathcal{D} = (\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{Y})$, where $\boldsymbol{S}$ denotes the audio wave sequence, $\boldsymbol{X}$ denotes the transcript of $\boldsymbol{S}$ and $\boldsymbol{Y}$ is the translation of $\boldsymbol{X}$. End-to-End ST aims to directly predict $\boldsymbol{Y}$ given $\boldsymbol{S}$ in the inference stage, without the necessity to generate $\boldsymbol{X}$.

As our method incorporates synthetic speech input during training, to avoid ambiguity, for a given raw natural wave sequence $\boldsymbol{S}$, we denote its synthetic counterpart as $\boldsymbol{S}'$.

4.2. Model architecture

Our proposed model consists of 5 essential modules. *Speech encoder* and *Text encoder* separately encode input audio waves/texts ($\boldsymbol{S}$ or $\boldsymbol{X}$) into dense representations, and pass them to a *shared encoder*, which projects the representations from different modalities into a shared latent space. The newly introduced module *alignment adapter* is stacked over the shared encoder, which treats the representation of $\boldsymbol{S}'$ as an anchor, and tries to pull the representation of $\boldsymbol{S}$ closer to the representation of $\boldsymbol{S}'$ in order to minimize the negative impact of speaker voice variation on speech translation. Finally, a *shared decoder* generates translation output based on the representation optimized by the alignment adapter.

Speech encoder We applied the same speech pre-trained model introduced in sub- Section 3.2. To comprehensively evaluate the impact of acoustic voice variation, we sequentially incorporate ASR and MT into the widely-adopted multi-task learning framework for E2E ST. The transcript and machine translation texts utilized in the evaluation were obtained solely from the MuST-C En-De task, with no external data being utilized.

Text encoder The text encoder is implemented by a word embedding layer. It takes transcript as input for the text translation task. It should be noticed that we treat text translation task as an auxiliary task, aiming to boost the performance of ST task. The encoder will not be utilized during inference.

Shared encoder/decoder Our shared encoder/decoder has the same architecture with vanilla Transformer (Vaswani et al., 2017). With the shared encoder, we can get the representations of $\boldsymbol{S}$, $\boldsymbol{S}'$ and $\boldsymbol{X}$(denoted as $\boldsymbol{s}$, $\boldsymbol{s}'$ and $\boldsymbol{x}$ respectively). For the shared decoder, in addition to the main task ST, we apply multi-task learning on it by adding two auxiliary tasks text translation (MT) and speech recognition (ASR). We adopt the cross-entropy loss on all the tasks, defined as follows:

$$\mathcal{L}_{\mathrm{ST}} = -\sum_{t=1}^{|\boldsymbol{Y}|} \log P(Y_t | \boldsymbol{Y}_{<t}, \boldsymbol{s}) \tag{1}$$

$$\mathcal{L}_{\mathrm{MT}} = -\sum_{t=1}^{|\boldsymbol{Y}|} \log P(Y_t | \boldsymbol{Y}_{<t}, \boldsymbol{x}) \tag{2}$$

[2] https://github.com/facebookresearch/fairseq.

[3] https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_small.pt.

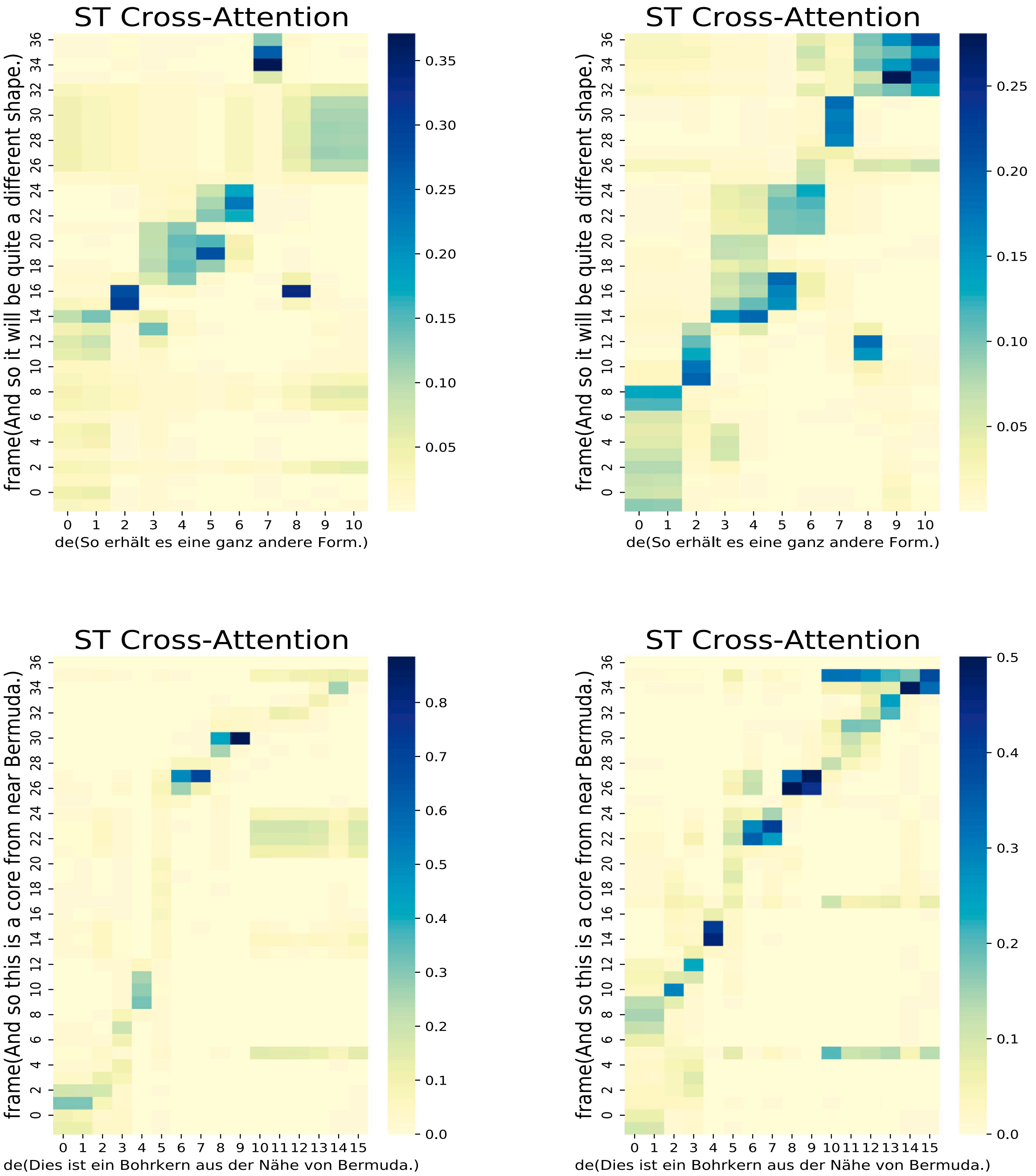


Fig. 1. Visualization of the encoder–decoder attention heatmap generated by the model on a raw speech input (left) and synthetic speech input (right).

$$\mathcal{L}_{\mathrm{ASR}} = -\sum_{t=1}^{|\boldsymbol{X}|} \log P(X_t|\boldsymbol{X}_{<t}, \boldsymbol{s}) \tag{3}$$

The incorporated synthetic speech inputs are also involved in the training of ST task, but are excluded from the training of ASR task, since adding too much synthetic speech data usually degrades the ASR performance (Bagchi, Wotherspoon, Jiang, & Muthukumar, 2020), which would further pose a negative impact on speech translation. The loss for the synthetic speech data is computed as:

$$\mathcal{L}_{\mathrm{ST}'} = -\sum_{t=1}^{|\boldsymbol{Y}|} \log P(Y_t|\boldsymbol{Y}_{<t}, \boldsymbol{s}') \tag{4}$$

Alignment adapter On the top of the shared encoder we incorporate a new module, alignment adapter, which has the same architecture with the shared encoder. The alignment adapter only takes the raw speech representation $\boldsymbol{s}$ generated by the shared encoder as its input, and outputs a synthetic-speech-aligned representation $\boldsymbol{s}_{\mathrm{align}}$. As the synthetic speeches do not have acoustic voice variation across speakers and our preliminary experiments show the positive effect of replacing raw speech inputs with corresponding synthetic speech inputs on speech translation, we introduce a loss $\mathcal{L}_{\mathrm{align}}$ to minimize the distance between $\boldsymbol{s}_{\mathrm{align}}$ and $\boldsymbol{s}'$, attempting to reduce acoustic voice variation by learning a new representation close to $\boldsymbol{s}'$. We use the Mean Squared Error (MSE) loss to calculate $\mathcal{L}_{\mathrm{align}}$.

$$\mathcal{L}_{\mathrm{align}} = \|\boldsymbol{s}_{\mathrm{align}} - \boldsymbol{s}'\|_2^2 \tag{5}$$

Normalized speech knowledge distillation In E2E ST, knowledge distillation is normally used to transfer knowledge from machine translation to speech translation. Unlike this KD methodology, we distill knowledge from the translation output of a synthetic speech input to that of a raw speech input at the decoder side. In doing so, we want the translations

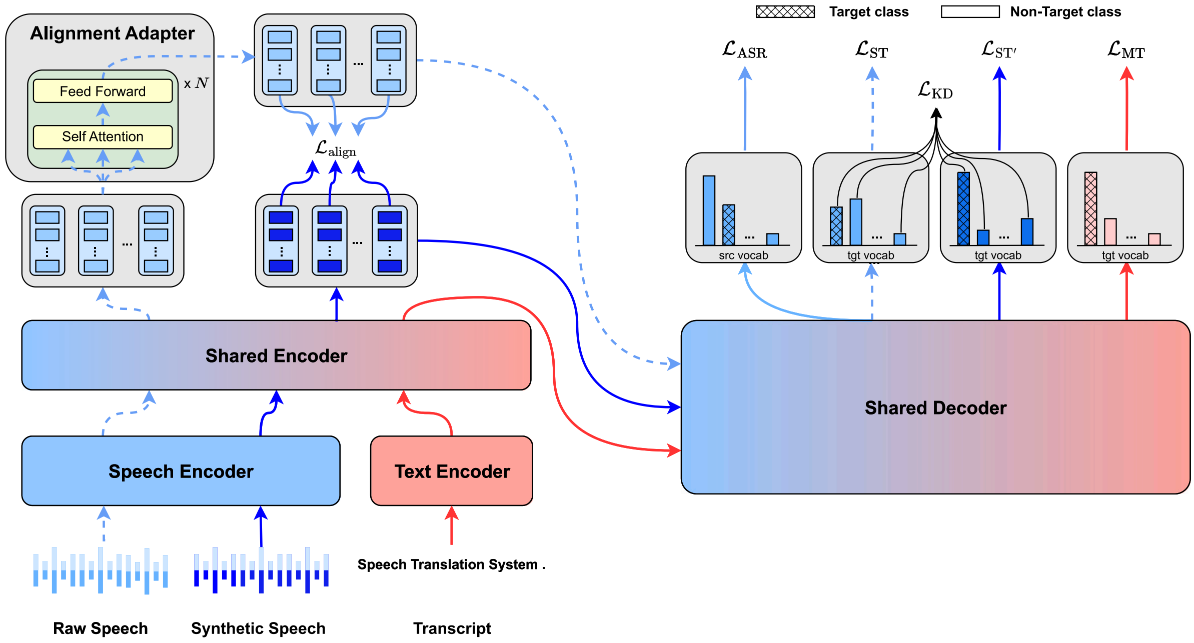


Fig. 2. Architecture of the proposed SVN-ST. Different colors of arrow lines distinguish different dataflows. Dashed lines indicate the dataflow during inference.

Table 2
BLEU scores of SVN-ST and state-of-the-art end-to-end speech translation models on the MuST-C *tst-common* set. Scores of other ST models are extracted from the corresponding papers.

Models	En-De	En-Fr	En-Es	En-Ru	En-It	En-Pt	En-Ro	En-Nl	AVG
Fairseq S2T (Wang, Tang et al., 2020)	22.7	32.9	27.2	15.3	22.7	28.1	21.9	27.3	24.8
XSTNet (Ye et al., 2021)	25.5	36.0	29.6	16.9	25.5	31.3	25.1	30.0	27.5
SpecRec (Chen et al., 2021)	22.4	32.6	26.9	14.9	22.1	27.2	24.5	26.5	24.6
STEMM (Fang et al., 2022)	25.6	36.1	30.3	17.1	25.6	31.0	24.3	30.1	27.5
E2E-ST-TDA (Du et al., 2022)	25.4	36.1	29.6	16.4	25.1	31.1	23.9	29.6	27.2
ConST (Ye et al., 2022)	25.7	36.8	30.4	17.3	26.3	32.0	24.8	30.6	28.0
MTL baseline	25.5	35.9	30.0	16.7	25.7	31.6	24.1	30.3	27.5
SVN-ST	**26.3***	**37.2***	**30.9***	**17.8***	**26.6***	**32.4***	**25.0***	**31.0***	**28.4**

* Indicate the improvement over MTL baseline is statistically significant ($p < 0.05$), estimated by paired bootstrap resampling (Koehn, 2004).

yielded from synthetic speech inputs to guide the translations from raw speech inputs. Specifically, we denote the logits generated by the shared decoder at time step t for token i as $z_i^{(t)}$, decoding temperature as τ, and the size of vocabulary as $|\mathcal{V}|$. When the input is a raw speech sequence, the estimated probability of the translated token i, $\hat{p}_i^{(t)}$, can be calculated as follows:

$$\begin{aligned}\hat{p}_i^{(t)} &= P(Y_t = i|\boldsymbol{Y}_{<t}, \boldsymbol{s}) \\ &= \frac{\exp\{\boldsymbol{z}_i^{(t)}/\tau\}}{\sum_{j=1}^{|\mathcal{V}|}\exp\{\boldsymbol{z}_j^{(t)}/\tau\}}\end{aligned} \tag{6}$$

Similarly, we denote $\hat{p}_i^{\prime(t)}$ as the estimated probability when the input is a synthetic speech sequence and compute it as:

$$\begin{aligned}\hat{p}_i^{\prime(t)} &= P(Y_t = i|\boldsymbol{Y}_{<t}, \boldsymbol{s}') \\ &= \frac{\exp\{\boldsymbol{z}_i^{\prime(t)}/\tau\}}{\sum_{j=1}^{|\mathcal{V}|}\exp\{\boldsymbol{z}_j^{\prime(t)}/\tau\}}\end{aligned} \tag{7}$$

Token-level knowledge distillation is applied and the corresponding loss is defined as:

$$\mathcal{L}_{\text{KD}} = -\sum_{t=1}^{|\boldsymbol{Y}|}\sum_{i=1}^{|\mathcal{V}|}\hat{p}_i^{\prime(t)}\log\hat{p}_i^{(t)} \tag{8}$$

Table 3
n_u values for each task. $\mathcal{L}_{\text{KD}}$ is introduced to the training process after n_u updates.

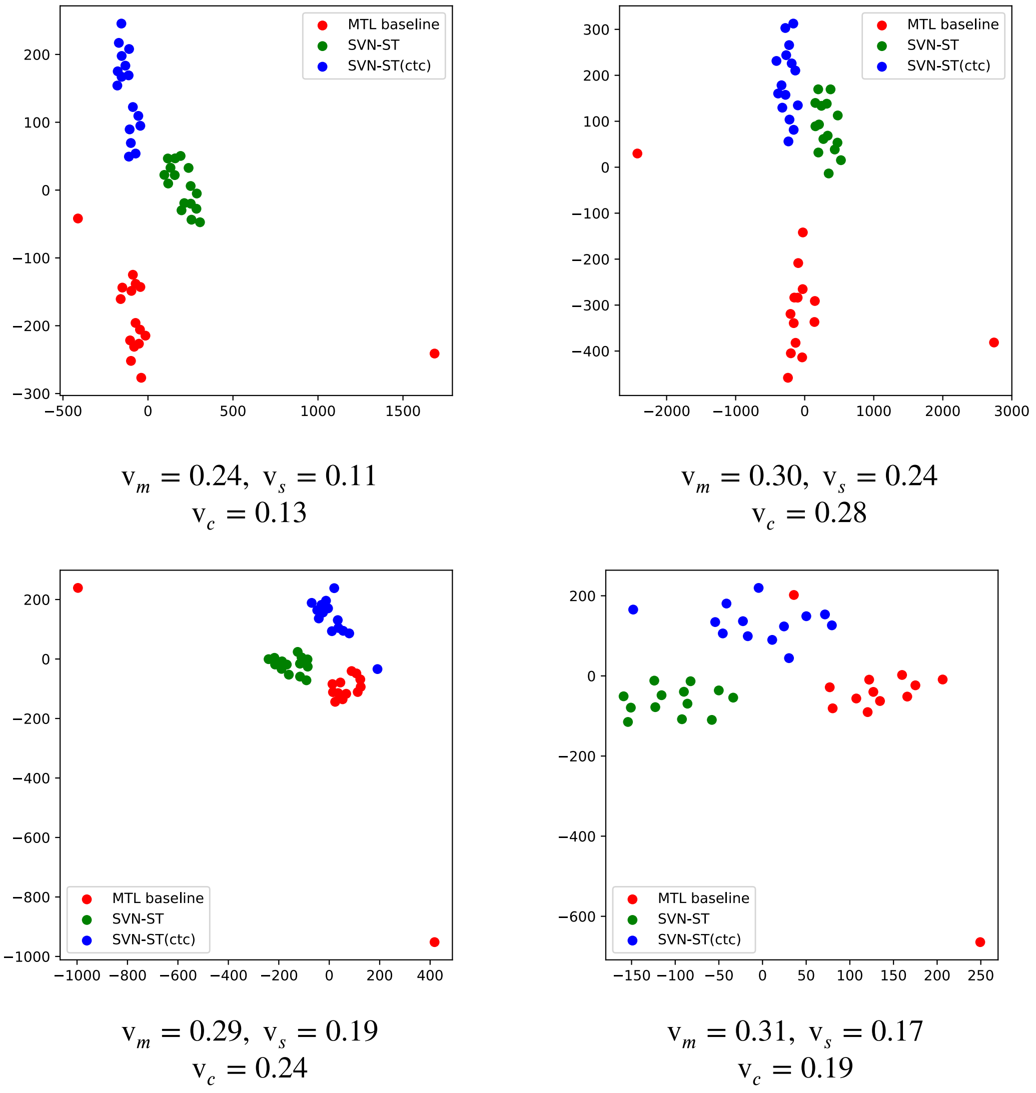
Task	n_u
En-De	85,000
En-Fr	94,000
En-Es	118,000
En-Ru	106,000
En-It	118,000
En-Pt	147,000
En-Ro	101,000
En-Nl	94,000

Table 4
BLEU scores on the MuST-C benchmark with $\mathcal{L}_{\text{ST}'}$, $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{KD}}$ being added step by step.

Models	En-De	En-De in-house TTS system	En-Fr	En-Es
MTL baseline	25.5	25.5	35.9	30.0
$+ \mathcal{L}_{\text{ST}'}$	25.8 (+0.3)	25.9 (+0.4)	36.1 (+0.2)	30.1 (+0.1)
$+ \mathcal{L}_{\text{ST}'} + \mathcal{L}_{\text{align}}$	26.0 (+0.5)	26.2 (+0.7)	36.2 (+0.3)	30.3 (+0.3)
$+ \mathcal{L}_{\text{ST}'} + \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{KD}}$	26.3 (+0.8)	26.4 (+0.9)	37.2 (+1.3)	30.9 (+0.9)

Our final objective for training is the summation of all the losses mentioned above:

$$\begin{aligned}\mathcal{L} = &\mathcal{L}_{\text{ST}} + \mathcal{L}_{\text{MT}} + \mathcal{L}_{\text{ASR}} + \\ &\mathcal{L}_{\text{ST}'} + \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{KD}}\end{aligned} \tag{9}$$



$v_m = 0.24,\ v_s = 0.11$
$v_c = 0.13$

$v_m = 0.30,\ v_s = 0.24$
$v_c = 0.28$

$v_m = 0.29,\ v_s = 0.19$
$v_c = 0.24$

$v_m = 0.31,\ v_s = 0.17$
$v_c = 0.19$

Fig. 3. Visualization of encoder representations generated by SVN-ST and the MTL baseline model. For each cluster in every subgraph, we compute the variance of the distances from each point to the centroid. The variance for the MTL baseline is denoted by v_m, while the variances for SVN-ST and SVN-ST$_{ctc}$ are denoted by v_s and v_c respectively.

4.3. Inference

During inference, SVN-ST takes a raw speech input, extracts the high-level features by the speech encoder. The extracted features are then processed sequentially by the shared encoder and alignment adapter to get the encoder representations, and finally accessed by the shared decoder to produce the corresponding target translation. This inference process is illustrated by the dashed lines in Fig. 2.

5. Experiments

We conducted extensive experiments to examine the effectiveness of SVN-ST.

5.1. Datasets

In addition to the MuST-C En-De dataset used in our preliminary experiments, we also evaluated our framework on the MuST-C English to Fr (French), Es (Spanish), Ru (Russian), It (Italian), Pt (Portuguese), Ro (Romanian) and Nl (Dutch) tasks. To be in line with previous studies, we used *dev* as the validation dataset and *tst-common* as the test dataset.

We used the same toolkit as in our preliminary experiments (see Section 3.1) to generate synthetic audio signals from transcripts. Han et al. (2021) introduce a shared semantic space to map text and speech representations of arbitrary lengths to representations of a fixed length. However, we argue that such operation has a tendency to lose information of the input. In order to keep all information of speech representations, we need to keep the duration of synthetic speech and raw speech consistent with each other. We hence performed steps to align them appropriately as follows.

1. First, we compute the duration of a raw audio, which has a sample ratio of 16K.
2. Next, we employ the aforementioned TTS service to generate an audio with a sample ratio of 22K from the transcript.
3. We then use FFmpeg[4] to convert the synthetic audio from 22K to 16K sample ratio.
4. The duration of the converted synthetic audio is subsequently computed.
5. We calculate the multiple relationship between the duration of the raw audio and that of the converted synthetic audio, and

[4] https://ffmpeg.org/.

Table 5
BLEU scores of SVN-ST vs. cascaded systems on the MuST-C *tst-common* set.

Models	En-De	En-Fr
ESPnet-ST (Inaguma et al., 2020)	23.7	33.8
In-house cascaded system	25.2	34.9
SVN-ST	**26.3**	**37.2**

Table 6
Impact of the differences in pitch on translation quality.

Data	Pitch	BLEU
MTL baseline (Raw speech)	189.5	25.5
SVN-ST (with ESPnet2 synthetic speech)	212.0	26.3
SVN-ST (with In-house synthetic speech)	238.7	26.4

Table 7
BLEU scores of SVN-ST with different pre-trained speech models on the MuST-C En-De task.

Pre-trained model	MTL baseline	SVN-ST	ΔBLEU
wav2vec 2.0	25.5	26.3	+0.8
HuBERT-base	25.4	26.6	+1.2
WavLM-base	26.6	27.6	+1.0

Table 8
Distribution of our multiple-speaker dataset. Each sentence in the dataset is uttered by at least two speakers.

# speakers	# sentences	# speakers	# sentences
2	8084	12	118
3	4381	13	49
4	874	14	30
5	790	15	18
6	1095	16	10
7	1486	17	4
8	1258	18	3
9	505	19	1
10	246	21	1
11	154		

leverage FFmpeg to scale the converted synthetic audio to match the duration of the raw audio.

6. If the scaled synthetic audio is still longer than its raw counterpart, we truncate it forcefully to achieve duration match, and vice versa.

We applied the same preprocessing pipeline, model configuration and implementation as described in Section 3.2, with the only difference that the length penalty of the En-Es and En-Nl task was set to 0.5 and En-It task was set to 0.4. For knowledge distillation, we set the temperature τ in Eqs. (6) and (7) to 1.0 for all the tasks except for the En-Fr task, which was set to 1.5. As shown by previous studies (Lee et al., 2023), a poor teacher model will degrade the performance of the student model. We hence did not perform knowledge distillation at the early stage of training, where we only trained the multi-task learning and alignment model. Instead, we implemented a warm-up strategy for incorporating the normalized speech KD into the training process for different tasks. Specifically, the normalized speech KD is not integrated into the training procedure until the model has converged after a certain number of updates denoted as n_u. The values of n_u for each task are listed in Table 3.

5.2. Baselines

To evaluate the effectiveness of SVN-ST, we compared against a series of state-of-the-art ST models.

- Fairseq S2T (Wang, Tang et al., 2020), which is an E2E ST model based on the Transformer architecture without explicit modifications.
- XSTNet (Ye et al., 2021), an E2E model that unifies automatic speech recognition and machine translation tasks through progressive multi-task training.
- STEMM (Fang et al., 2022), which combines the representations from different modalities via a mixup strategy and uses Jensen–Shannon Divergence to minimize the difference between the two output distributions from each modality.
- E2E-ST-TDA (Du et al., 2022), which introduces two different paths in the decoding process and regularizes them with a contrastive loss.
- ConST (Ye et al., 2022), a cross-modal contrastive learning method aimed at bridging the gap between the representations from different modalities.
- SpecRec (Chen et al., 2021), which learns better speech representation via recovering the missing speech frames and provides an alternative approach to improving E2E-ST.

For a comprehensive comparison, we also compared our system with two cascaded systems: **ESPnet-ST** (Inaguma et al., 2020) and an in-house cascaded ST system. In the in-house system, the ASR module is the same architecture as SVN-ST speech encoder, trained on the MuST-C speech data only with a Connectionist Temporal Classification loss. The MT module is a Vanilla Transformer Base model trained on the MuST-C parallel data with a standard MT loss. Similar to the E2E systems mentioned above, neither of the cascaded systems were trained on external machine translation data.

5.3. Results

Table 2 presents the comparison results of our SVN-ST against previous end-to-end ST models on the MuST-C tasks. In the constrained scenario where no external machine translation data are used, our proposed SVN-ST outperforms the baseline Fairseq ST model by 3.6 BLEU on average, the SpecRec by 3.8 BLEU on average, the strongest E2E ST model ConST by 0.4 BLEU on average, achieving the new state-of-the-art results. It is worth noting that our approach is better than ConST in all translation directions, indicating the stable and general effectiveness of our approach across different language pairs.

The comparison results of SVN-ST against the two cascaded systems are shown in Table 5. SVN-ST leads the in-house cascaded system by 1.1 BLEU on the En-De task and 2.3 BLEU on the En-Fr task over the in-house cascaded system. The improvements over Espnet are even larger, which are 2.6 and 3.4 BLEU respectively. These results clearly show that our model is superior to cascaded ST systems.

We also consider the impact of the differences in pitch between synthetic and raw data. The statistical information is listed in Table 6, which reveals that our improvement on BLEU score is stable, despite the variation of pitch.

5.4. Ablation study

Table 4 shows the ablation study results on the MuST-C dataset. We observe that simply incorporating synthetic speech data obtains slight improvements. This suggests that our SVN-ST does not boost the performance by data augmentation. For the En-De task, $\mathcal{L}_{\mathrm{ST}'}$, $\mathcal{L}_{\mathrm{align}}$ loss and $\mathcal{L}_{\mathrm{KD}}$ loss contribute equally, but for the En-Fr and En-Es tasks $\mathcal{L}_{\mathrm{KD}}$ loss contributes the most, with a gain of 0.6 to 1.0 BLEU.

To examine whether the improvements are robust, we experimented SVN-ST with different TTS systems and pre-trained speech models. Table 4 shows that speeches synthesized by an in-house TTS system can get comparable results (Column 1 vs. Column 2 in Table 3). Table 7 demonstrates that SVN-ST is able to obtain consistent improvements

Table 9
Win-ratios of SVN-ST vs. the MTL baseline model on the extended test dataset. The sum of win-ratios for each two compared rows is not 100% since in some cases SVN-ST ties with the MTL baseline.

Comparisons	Models	Avg. BLEU	Δ	Max. BLEU	Δ	Min. BLEU	Δ
A	**MTL baseline**	46.9%		41.5%		**49.6%**	
	+ $\mathcal{L}_{\mathrm{ST'}}$	**51.0%***	+4.1%	**41.7%***	+0.2%	40.3%*	−9.3%
B	**MTL baseline**	44.6%		37.5%		42.5%	
	+ $\mathcal{L}_{\mathrm{ST'}}$+$\mathcal{L}_{\mathrm{align}}$	**52.9%***	+8.3%	**43.4%***	+5.9%	**46.1%***	+3.6%
C	**MTL baseline**	43.6%		36.0%		42.5%	
	+ $\mathcal{L}_{\mathrm{ST'}}$+$\mathcal{L}_{\mathrm{align}}$+$\mathcal{L}_{\mathrm{KD}}$	**54.1%***	+10.5%	**44.7%***	+8.7%	**46.3%***	+3.8%
D	**MTL baseline**	47.6%		39.6%		**45.3%**	
	+ $\mathcal{L}_{\mathrm{ST'}}$+$\mathcal{L}_{\mathrm{ctc}}$	**50.2%***	+2.6%	**41.7%***	+2.1%	44.3%*	−1.0%
E	**MTL baseline**	46.2%		37.1%		**46.1** %	
	+ $\mathcal{L}_{\mathrm{ST'}}$+$\mathcal{L}_{\mathrm{ctc}}$+$\mathcal{L}_{\mathrm{KD}}$	**52.0%***	+5.8%	**44.0%***	+6.9%	44.6%	−1.5%

* Indicates that the improvement over the MTL baseline is statistically significant ($p < 0.05$), as estimated by Wilcoxon significance testing.

Table 10
BLEU scores on the extended test dataset selected from the CoVoST 2 En-De training dataset, with different losses.

Models	BLEU	Δ
ConST (Ye et al., 2022)	16.5	−1.5
MTL baseline	18.0	–
+ $\mathcal{L}_{\mathrm{ST'}}$	17.9	−0.1
+ $\mathcal{L}_{\mathrm{ST'}}$ + $\mathcal{L}_{\mathrm{align}}$	18.8	+0.8
+ $\mathcal{L}_{\mathrm{ST'}}$ + $\mathcal{L}_{\mathrm{align}}$ + $\mathcal{L}_{\mathrm{KD}}$	19.1	+1.1
+ $\mathcal{L}_{\mathrm{ST'}}$ + $\mathcal{L}_{\mathrm{ctc}}$	18.5	+0.5
+ $\mathcal{L}_{\mathrm{ST'}}$ + $\mathcal{L}_{\mathrm{ctc}}$ + $\mathcal{L}_{\mathrm{KD}}$	18.7	+0.7

Table 11
BLEU scores of SVN-ST and SVN-ST$_{\mathrm{ctc}}$ with HuBERT-base on the MuST-C En-De task.

Pre-trained model	MTL baseline	SVN-ST	SVN-ST$_{\mathrm{ctc}}$
HuBERT-base	25.4	26.6	26.3

with different pre-trained speech models, i.e., wav2vec 2.0,[5] HuBERT[6] (Hsu et al., 2021) and WavLM[7] (Chen et al., 2022).

6. Analysis

In this section, we carried out a series of in-depth analyses to further investigate the impact of synthetic data and newly introduced losses on SVN-ST. As our motivation is to utilize synthetic speech to eliminate the negative impact of acoustic voice variation, we need to collect a dataset, which contains audio recordings that have identical content but spoken by multiple speakers with distinct vocal characteristics. For avoiding data leakage, we did not use MuST-C dataset, but selected some audios from the CoVoST 2 (Wang et al., 2021) En-De training dataset as an extension to our test dataset. We filter out 19,107 $(\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{Y})$ triplets from CoVoST 2 En-De training data, where each unique transcript-translation pair $(\boldsymbol{X}, \boldsymbol{Y})$ has at least 2 different corresponding audios spoken by different speakers. The detailed distribution is listed in Table 8. We selected the MTL baseline as a strong baseline. We also chose the state-of-the-art ConST as a representative of non-normalized speech translation model. The latest four related works (Gat et al., 2023; Huang et al., 2022; Lee et al., 2021; Qian et al., 2022) on speech normalization have not been experimented on the speech-to-text task, so we cannot make direct comparisons. These four works have one thing in common: They all use pre-trained models (HuBERT) to convert speech representations into discrete units and use CTC loss for fine-tuning to eliminate the uncertainty of speech representation and enhance robustness. Therefore, we stack a pre-trained HuBERT model at the top of the shared encoder of Synthetic Speech to obtain discrete units and replace the alignment loss with CTC on the discrete units.

6.1. Validation of the ablation study

We conducted the ablation study on the extended test dataset again to verify the robustness of SVN-ST. Results are shown in Table 10. It can be seen that our proposed SVN-ST performs in the trend as on the MuST-C tasks. Furthermore, adding only the $\mathcal{L}_{\mathrm{ST'}}$ loss in MTL drops performance, once again suggesting that simply augmenting training data with synthetic speech data is not the main reason for improvements. The ConST (Ye et al., 2022) system exhibits the poorest robustness, potentially due to its approach of closing the modal gap between text and speech through contrastive loss, which in turn impairs the robustness of the speech representation when compared to the MTL baseline. Nevertheless, the utilization of CTC loss with discrete units yields a positive effect and can enhance robustness by approximately 0.5 ∼0.7 BLEU in comparison to the MTL baseline. Notably, our alignment loss with continuous speech representations surpasses the performance of CTC loss with discrete units. Furthermore, when compared to the MTL baseline, SVN-ST achieves an improvement in robustness by approximately 0.8 ∼1.1 BLEU. We also conduct CTC experiment on MuST-C En-De dataset, As shown in Table 11, SVN-ST with CTC loss outperforms MTL baseline about by 0.9 BLEU, but is weaker than SVN-ST.

6.2. Visualization of encoder representations

As we propose the alignment adapter to bridge the gap between the representation of a raw speech and synthetic speech input, we expect their distance in the shared space to be smaller than that obtained by the MTL baseline model. We hence selected four transcript-translation pairs from our extended test dataset. Each pair has more than 10 corresponding speeches. We utilized t-SNE for reducing the dimension of data points. Fig. 3 illustrates the distribution of the dimension-reduced data points, which shows that SVN-ST learns more compact representations for utterances of different speakers. We computed the variance of the all 19,107 instances in the whole extended test dataset from each point to the centroid. The estimated variances of the MTL baseline and SVN-ST$_{ctc}$ are 0.45 and 0.39 respectively, while that of SVN-ST is 0.35.

6.3. Win-ratios of sentence-level translation quality

We further conducted fine-grained sentence-level translation quality comparison of SVN-ST against the MTL baseline. For each transcription-translation pair, we calculated the mean BLEU score that is averaged across different speeches, and also kept the maximum/minimum BLEU

[5] https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_small.pt.

[6] https://dl.fbaipublicfiles.com/hubert/hubert_base_ls960.pt.

[7] https://drive.google.com/open?id=1BhTPLUkfN6e2xkqR8LEm9lByXbLY1IYd.

Table 12
Translation results generated by the MTL baseline, SVN-ST$_{ctc}$ and SVN-ST. Speeches are from the CoVoST 2 training dataset, whose transcript is “Team Four will meet up at point B with team Five.” and the corresponding reference translation is “Team Vier trifft am Punkt B mit Team Fünf zusammen.” BLEU scores are evaluated at the sentence level.

System	Output	BLEU
MTL baseline	“Teme Four” wird sich an diesem Punkt B mit "Team 5 treffen.	10.3
	“Teme Four” treffen sich mit “Team Five” mit “Teme Five”.	3.0
	“TEM4” werden wir uns auf Point B mit "TEM 5 treffen.	5.8
	“Team Four”, werden wir uns bei “P” mit "Team 5 treffen.	3.0
	Der “Teme Four” wird sich mit “Team Five” mit “Team Five” treffen.	2.9
	“Team Four” werden uns bei Point B mit “Teen 5” treffen.	5.8
	eo): “Team 4” wird sich mit “Team 5” treffen.	3.2
	TEM4 wird sich mit “Team Four” mit “Team Five” treffen.	3.7
	Die 4-Team B werden sich mit “Teme Five” treffen.	4.0
	Die “Teme Four” werden sich bei Point B mit “Team Five” treffen.	5.4
	TEM4 wird bei P. B. mit “Team Five” gemacht werden.	3.7
	Zeigen Sie mir auf, B mit Teamsides.	8.4
SVN-ST$_{ctc}$	“Team 4 wird sich bei Punkt B mit Team 5 treffen.”	19.7
	Das Team 4 wird sich mit Ting Five treffen.	5.0
	“Team 4” wird sich mit “Team 5” treffen.	4.4
	Das Team 4: Wir treffen uns auf Point P und TM 5.	3.4
	Das Team 4 wird sich bei Point B mit “Team 5” treffen.	6.6
	Das Team 4 wird sich bei Punkt B mit Team 5 treffen.	21.4
	Das Team 4 wird sich auf Point B-Will-Team 5 treffen.	4.5
	Das Team 4 wird sich bei Point B mit Team 5 treffen.	13.1
	Das Team 4 wird sich bei Punkt B mit Team 5 treffen.	21.4
	Das Team 4 wird sich bei Point B mit Team 5 treffen.	13.1
	Vierten wir uns auf Point B mit “Team Fire”.	8.1
	Temple me Dabrett, Point B mit Team V.	15.2
SVN-ST	Ein Team für vier wird sich zu Punkt B treffen, mit dem Team 5.	6.4
	Das Team IV wird sich zu Punkt B treffen, mit dem Team 5.	6.9
	Das Team Vier wird sich zu Punkt B mit Team 5 treffen.	23.9
	Das Team IV wird sich zu Punkt B mit fünf treffen.	14.3
	Das Team 4 wird sich zu Punkt B treffen, mit dem Team 5.	6.9
	Das Team vier wird sich zu Punkt B mit Teen 5 treffen.	13.1
	Das Team IV wird sich zu Punkt B mit Team 5 treffen.	21.4
	Das Team Vier wird sich zu Punkt B mit Team 5 treffen.	23.9
	Ein Team vier wird sich zu Punkt B mit Team 5 treffen.	21.4
	Das Team Vier wird sich zu Punkt B mit Team 5 treffen.	23.9
	Das Team vier wird sich zu Punkt B mit Team Five ausdenken.	21.4
	(Video) NR: Temple Meatre, B mit Teme Side.	6.8

score for the pair. For each recorded BLEU (i.e., avg/min/max BLEU), we checked the win-ratios (Zhang, Williams, Titov and Sennrich, 2020), which calculates the proportion of samples over all instances for which a method achieves a higher BLEU score. Results are shown in Table 9. Comparison A shows that $\mathcal{L}_{ST'}$ alone slightly improves the win-ratios in Avg./Max. BLEU but drops the win-ratio in Min. BLEU. This again indicates that simply using synthetic speech data is not a good option. Comparison B demonstrates that Using both $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{ST'}$, SVN-ST gains 52.9% in the average BLEU metric, outperforming the MTL baseline by 8.3 points. When the normalized speech KD is also used (Comparison C in Table 9), the improvements are upgraded to 10.5, 8.7 and 3.8 points in Avg., Min., and Max. metrics respectively. Comparison between D and E demonstrates that replacing alignment loss with CTC improves the win-ratios in Avg./Max. BLEU for SVN-ST, but slightly drops the win-ratios in Min. BLEU(−1.0% and −1.5%). Table 9 also shows that alignment loss on continuous representation outperforms the CTC loss on discrete units.

Table 12 illustrates the examples of translations yielded by the MTL baseline, SVN-ST$_{\text{ctc}}$ and SVN-ST for an audio file chosen from the CoVoST 2 training dataset, whose transcript is “Team Four will meet up at point B with team Five”. delivered by 12 different speakers. Translations yielded by SVN-ST and SVN-ST$_{\text{ctc}}$ consist of 29 and 35 unique words respectively, while translations by the MTL baseline include 42. The baseline system yields words such as “TEM”, “eo”, and “gemacht”, which should not appear in the translation, resulting in hallucination and potentially explaining why it has a more diverse lexicon. On the other hand, translations generate by SVN-ST and SVN-ST$_{\text{ctc}}$ are more consistent across different speakers and have higher average BLEU scores than those of the MTL baseline (15.9 vs. 11.3 vs. 4.9). The example shows that for speeches which are corresponding to the same content but delivered by different speakers, translations yielded by SVN-ST and SVN-ST$_{\text{ctc}}$ are more consistent with each other and less error-prone to hallucination, than those generated by the MTL baseline. Compared to SVN-ST$_{\text{ctc}}$, SVN-ST yields better results.

7. Conclusion

In this paper, we have presented SVN-ST that normalizes speaker voice for end-to-end ST. We explore synthetic speeches from TTS systems with an alignment adapter at the encoder side and a normalized speech KD module at the decoder side. This technique allows us to normalize speech inputs from different speakers, reducing the negative impact of acoustic voice variation and enhancing the model's robustness. Experimental results on the MuST-C and CoVoST 2 datasets demonstrate consistent improvements over the state-of-the-art ST systems.

CRediT authorship contribution statement

Zhengshan Xue: Conceptualization, Methodology, Software, Validation. **Tingxun Shi:** Formal analysis, Writing – original draft. **Xiaolei Zhang:** Visualization, Data curation. **Deyi Xiong:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in the paper is all open source and can be easily obtained from the internet.

Acknowledgments

This work was supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region (No. 2022D01D43) and the Key Research and Development Program of Yunnan Province (No. 202203AA080004).

References

Anastasopoulos, A., & Chiang, D. (2018). Tied multitask learning for neural speech translation. arXiv preprint arXiv:1802.06655.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460.

Bagchi, D., Wotherspoon, S., Jiang, Z., & Muthukumar, P. (2020). Speech synthesis as augmentation for low-resource ASR. arXiv preprint arXiv:2012.13004.

Bansal, S., Kamper, H., Livescu, K., Lopez, A., & Goldwater, S. (2018). Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. arXiv preprint arXiv:1809.01431.

Bonnici, R. S., Benning, M., & Saitis, C. (2022). Timbre transfer with variational auto encoding and cycle-consistent adversarial networks. In *2022 international joint conference on neural networks* (pp. 1–8). IEEE.

Chen, J., Ma, M., Zheng, R., & Huang, L. (2021). SpecRec: An alternative solution for improving end-to-end speech-to-text translation via spectrogram reconstruction.. In *Interspeech* (pp. 2232–2236).

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., et al. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1505–1518.

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., & Turchi, M. (2019). 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). Association for Computational Linguistics.

Di Gangi, M. A., Negri, M., & Turchi, M. (2019). Adapting transformer to end-to-end spoken language translation. In *Proceedings of interspeech 2019* (pp. 1133–1137). International Speech Communication Association (ISCA).

Dong, Q., Ye, R., Wang, M., Zhou, H., Xu, S., Xu, B., et al. (2021). Listen, Understand and Translate: Triple supervision decouples end-to-end speech-to-text translation. *Vol. 35*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 12749–12759). (14).

Du, Y., Zhang, Z., Wang, W., Chen, B., Xie, J., & Xu, T. (2022). Regularizing end-to-end speech translation with triangular decomposition agreement. *Vol. 36*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 10590–10598). (10).

Duong, L., Anastasopoulos, A., Chiang, D., Bird, S., & Cohn, T. (2016). 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).

Fang, Q., Ye, R., Li, L., Feng, Y., & Wang, M. (2022). STEMM: Self-learning with speech-text manifold mixup for speech translation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 7050–7062).

Gat, I., Kreuk, F., Nguyen, T. A., Lee, A., Copet, J., Synnaeve, G., et al. (2023). Augmentation invariant discrete representation for generative spoken language modeling. In *Proceedings of the 20th international conference on spoken language translation (IWSLT 2023)* (pp. 465–477). Association for Computational Linguistics.

Han, C., Wang, M., Ji, H., & Li, L. (2021). Learning shared semantic space for speech-to-text translation. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021* (pp. 2214–2225). Association for Computational Linguistics, Online.

Hayashi, T., Yamamoto, R., Yoshimura, T., Wu, P., Shi, J., Saeki, T., et al. (2021). ESPnet2-TTS: Extending the edge of TTS research. arXiv preprint arXiv:2110.07840.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451–3460.

Huang, R., Liu, J., Liu, H., Ren, Y., Zhang, L., He, J., et al. (2022). Transpeech: Speech-to-speech translation with bilateral perturbation. arXiv preprint arXiv:2205.12523.

Inaguma, H., Kiyono, S., Duh, K., Karita, S., Yalta, N., Hayashi, T., et al. (2020). 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). Association for Computational Linguistics, Online..

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 388–395).

Kudo, T., & Richardson, J. (2018). 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).

Lam, T. K., Schamoni, S., & Riezler, S. (2022). Sample, translate, recombine: Leveraging audio alignments for data augmentation in end-to-end speech translation. arXiv preprint arXiv:2203.08757.

Lee, A., Gong, H., Duquenne, P.-A., Schwenk, H., Chen, P.-J., Wang, C., et al. (2021). Textless speech-to-speech translation on real data. arXiv preprint arXiv:2112.08352.

Lee, H., Hou, R., Kim, J., Liang, D., Hwang, S. J., & Min, A. (2023). A study on knowledge distillation from weak teacher for scaling up pre-trained language models. arXiv preprint arXiv:2305.18239.

Liu, Y., Xiong, H., He, Z., Zhang, J., Wu, H., Wang, H., et al. (2019). End-to-end speech translation with knowledge distillation. arXiv preprint arXiv:1904.08075.

Liu, Y., Xiong, H., Zhang, J., He, Z., Wu, H., Wang, H., et al. (2019). End-to-end speech translation with knowledge distillation. In G. Kubin, Z. Kacic (Eds.), *Interspeech 2019, 20th annual conference of the international speech communication association, Graz, Austria, 15-19 september 2019* (pp. 1128–1132). ISCA.

Liu, Y., Zhu, J., Zhang, J., & Zong, C. (2020). Bridging the modality gap for speech-to-text translation. arXiv preprint arXiv:2010.14920.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., et al. (2019). Fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics (demonstrations)* (pp. 48–53).

Post, M. (2018). A call for clarity in reporting BLEU scores. arXiv preprint arXiv:1804.08771.

Qian, K., Zhang, Y., Gao, H., Ni, J., Lai, C.-I., Cox, D., et al. (2022). Contentvec: An improved self-supervised speech representation by disentangling speakers. In *International conference on machine learning* (pp. 18003–18017). PMLR.

Salesky, E., Sperber, M., & Black, A. W. (2019). Exploring phoneme-level speech representations for end-to-end speech translation. arXiv preprint arXiv:1906.01199.

Sperber, M., & Paulik, M. (2020). Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7409–7421). Association for Computational Linguistics, Online..

Tang, Y., Pino, J., Li, X., Wang, C., & Genzel, D. (2021). Improving speech translation by understanding and learning from the auxiliary text translation task. 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. 4252–4261).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), *Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, december 4-9, 2017, Long Beach, CA, USA* (pp. 5998–6008).

Vila, L. C., Escolano, C., Fonollosa, J. A. R., & Costa-jussà, M. R. (2018). End-to-end speech translation with the Transformer. In J. Luque, A. Bonafonte, F. A. Pujol, & A. J. S. Teixeira (Eds.), *Fourth international conference, IberSPEECH 2018, Barcelona, Spain, 21-23 november 2018, proceedings* (pp. 60–63). ISCA.

Wang, C., Tang, Y., Ma, X., Wu, A., Okhonko, D., & Pino, J. (2020). 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: Association for Computational Linguistics.

Wang, C., Wu, A., Gu, J., & Pino, J. (2021). CoVoST 2 and massively multilingual speech translation.. In *Interspeech* (pp. 2247–2251).

Wang, C., Wu, Y., Liu, S., Yang, Z., & Zhou, M. (2020). Bridging the gap between pre-training and fine-tuning for end-to-end speech translation. *Vol. 34*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 9161–9168). (05).

Xu, C., Hu, B., Li, Y., Zhang, Y., Huang, S., Ju, Q., et al. (2021). 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).

Ye, R., Wang, M., & Li, L. (2021). End-to-end speech translation via cross-modal progressive training. In H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, & P. Motlícek (Eds.), *Interspeech 2021, 22nd annual conference of the international speech communication association, Brno, Czechia, 30 august - 3 september 2021* (pp. 2267–2271). ISCA.

Ye, R., Wang, M., & Li, L. (2022). Cross-modal contrastive learning for speech translation. In *Proceedings of the 2022 conference of the north American chapter of the association for computational linguistics: human language technologies* (pp. 5099–5113).

Zhang, B., Haddow, B., & Sennrich, R. (2022). Revisiting end-to-end speech-to-text translation from scratch. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, & S. Sabato (Eds.), *Proceedings of machine learning research: Vol. 162, International conference on machine learning, ICML 2022, 17-23 july 2022, Baltimore, Maryland, USA* (pp. 26193–26205). PMLR.

Zhang, B., Titov, I., Haddow, B., & Sennrich, R. (2020). Adaptive feature selection for end-to-end speech translation. arXiv preprint arXiv:2010.08518.

Zhang, B., Williams, P., Titov, I., & Sennrich, R. (2020). Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1628–1639).

Zhao, J., Luo, W., Chen, B., & Gilman, A. (2021). Mutual-learning improves end-to-end speech translation. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 3989–3994). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.