Back Translation for Speech-to-text Translation Without Transcripts

Qingkai Fang[1,2], **Yang Feng**[1,2*]
[1]Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2]University of Chinese Academy of Sciences, Beijing, China
{fangqingkai21b, fengyang}@ict.ac.cn

Abstract

The success of end-to-end speech-to-text translation (ST) is often achieved by utilizing *source transcripts*, *e.g.*, by pre-training with automatic speech recognition (ASR) and machine translation (MT) tasks, or by introducing additional ASR and MT data. Unfortunately, transcripts are only sometimes available since numerous unwritten languages exist worldwide. In this paper, we aim to utilize large amounts of target-side monolingual data to enhance ST without transcripts. Motivated by the remarkable success of back translation in MT, we develop a back translation algorithm for ST (**BT4ST**) to synthesize pseudo ST data from monolingual target data. To ease the challenges posed by short-to-long generation and one-to-many mapping, we introduce self-supervised discrete units and achieve back translation by cascading a *target-to-unit* model and a *unit-to-speech* model. With our synthetic ST data, we achieve an average boost of 2.3 BLEU on MuST-C En→De, En→Fr, and En→Es datasets. More experiments show that our method is especially effective in low-resource scenarios.[1][2]

1 Introduction

End-to-end speech-to-text translation (ST) means directly translating speech in the source language to target text without generating source transcripts (Bérard et al., 2016; Duong et al., 2016). Different from traditional cascading methods which first transcribe the speech with automatic speech recognition (ASR) and then translate the transcripts into the target text with machine translation (MT), end-to-end ST has the potential to reduce latency and avoid error propagation. Hence, it has drawn much attention and achieved great success in recent years (Anastasopoulos et al., 2021, 2022).

However, it is challenging to train an end-to-end ST model with only speech-translation pairs. Traditional cascaded models learn cross-modal mapping with ASR and cross-lingual mapping with MT. In contrast, end-to-end ST requires simultaneous cross-modal and cross-lingual mapping, which is more complicated and usually relies on more training data. However, the amount of ST data is usually limited due to the high cost of data collection, so the ST model trained with only speech-translation pairs is usually unsatisfactory.

To tackle these problems, researchers often utilize *source transcripts* to assist ST training by introducing auxiliary ASR and MT tasks. With abundant ASR and MT data, the ASR task can help the model learn better cross-modal mapping, while the MT task can help learn better cross-lingual mapping, which can significantly improve ST as shown in recent ST studies (Wang et al., 2020a; Xu et al., 2021; Ye et al., 2021; Fang et al., 2022). Unfortunately, source transcripts are not always available. It is estimated that there are around 3000 unwritten languages in the world which have no orthography for transcription[3]. For those languages, we can no longer leverage source transcripts to help with training, so many of the latest techniques fail to benefit them.

How to train a stronger ST model without transcripts? In this paper, we address this question from the perspective of data augmentation. Although ASR and MT data are unavailable, a large amount of target-side monolingual data is still easily accessible. Motivated by the success of back translation in MT (Sennrich et al., 2016; Edunov et al., 2018), we aim to develop a back translation algorithm for ST (**BT4ST**) to synthesize pseudo ST data from monolingual target data. However, compared to text-to-text back translation, gener-

* Corresponding author: Yang Feng.

[1]Code is publicly available at https://github.com/ictnlp/BT4ST.

[2]Examples of synthetic ST data are available at https://bt4st.github.io/ and Appendix B.

[3]https://www.ethnologue.com/

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4567–4587
July 9-14, 2023 ©2023 Association for Computational Linguistics

ating source speech from the target text without source transcripts is much more challenging[4]. First, the length of text is usually only tens or hundreds, but the length of speech is about tens of thousands[5]. Therefore, the model is required to generate an extremely long sequence from a short sequence *without* the assumption of monotonic alignment, which is a more difficult sequential decision problem. Second, the conversion from text to speech is a one-to-many mapping problem due to the variations in speech (Chen et al., 2021; Ren et al., 2021). For example, the pronunciation of the same content may differ among speakers.

To address these challenges, we introduce self-supervised discrete units of the source speech as intermediate representations, and achieve back translation by cascading a *target-to-unit* model and a *unit-to-speech* model. Since the length of unit sequences is similar to the length of target characters[6], and there is a monotonic assumption between the unit sequence and the source speech, the short-to-long generation problem in *target-to-unit* and *unit-to-speech* models can be greatly alleviated. Besides, discrete units can disentangle content information from other variation information (*e.g.*, pitch and speaker) (Polyak et al., 2021), which eases the one-to-many mapping problem in the *target-to-unit* model[7]. We also introduce a speaker encoder to provide speaker information as input to the *unit-to-speech* model, which allows us to generate diverse source speeches by coupling different speaker representations. Following this pipeline, we can synthesize large amounts of pseudo ST data from monolingual target data without requiring transcripts. Finally, we train our ST model with both synthetic and real data.

We conduct experiments on MuST-C En→De, En→Fr, and En→Es datasets. By leveraging about 5M additional monolingual target data for each language pair, we achieve an average improvement of 2.3 BLEU compared with the strong baseline. We also observe that our approach is more effective in low-resource scenarios, yielding a boost of 5.6 BLEU when only 100 hours of ST data are available. In addition, we generate multiple different pseudo datasets with a simple ***Diverse*** **BT4ST** method, train several models separately with each dataset, and ensemble them together. By ensembling five models, we achieve an average boost of 4.0 BLEU in three translation directions.

[4]When transcripts are available, we can decompose it into two sub-tasks: MT for generating transcripts and text-to-speech (TTS) for speech synthesis, which becomes much more manageable.

[5]For 16kHz audio waveform, 1 second of speech corresponds to a sequence of 16,000 samples.

[6]1 second of speech corresponds to 50 discrete units.

[7]It should be noted that the one-to-many mapping problem still exist due to the many-to-many mappings between transcripts and translations, but it is not the focus of this work.

2 Background

Our work focuses on developing a back translation algorithm for ST. We first introduce the task definition and model architecture of ST in Section 2.1, and then introduce the concept of back translation in Section 2.2.

2.1 Speech-to-text Translation

Task Definition The goal of speech-to-text translation (ST) is to translate speech in one language into text in another language. We denote the source speech as $\mathbf{x} = (x_1, ..., x_I)$, where I is the length of the audio waveform. The model generates the target sentence $\mathbf{y} = (y_1, ..., y_J)$, where J is the length of the target text. In this paper, we assume that transcripts of the source speech are not available, which is a more general scenario considering about 3000 unwritten languages in the world.

Model Architecture Our ST model consists of three stacked modules: *acoustic encoder*, *length adaptor*, and *translation model*. The *acoustic encoder* is a HuBERT (Hsu et al., 2021) model pretrained on unlabelled audio data, which can generate meaningful representations for the source speech. The *length adaptor* (Li et al., 2021) is a series of convolutional layers to shrink the length of speech representations by a factor of 4. The *translation model* is a Transformer (Vaswani et al., 2017) with N encoder layers and N decoder layers, which takes the shrunken speech representations as input and outputs the target sentence. We train the ST model by minimizing the cross-entropy loss:

$$\mathcal{L}_{\mathrm{ST}} = -\sum_{j=1}^{J} \log p(y_j|\mathbf{x}, \mathbf{y}_{<j}). \tag{1}$$

2.2 Back Translation

Back translation (BT) is a simple and effective method to leverage target-side monolingual data in neural machine translation (NMT). Formally, given a parallel corpus $\mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$, and a monolingual corpus of the target language

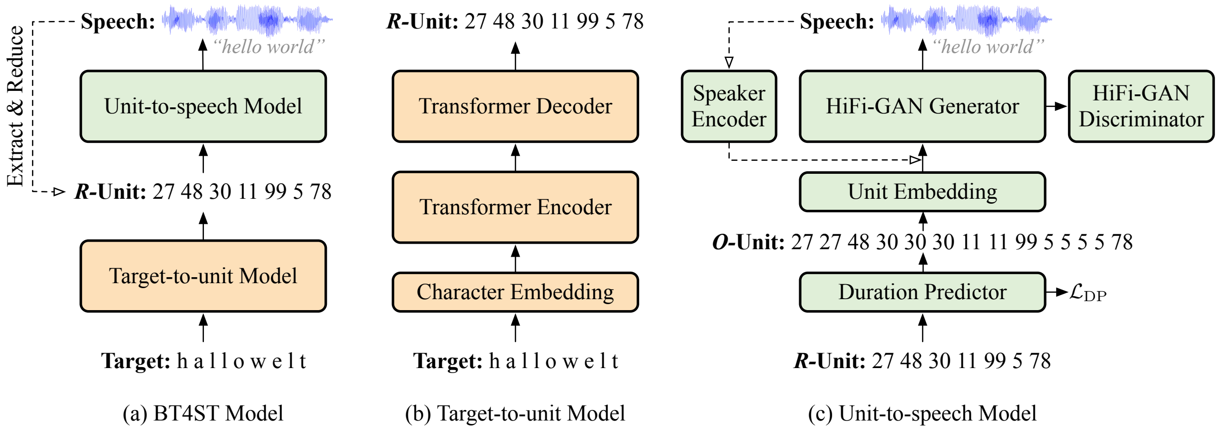


Figure 1: The overall architecture of our model. ***R*-Unit**: reduced discrete units; ***O*-Unit**: original discrete units.

$\mathcal{T} = \{\mathbf{y}^{(m)}\}_{m=1}^{M}$, we first train a *target-to-source* model on $\mathcal{D}$. Next, we use this model to generate additional pseudo parallel data $\widetilde{\mathcal{D}} = \{(\widetilde{\mathbf{x}}^{(m)}, \mathbf{y}^{(m)})\}_{m=1}^{M}$ from the monolingual corpus $\mathcal{T}$. Finally, $\widetilde{\mathcal{D}}$ can be used as a complement to $\mathcal{D}$ to train a stronger *source-to-target* model.

3 Method: BT4ST

How to acquire a *target-to-source* model for ST given a ST parallel corpus? Directly generating source speech from the target text is a challenging problem. Inspired by recent success in self-supervised discrete representation learning for speech (Baevski et al., 2020; Hsu et al., 2021; Lakhotia et al., 2021), we first transform the source speech into a sequence of discrete units with a speech pre-trained model (Section 3.1), which is used as an intermediate representation in the back translation process. With discrete units, we train a *target-to-unit* model (Section 3.2) and a *unit-to-speech* model (Section 3.3) on the parallel corpus, where the former predicts the sequence of discrete units corresponding to the source speech, and the latter converts discrete units into waveform. The model architecture is illustrated in Figure 1.

3.1 Unit-based Speech Representation

We use the pre-trained HuBERT (Hsu et al., 2021) model to generate discrete units corresponding to the source speech following Lee et al. (2022a,b). HuBERT generates 50Hz continuous representations for the input speech. We apply the K-means clustering algorithm to the continuous representations of the training data, and then transform the continuous representations into the corresponding cluster indices, *i.e.*, discrete units. Finally, the input speech $\mathbf{x} = (x_1, ..., x_I)$ is converted into a sequence of discrete units $\mathbf{z} = (z_1, ..., z_T), z_t \in \{0, 1, ..., K-1\}, \forall 1 \le t \le T$, where K is the number of clusters, and T is the number of frames where $T = \lfloor \frac{I}{320} \rfloor$. There are two advantages to use discrete units as intermediate representations rather than predicting audio waveform directly. First, the sequence of discrete units is much shorter than the audio waveform, alleviating the difficulty of short-to-long generation. Second, discrete units can disentangle speech content from the pitch and speaker information (Polyak et al., 2021), which eases the one-to-many mapping problem.

3.2 Target-to-unit Model

Our *target-to-unit* model is a Transformer-based sequence-to-sequence model, which predicts discrete units of the source speech based on the target text. The target text $\mathbf{y}$ is tokenized as characters and fed to the encoder. For the discrete units $\mathbf{z}$, we first merge repeating units into a single one to obtain the *reduced* discrete units $\mathbf{z}' = (z'_1, ..., z'_{T'})$ following Lee et al. (2022a). For example, $(1, 1, 2, 2, 2, 3, 4, 4)$ will collapse to $(1, 2, 3, 4)$. We then train the model with *reduced* discrete units as the target, which can accelerate training and inference. The training objective is as follows:

$$\mathcal{L}_{\text{T2U}} = -\sum_{t=1}^{T'} \log p(z'_t | \mathbf{y}, \mathbf{z}'_{<t}). \qquad (2)$$

3.3 Unit-to-speech Model

Our *unit-to-speech* model is a unit-based HiFi-GAN vocoder (Kong et al., 2020) as proposed in Polyak et al. (2021). It takes the *reduced* discrete units as input and generates the waveform. The model consists of four modules: *duration predictor*, *speaker encoder*, *generator*, and *discriminator*.

Duration Predictor As the output of the *target-to-unit* model is *reduced* discrete units, we add a duration predictor (Ren et al., 2021) to predict the duration of each unit. It consists of two 1D-convolutional layers with ReLU activation, each followed by layer normalization and dropout. Finally, a linear layer projects the hidden state into a scalar as duration. We denote the predicted duration vector as $\mathbf{d} = (d_1, ..., d_{T'})$, and the ground truth as $\mathbf{d}^* = (d_1^*, ..., d_{T'}^*)$. The duration predictor is optimized with Mean Squared Logarithmic Error (MSLE) between ground truth and predictions as:

$$\mathcal{L}_{\text{DP}} = \frac{1}{T'} \sum_{t=1}^{T'} (\log(1 + d_t) - \log(1 + d_t^*))^2. \quad (3)$$

Given the duration vector, we expand *reduced* discrete units by repeating each unit. For example, given $\mathbf{z}' = (1, 2, 3, 4)$ and the corresponding duration vector $\mathbf{d} = (2, 3, 1, 2)$, the expanded sequence becomes $\mathbf{z} = (1, 1, 2, 2, 2, 3, 4, 4)$. We use ground truth duration during training and predicted one during inference. Finally, the units are converted to continuous representations via a look-up table.

Speaker Encoder Discrete units contain little speaker information, but ST corpus usually contains speech from multiple speakers. To ease the one-to-many mapping problem in speech synthesis, we introduce a speaker encoder to extract the speaker information. The speaker encoder is a pre-trained speaker verification network (Wan et al., 2018), which extracts a single 256-dimensional speaker embedding from the speech. The speaker embedding is then concatenated to the representation of each unit. During inference, we randomly select a speaker embedding from the training set for each sample. It allows us to synthesize a pseudo ST dataset containing multiple speakers.

Generator and Discriminator The generator and discriminator are the same as the original HiFi-GAN (Kong et al., 2020). The generator consists of several stacked blocks, each containing a transposed convolution layer followed by multiple residual blocks. It takes the concatenated features as input and outputs the waveform $\mathbf{x}$. The discriminator contains a Multi-Period Discriminator (MPD) and a Multi-Scale Discriminator (MSD), which are used to identify the periodic or consecutive patterns in the audio. The generator and discriminator are trained adversarially. More details about the HiFi-GAN can be found in Appendix A.

Algorithm 1: Back Translation for ST

Input : ST data $\mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$,
Target data $\mathcal{T} = \{\mathbf{y}^{(m)}\}_{m=1}^{M}$,
Data selection ratio ρ

Output: ST model $M_{x \to y}$

Procedure `BT4ST`($\mathcal{D}$, $\mathcal{T}$, ρ)
- Get *reduced* discrete units $\mathbf{z}'$ from $\mathbf{x}$ to create $\mathcal{D}' = \{(\mathbf{x}^{(n)}, \mathbf{z}'^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$, for every $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}$
- Train *target-to-unit* model $M_{z' \leftarrow y}$ and *unit-to-target* model $M_{z' \to y}$ with paired data $(\mathbf{z}', \mathbf{y})$ in $\mathcal{D}'$
- Train *unit-to-speech* model $M_{x \leftarrow z'}$ with paired data $(\mathbf{x}, \mathbf{z}')$ in $\mathcal{D}'$
- Use $M_{z' \leftarrow y}$ and $M_{x \leftarrow z'}$ to create $\widetilde{\mathcal{D}} = \{(\widetilde{\mathbf{x}}^{(m)}, \widetilde{\mathbf{z}}'^{(m)}, \mathbf{y}^{(m)})\}_{m=1}^{M}$, for every $\mathbf{y} \in \mathcal{T}$
- Use $M_{z' \to y}$ to translate $\widetilde{\mathbf{z}}'$ into $\widetilde{\mathbf{y}}$, and compute $\text{BLEU}(\widetilde{\mathbf{y}}, \mathbf{y})$, for every $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{z}}', \mathbf{y}) \in \widetilde{\mathcal{D}}$
- Select top $\rho \cdot M$ samples from $\widetilde{\mathcal{D}}$ based on BLEU scores, denoted as $\widetilde{\mathcal{D}}_S$
- `Training`($\mathcal{D}$, $\widetilde{\mathcal{D}}_S$)

End

Procedure `Training`($\mathcal{D}$, $\widetilde{\mathcal{D}}_S$)
- Pre-train ST model $M_{x \to y}$ on $\widetilde{\mathcal{D}}_S$
- Fine-tune ST model $M_{x \to y}$ on $\mathcal{D}$

End

3.4 Data Selection and Model Training

By cascading the *target-to-unit* model and *unit-to-speech* model, we can synthesize ST data $\widetilde{\mathcal{D}} = \{(\widetilde{\mathbf{x}}^{(m)}, \widetilde{\mathbf{z}}'^{(m)}, \mathbf{y}^{(m)})\}_{m=1}^{M}$ from the target-side monolingual corpus $\mathcal{T} = \{\mathbf{y}^{(m)}\}_{m=1}^{M}$, where $\widetilde{\mathbf{x}}$ is the synthetic speech and $\widetilde{\mathbf{z}}'$ is the corresponding *reduced* discrete units. However, the synthetic corpus may contain some low-quality data, which may hurt model training. Therefore, we introduce *data selection* and *2-stage model training* to better utilize the synthetic data.

Data Selection To identify the low-quality data, we train a *unit-to-target* model to translate the generated units $\widetilde{\mathbf{z}}'$ back into target text $\widetilde{\mathbf{y}}$, and compute the BLEU score (Papineni et al., 2002) between $\widetilde{\mathbf{y}}$ and the ground truth $\mathbf{y}$, *i.e.,* $\text{BLEU}(\widetilde{\mathbf{y}}, \mathbf{y})$. Finally, we only keep the top $\rho \cdot M$ samples according to the BLEU score, where ρ is the selection ratio. The remaining low-quality samples are discarded.

2-stage Model Training After we get the selected synthetic ST data, we train the model in a 2-stage manner following Abdulmumin et al. (2021). We first pre-train the model on synthetic data, and then fine-tune the model on real data. This can prevent the noise-infested synthetic data from overwhelming real data during training. Algorithm 1 describes the whole process of our proposed method.

4 Experiments

4.1 Datasets

MuST-C MuST-C (Di Gangi et al., 2019) is a multilingual speech translation dataset, which contains about 400 hours of English (En) audio clips and corresponding translations in 8 languages: German (De), French (Fr), Spanish (Es), Italian (It), Portuguese (Pt), Dutch (Nl), Romanian (Ro), and Russian (Ru). We conduct experiments on En→De, En→Fr, and En→Es, because these three languages have larger public monolingual corpora.

Monolingual Target Data We use the target text in WMT (Buck and Koehn, 2016) En→De, En→Fr, and En→Es datasets as monolingual target data. Specifically, we use *europarl v7*[8], *commoncrawl*[9], *news commentary v12*[10] subsets for all three languages. The source text is never used in the experiments. The detailed statistics of data we used are shown in Table 1.

Target	MuST-C (En→)		Monolingual Data			
	hours	#samples	Euro.	CC.	NC.	All
De	408	234k	1.9M	2.4M	0.3M	4.6M
Fr	492	280k	2.0M	3.2M	0.3M	5.5M
Es	504	270k	2.0M	1.8M	0.3M	4.1M

Table 1: Statistics of all datasets. Euro.: *europarl v7*; CC.: *commoncrawl*; NC.: *news commentary v12*.

4.2 Model Settings

Target-to-unit Model The *target-to-unit* model is a Transformer with 6 encoder layers and 6 decoder layers. Each layer comprises 512 hidden states, 4 attention heads, and 2048 feed-forward hidden states. The dropout is 0.3, and the label smoothing is 0.1. For target text, we first lowercase the text and segment it into characters using SentencePiece[11]. For discrete units, we use the pre-trained quantized model[12], which learns $K = 100$ clusters from the 6th layer representations of pre-trained HuBERT-Base model[13], to convert the speech into discrete units. During training, the batch size is 400, and the maximum learning rate is 5e-4. We train the model up to 100k steps with Adam optimizer (Kingma and Ba, 2015). During inference, we use beam search with a beam size of 8 to generate the *reduced* discrete units.

The *unit-to-target* model is the same as the *target-to-unit* model except for the translation direction, which is used for data selection. During inference, we use greedy search to save time.

Unit-to-speech Model For the *unit-to-speech* model, the configurations of the generator and discriminator are the same as Polyak et al. (2021). The 1D-convolutional layers in the duration predictor are set to kernel size 3, padding 1, and hidden dimension 256. We use the pre-trained *d-vector* model[14] as the speaker encoder to extract the 256-dimensional speaker embedding. We train the model up to 100k steps with a batch size of 128.

ST Model The ST model contains three stacked modules. We use the pre-trained HuBERT-Base model as the acoustic encoder, which takes the raw 16-bit 16kHz audio wave as input. The length adaptor comprises two 1D-convolutional layers with kernel size 5, stride size 2, padding 2, and hidden dimension 1024. The translation model follows Transformer-Base architecture, containing 6 encoder layers and 6 decoder layers. Each layer comprises 512 hidden states, 8 attention heads, and 2048 feed-forward hidden states. The dropout is 0.1, and the label smoothing is 0.1. We refer to this architecture as **HU-TRANSFORMER**. We learn a vocabulary of size 8k from the target texts in the MuST-C dataset to segment the target texts into subwords for training the ST model.

To augment the ST model with back translation, we first select the top $\rho = 75\%$ of synthetic data

[8] http://statmt.org/wmt13/training-parallel-europarl-v7.tgz

[9] http://statmt.org/wmt13/training-parallel-commoncrawl.tgz

[10] http://data.statmt.org/wmt17/translation-task/training-parallel-nc-v12.tgz

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

[12] https://dl.fbaipublicfiles.com/textless_nlp/gslm/hubert/km100/km.bin

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

[14] https://github.com/yistLin/dvector

Models	External Data		En→De		En→Fr		En→Es		Average	
	Audio	Target	BLEU	Δ	BLEU	Δ	BLEU	Δ	BLEU	Δ
Previous ST baselines (No transcripts used)										
REVISIT-ST (Zhang et al., 2022a)	×	×	23.0		33.5		28.0		28.2	
W-TRANSF. (Ye et al., 2021)	✓	×	23.6		34.6		28.4		28.9	
Our implementations										
HU-TRANSFORMER	✓	×	24.3		34.9		28.7		29.3	
BT4ST	✓	✓	**26.6***	*+2.3*	**36.9***	*+2.0*	**31.2***	*+2.5*	**31.6**	*+2.3*

Table 2: BLEU scores on MuST-C En→De, En→Fr, and En→Es tst-COMMON set. * means the improvements over **HU-TRANSFORMER** are statistically significant ($p < 0.01$).

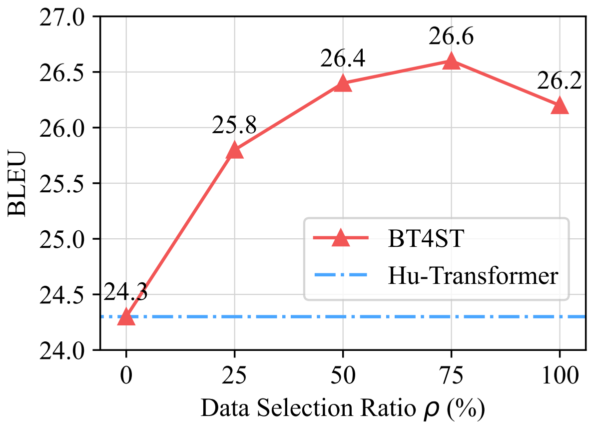


Figure 2: BLEU scores on MuST-C En→De tst-COMMON set with different data selection ratio ρ.

based on BLEU scores. The target texts in the synthetic data are segmented into subwords using the same vocabulary. We use Adam optimizer with 4k warm-up steps to pre-train the model on synthetic data up to 300k steps, and fine-tune the model on MuST-C up to 20 epochs. During both pre-training and fine-tuning, each batch contains at most 16M audio frames, and the maximum learning rate is 1e-4. During inference, we average the checkpoints of the last 10 epochs for evaluation. We use beam search with a beam size of 8. SacreBLEU[15] (Post, 2018) is used to compute case-sensitive detokenized BLEU scores and the statistical significance of translation results with paired bootstrap resampling[16] (Koehn, 2004). The length penalty is set to 1.2, 1.8, and 0.6 for En→De, En→Fr, and En→Es, respectively. We implement our model with *fairseq*[17] (Ott et al., 2019). All models are trained on 4 Nvidia Tesla V100 GPUs.

[15] https://github.com/mjpost/sacrebleu

[16] sacreBLEU signature: nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:no | tok:13a | smooth:exp | version:2.0.0

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

Baseline Systems We include two baseline systems: REVISIT-ST (Zhang et al., 2022a) and W-TRANSF. (Ye et al., 2021) for comparison. REVISIT-ST is a carefully designed ST baseline including several techniques like parameterized distance penalty (PDP) and CTC-based regularization. W-TRANSF. is a stronger ST baseline with a pre-trained acoustic model, which combines Wav2vec 2.0 (Baevski et al., 2020) and Transformer. Both of them are only trained on speech-translation pairs without using any transcripts. Our **HU-TRANSFORMER** is also a strong baseline model trained on speech-translation pairs from MuST-C, and we examine our proposed **BT4ST** on top of this by adding synthetic ST data.

5 Results and Analysis

5.1 Results on MuST-C Dataset

Table 2 shows the results on MuST-C En→De, En→Fr, and En→Es tst-COMMON set. First, we observe that our **HU-TRANSFORMER** is a strong baseline compared with previous baselines. Second, by synthesizing pseudo ST data with our proposed **BT4ST** and using it to pre-train the model, we achieve an average boost of 2.3 BLEU in three translation directions. It demonstrates that our approach can effectively utilize external monolingual target data to improve the performance of ST.

5.2 Impact of Data Selection Ratio

Although previous work (Edunov et al., 2018) found that noisy synthetic data can benefit training, we argue that the extremely low-quality data also hurt performance. As described in Section 3.4, we select top $\rho \cdot M$ synthetic data based on BLEU scores. We constrain ρ in $[0\%, 25\%, 50\%, 75\%, 100\%]$ for experiments, and the results on MuST-C En→De tst-COMMON set are shown in Figure 2. We find that filtering out

Methods	BLEU
Beam search	26.6
Greedy search	26.2
Top-10 sampling	26.4
Sampling	26.1

Table 3: BLEU scores on MuST-C En→De `tst-COMMON` set with different unit generation methods.

Synthetic Data Types	BLEU
Multi-speaker	26.6 (± 0.1)
Single-speaker	26.4

Table 4: BLEU scores on MuST-C En→De `tst-COMMON` set with different types of pseudo data. For multi-speaker data, we report the mean value and standard variance of BLEU scores derived from 3 independent experiments with different random seeds.

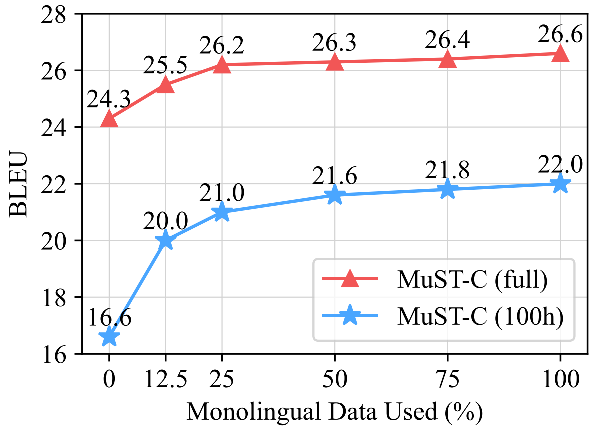


Figure 3: BLEU scores on MuST-C En→De `tst-COMMON` set with different amount of monolingual and parallel data.

the last 25% of samples (*i.e.*, $\rho = 75\%$) gives a 0.4 BLEU improvement (26.2→26.6) and performs best. We use $\rho = 75\%$ for all experiments.

5.3 Impact of Unit Generation Methods

The generation method of the *target-to-unit* model also influences the performance of back translation, as it determines the quality and diversity of synthetic data. We conduct experiments with beam search, greedy search, top-10 sampling[18], and sampling. As shown in Table 3, beam search performs best over all methods on MuST-C En→De `tst-COMMON` set. We consider two reasons for this. First, our ST dataset is actually a low-resource setting, containing only less than 300k parallel data[19]. Second, target-to-unit generation is a more difficult task compared to text translation. Therefore, beam search is more effective as it generates high probability outputs, while sampling from the model distribution may produce harmful low-quality data. We use beam search for all experiments.

5.4 Single-speaker vs. Multi-speaker Synthetic Data

As described in Section 3.3, we provide speaker embedding as input during speech synthesis, which allows us to synthesize pseudo ST datasets containing multiple speakers. To examine whether speaker diversity benefits back translation, we synthesize both multi-speaker pseudo data and single-speaker pseudo data for experiments. For multi-speaker data, we randomly select a speaker embedding from the training set for each sample. For single-speaker data, we use the average speaker embedding on the training set for all samples. As shown in Table 4, we find that the multi-speaker data is slightly better than the single-speaker data, probably because it is closer to the real ST dataset.

5.5 Results with Different Amounts of Monolingual/Parallel Data

In this section, we investigate the results under different amounts of monolingual and parallel data. First, we randomly sample 100 hours of ST data (corresponding to about 58K speech-translation pairs) from MuST-C En→De `train` set to simulate the low-resource setting. We then vary the amount of monolingual data used for back translation. As shown in Figure 3, we observe that (i) the BLEU score keeps increasing with the number of monolingual data regardless of the size of parallel data, and (ii) our approach is particularly effective in the low-resource setting. With only 12.5% monolingual data (about 0.6M), we observe a 3.4 BLEU improvement compared with the baseline. When all 4.6M monolingual data is used, we achieve a more significant boost of 5.4 BLEU.

5.6 Diverse BT4ST and Model Ensemble

Model ensemble is a widely-used technique in state-of-the-art MT systems (Barrault et al., 2020; Akhbardeh et al., 2021), which can combine differ-

[18]Top-10 sampling refers to sampling over the ten most likely words.

[19]Edunov et al. (2018) find that beam search is more effective than sampling in low-resource settings, while the opposite is true for high-resource settings.

Models	#Models	En→De BLEU	En→De Δ	En→Fr BLEU	En→Fr Δ	En→Es BLEU	En→Es Δ	Average BLEU	Average Δ
Hu-Transformer	1	24.3		34.9		28.7		29.3	
BT4ST	1	26.6	*+2.3*	36.9	*+2.0*	31.2	*+2.5*	31.6	*+2.3*
Diverse BT4ST + Ensemble	2	27.5	*+3.2*	38.3	*+3.4*	32.3	*+3.6*	32.7	*+3.4*
	3	27.9	*+3.6*	38.8	*+3.9*	32.4	*+3.7*	33.0	*+3.7*
	4	**28.1**	***+3.8***	**39.0**	***+4.1***	32.6	*+3.9*	33.2	*+3.9*
	5	28.0	*+3.7*	**39.0**	***+4.1***	**32.8**	***+4.1***	**33.3**	***+4.0***

Table 5: BLEU scores on MuST-C En→De, En→Fr, and En→Es `tst-COMMON` set with model ensemble.

ent single models (*e.g.*, models trained on different data) to achieve stronger performance. To synthesize multiple diverse pseudo datasets from a single monolingual corpus, we introduce a simple ***Diverse* BT4ST** method. For the *target-to-unit* model, we activate the dropout modules during beam search decoding. For the *unit-to-speech* model, we randomly select the speaker embedding as above. By setting different random seeds, we can generate source speeches with different content and different speakers from the target text. In this way, we obtain multiple different pseudo datasets and train several models individually. We then combine these models by computing the token-level average log probability during decoding. As shown in Table 5, model ensemble can significantly boost performance. We achieve an average boost of 4.0 BLEU in three directions by ensembling five models.

Metrics	De→En	Fr→En	Es→En
Unit-BLEU	26.5	27.9	27.2
ASR-BLEU	19.0	24.5	20.8

Table 6: Unit-BLEU and ASR-BLEU scores of the target-to-unit models.

5.7 Performance of the Target-to-unit Model

In this section, we report the performance of our target-to-unit models. We evaluate the performance with two metrics: Unit-BLEU and ASR-BLEU. Unit-BLEU is the BLEU score calculated on the reduced discrete unit sequence. ASR-BLEU is the BLEU score calculated on the transcribed text with the open-source ASR-BLEU toolkit[20]. As shown in Table 6, our target-to-unit models achieve promising results on MuST-C De→En, Fr→En, and Es→En `tst-COMMON` set, indicating that our proposed method can synthesize reasonable ST data.

6 Related Work

End-to-end ST End-to-end ST is theoretically attractive due to its advantages in alleviating error propagation and reducing latency, but it also faces many challenges because of data scarcity. Therefore, researchers often leverage *source transcripts* to help train with auxiliary tasks. Plenty of existing work first pre-train the model with the ASR task (Bansal et al., 2019; Stoian et al., 2020; Wang et al., 2020b), MT task (Han et al., 2021; Fang et al., 2022; Ye et al., 2022; Fang and Feng, 2023; Zhou et al., 2023), or both together (Wang et al., 2020a; Alinejad and Sarkar, 2020; Le et al., 2021; Dong et al., 2021a; Xu et al., 2021), and then fine-tune the model with the ST task, which becomes the *de-facto* paradigm in recent ST studies. Le et al. (2020); Indurthi et al. (2021); Tang et al. (2021a,b); Dong et al. (2021b); Ye et al. (2021) adopt multi-task learning to share knowledge among different tasks to improve ST. Bapna et al. (2021, 2022); Chen et al. (2022); Cheng et al. (2022); Ao et al. (2022); Tang et al. (2022); Zhang et al. (2022b) jointly pre-train the model with speech and text data to learn a unified space for both modalities, which achieve competitive results in ST. Jia et al. (2019); Lam et al. (2022) synthesize ST data with the help of MT model, TTS model and forced alignment tools. However, all of these studies assume that transcripts are available, which does not hold true for large numbers of unwritten languages in the world. Zhang et al. (2022a) first challenge this assumption and propose a set of practices to train a better ST model with only speech-translation pairs. In this paper, we extend this line of research and propose a back translation algorithm to utilize large-scale monolingual target data to improve ST

[20] https://github.com/facebookresearch/fairseq/tree/ust/examples/speech_to_speech/asr_bleu

without transcripts.

Back Translation Back translation for NMT was first proposed by Sennrich et al. (2016) and is widely used in state-of-the-art NMT systems (Akhbardeh et al., 2021). Since then, many techniques have been proposed to improve BT such as Iterative BT (Hoang et al., 2018; Dou et al., 2020), Tagged BT (Caswell et al., 2019; Marie et al., 2020), Tag-less BT (Abdulmumin et al., 2021), MetaBT (Pham et al., 2021), HintedBT (Ramnath et al., 2021), and so on. Edunov et al. (2018) investigates different generation methods of BT in large-scale settings. Huang et al. (2021); Liu et al. (2021) focus on combining back translation with pre-training. Xu et al. (2022) combines synthetic data generated by beam search and sampling to better trade off the importance and quality of synthetic data. Despite the success of BT in MT, BT for ST is still a challenging problem. Nguyen et al. (2022b) introduces a pipeline BT method for speech-to-speech translation which cascades an unsupervised MT model and a TTS model. In contrast, our approach can synthesize pseudo ST data from the target-side monolingual corpus without relying on source transcripts.

Discrete Speech Units Discrete units, as a self-supervised discrete representation of speech, have proved effective on many tasks, such as spoken language modeling (Lakhotia et al., 2021; Kharitonov et al., 2022; Gat et al., 2022; Borsos et al., 2022), speech-to-speech translation (Lee et al., 2022a,b; Popuri et al., 2022; Inaguma et al., 2022; Chen et al.; Li et al., 2022), speech emotion conversion (Kreuk et al., 2021), speech dialogue (Nguyen et al., 2022a), speech resynthesis (Polyak et al., 2021), speaking style conversion (Maimon and Adi, 2022), and so on. In this paper, we achieve back translation for ST by leveraging discrete units and further prove its effectiveness.

7 Conclusion and Future Work

In this paper, we develop a back translation algorithm for speech translation, which can synthesize pseudo ST data from monolingual target data without relying on transcripts. We utilize self-supervised discrete units and achieve back translation by cascading a *target-to-unit* model and a *unit-to-speech* model. Experimental results on the MuST-C benchmark demonstrate the superiority of our approach, especially in low-resource settings.

This work focuses on enhancing ST when the source transcripts are unavailable, which is an essential but under-explored issue. We hope our work will draw more attention to this issue from researchers, which will benefit more real-world unwritten languages. In the future, we are interested in exploring how to combine advanced BT techniques (*e.g.*, Iterative BT) with our approach.

Limitations

Our work provides an effective solution to augment ST when source transcripts are unavailable, which could benefit many unwritten languages. However, limited by the publicly available ST datasets, we use English as an unwritten language for experiments, which may slightly differ from real-world unwritten languages. Since we never use transcripts in our approach, we believe our work can shed some light on ST for real-world unwritten languages. We are glad to explore this if there are available datasets in the future.

Ethics Statement

Our model is developed and evaluated with publicly available datasets: MuST-C and WMT. The pre-trained models we use, like HuBERT and d-vector models, are open and permitted for research purposes. Our use of the above artifacts is consistent with their intended use since they are widely used in the speech research community. Although our method could help the speech translation of unwritten languages like some dialects, the performance of the ST model still heavily relies on the amount of ST training data. Therefore, the output of the model is not always reliable and it would be better to be assisted by professional human translators in real applications.

Acknowledgements

We thank all the anonymous reviewers for their insightful and valuable comments. This work was supported by National Key R&D Program of China (NO. 2018AAA0102502)

References

Idris Abdulmumin, Bashir Shehu Galadanci, and Aliyu Garba. 2021. Tag-less back-translation. *Mach. Transl.*, 35(4):519–549.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Ashkan Alinejad and Anoop Sarkar. 2020. Effectively pretraining a speech translation decoder with machine translation data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8014–8020, Online. Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, Dublin, Ireland. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proc. of NAACL-HLT*, pages 58–68.

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. 2022. mslam: Massively multilingual joint pre-training for speech and text. *CoRR*, abs/2202.01374.

Ankur Bapna, Yu-an Chung, Nan Wu, Anmol Gulati, Ye Jia, Jonathan H Clark, Melvin Johnson, Jason Riesa, Alexis Conneau, and Yu Zhang. 2021. Slam: A unified encoder for speech and language modeling via speech-text joint pre-training. *arXiv preprint arXiv:2110.10329*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. 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*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2022. Audiolm: a language modeling approach to audio generation.

Christian Buck and Philipp Koehn. 2016. Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 554–563, Berlin, Germany. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, sheng zhao, and Tie-Yan Liu. 2021. Adaspeech: Adaptive text to speech for custom voice. In *International Conference on Learning Representations*.

Peng-Jen Chen, Kevin Tran, Yilin Yang, Jingfei Du, Justine Kao, Yu-An Chung, Paden Tomasello, Paul-Ambroise Duquenne, Holger Schwenk, Hongyu Gong, Hirofumi Inaguma, Sravya Popuri, Changhan Wang, Juan Pino, Wei-Ning Hsu, and Ann Lee. Speech-to-speech translation for a real-world unwritten language.

Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro J. Moreno, Ankur Bapna, and Heiga Zen. 2022. Maestro: Matched speech text representations through modality matching. In *INTERSPEECH*.

Yong Cheng, Yu Zhang, Melvin Johnson, Wolfgang Macherey, and Ankur Bapna. 2022. Mu^{2}slam: Multitask, multilingual speech and language models.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 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)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021a. Consecutive decoding for speech-to-text translation. In *The Thirty-fifth AAAI Conference on Artificial Intelligence, AAAI*.

Qianqian Dong, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021b. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5894–5904, Online. Association for Computational Linguistics.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 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*, pages 949–959, San Diego, California. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Qingkai Fang and Yang Feng. 2023. Understanding and bridging the modality gap for speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 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)*, pages 7050–7062, Dublin, Ireland. Association for Computational Linguistics.

Itai Gat, Felix Kreuk, Ann Lee, Jade Copet, Gabriel Synnaeve, Emmanuel Dupoux, and Yossi Adi. 2022. On the robustness of self-supervised representations for spoken language modeling.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2214–2225, Online. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Dandan Huang, Kun Wang, and Yue Zhang. 2021. A comparison between pre-training and large-scale back-translation for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1718–1732, Online. Association for Computational Linguistics.

Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe, and Juan Pino. 2022. Unity: Two-pass direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2212.08055*.

Sathish Indurthi, Mohd Abbas Zaidi, Nikhil Kumar Lakumarapu, Beomseok Lee, Hyojung Han, Seokchan Ahn, Sangha Kim, Chanwoo Kim, and Inchul Hwang. 2021. Task aware multi-task learning for speech to text tasks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7723–7727.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *Proc. of ICASSP*, pages 7180–7184.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. 2022. Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8666–8681, Dublin, Ireland. Association for Computational Linguistics.

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

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc.

Felix Kreuk, Adam Polyak, Jade Copet, Eugene Kharitonov, Tu Anh Nguyen, Morgane Rivière, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, and Yossi Adi. 2021. Textless speech emotion conversion using decomposed and discrete representations. *CoRR*, abs/2111.07402.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2022. Sample, translate, recombine: Leveraging audio alignments for data augmentation in end-to-end speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 245–254, Dublin, Ireland. Association for Computational Linguistics.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3520–3533, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. 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)*, pages 817–824, Online. Association for Computational Linguistics.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. 2022a. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3327–3339, Dublin, Ireland. Association for Computational Linguistics.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2022b. Textless speech-to-speech translation on real data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 860–872, Seattle, United States. Association for Computational Linguistics.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Multilingual speech translation from efficient finetuning of pretrained models. 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)*, pages 827–838, Online. Association for Computational Linguistics.

Xinjian Li, Ye Jia, and Chung-Cheng Chiu. 2022. Textless direct speech-to-speech translation with discrete speech representation.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2021. On the complementarity between pre-training and back-translation for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2900–2907, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gallil Maimon and Yossi Adi. 2022. Speaking style conversion with discrete self-supervised units. *ArXiv*, abs/2212.09730.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997, Online. Association for Computational Linguistics.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2022a. Generative spoken dialogue language modeling.

Xuan-Phi Nguyen, Sravya Popuri, Changhan Wang, Yun Tang, Ilia Kulikov, and Hongyu Gong. 2022b. Improving speech-to-speech translation through unlabeled text. *arXiv preprint arXiv:2210.14514*.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Hieu Pham, Xinyi Wang, Yiming Yang, and Graham Neubig. 2021. Meta back-translation. In *International Conference on Learning Representations*.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Proc. Interspeech 2021*.

Sravya Popuri, Peng-Jen Chen, Changhan Wang, Juan Miguel Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee. 2022. Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation. In *Interspeech*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Sahana Ramnath, Melvin Johnson, Abhirut Gupta, and Aravindan Raghuveer. 2021. HintedBT: Augmenting Back-Translation with quality and transliteration hints. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1717–1733, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing asr pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7909–7913. IEEE.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, and Juan Pino. 2022. Unified speech-text pre-training for speech translation and recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1488–1499, Dublin, Ireland. Association for Computational Linguistics.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021a. 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)*, pages 4252–4261, Online. Association for Computational Linguistics.

Yun Tang, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. 2021b. A general multi-task learning framework to leverage text data for speech to text tasks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6209–6213.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez-Moreno. 2018. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 4879–4883. IEEE.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2020a. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation. In *Proc. of AAAI*, volume 34, pages 9161–9168.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020b. Curriculum pre-training for end-to-end speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3728–3738, Online. Association for Computational Linguistics.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 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)*, pages 2619–2630, Online. Association for Computational Linguistics.

Jiahao Xu, Yubin Ruan, Wei Bi, Guoping Huang, Shuming Shi, Lihui Chen, and Lemao Liu. 2022. On synthetic data for back translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 419–430, Seattle, United States. Association for Computational Linguistics.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. In *Proc. of INTERSPEECH*.

Rong Ye, Mingxuan Wang, and Lei Li. 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*, pages 5099–5113, Seattle, United States. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Rico Sennrich. 2022a. Revisiting end-to-end speech-to-text translation from scratch. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 26193–26205. PMLR.

Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Lirong Dai, Jinyu Li, and Furu Wei. 2022b. Speechut: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Online and Abu Dhabi.

Yan Zhou, Qingkai Fang, and Yang Feng. 2023. CMOT: Cross-modal mixup via optimal transport for speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

A Details about the HiFi-GAN

The HiFi-GAN vocoder (Kong et al., 2020) consists of one generator and two discriminators. Next, we describe the model architecture and training objectives of HiFi-GAN.

Generator The generator is a convolutional neural network containing several stacked blocks. Each block comprises a transposed convolution layer followed by a multi-receptive field fusion (MRF) module. The transposed convolution layers upsample the input sequence to match the length of output waveforms. The MRF module includes multiple residual blocks with different receptive fields to model different patterns in parallel. The input to the generator is concatenated unit embedding $\mathrm{Emb}(\mathbf{z})$ and speaker embedding $\mathbf{e}_{\text{spkr}}$.

Discriminator The discriminator consists of a Multi-Period Discriminator (MPD) and a Multi-Scale Discriminator (MSD). MPD is a mixture of sub-discriminators operating on equally spaced samples of the input audio. We adopt 5 sub-discriminators and set the space between samples to $[2, 3, 5, 7, 11]$ respectively. MSD is also a mixture of sub-discriminators operating on different input scales: raw audio, $\times 2$ downsampled audio, and $\times 4$ downsampled audio. MPD and MSD can identify different periodic and consecutive patterns in the audio.

Training Objectives Using G to denote the generator and D_j to denote each sub-discriminator, we define the adversarial loss of G and D_j as:

$$\mathcal{L}_{\text{adv}}(G; D_j) = \mathbb{E}_{\mathbf{x}}\left[(1 - D_j(\widehat{\mathbf{x}}))^2\right], \quad (4)$$

$$\mathcal{L}_{\text{adv}}(D_j; G) = \mathbb{E}_{\mathbf{x}}\left[(1 - D_j(\mathbf{x}))^2 + D_j(\widehat{\mathbf{x}})^2\right], \quad (5)$$

where $\mathbf{x}$ denotes the ground truth audio and $\widehat{\mathbf{x}} = G(\mathrm{Emb}(\mathbf{z}), \mathbf{e}_{\text{spkr}})$ denotes the synthetic audio.

Besides, there are two auxiliary training objectives. The first term measures the L1 distance between mel-spectrogram of the ground truth audio and synthetic audio:

$$\mathcal{L}_{\text{mel}}(G) = \mathbb{E}_{\mathbf{x}}\left[\|\phi(\mathbf{x}) - \phi(\widehat{\mathbf{x}})\|_1\right], \quad (6)$$

where ϕ is the function to compute mel-spectrogram of the audio. The second term is a feature-matching loss which measures the difference in discriminator features between ground truth audio and synthetic audio:

$$\mathcal{L}_{\text{fm}}(G; D_j) = \mathbb{E}_{\mathbf{x}}\left[\sum_{i=1}^{L} \frac{1}{N_i}\|\delta_i(\mathbf{x}) - \delta_i(\widehat{\mathbf{x}})\|_1\right], \quad (7)$$

where L denotes the number of layers in D_j, δ_i denotes the feature extractor of i-th layer, and N_i denotes the number of features in i-th layer.

Considering all sub-discriminators, the final training objectives are as follows:

$$\mathcal{L}_G = \sum_{j=1}^{J}[\mathcal{L}_{\text{adv}}(G; D_j) + \lambda_{\text{fm}}\mathcal{L}_{\text{fm}}(G; D_j)] \quad (8)$$

$$+ \lambda_{\text{mel}}\mathcal{L}_{\text{mel}}(G), \quad (9)$$

$$\mathcal{L}_D = \sum_{j=1}^{J}\mathcal{L}_{\text{adv}}(D_j; G), \quad (10)$$

where J is the number of sub-discriminators. We set $\lambda_{\text{fm}} = 2$ and $\lambda_{\text{mel}} = 45$.

B Examples of Synthetic ST Data

To understand our approach more intuitively, we provide some examples of synthetic ST data in this section. Table 7, 8, and 9 show some examples of synthetic En→De, En→Fr, and En→Es ST data, respectively. For each sample, we give the original target text, generated *reduced* discrete units, generated source speech, and the corresponding transcript obtained with a state-of-the-art ASR model Whisper-Large[21] (Radford et al., 2022). We observe that our method can generate reasonable source speech, though it may contain minor errors or duplicates. This explains why our method can enhance ST successfully.

We also provide an example of our proposed ***Diverse*** **BT4ST** method in Table 10, which generates multiple diverse source speeches from only one target text. We observe that the four outputs of our model are all reasonable and differ from each other in some way, which confirms that ***Diverse*** **BT4ST** is a simple and effective method to generate diverse pseudo data.

We provide the corresponding audio files of the above samples at `https://bt4st.github.io/`.

[21] `https://github.com/openai/whisper`

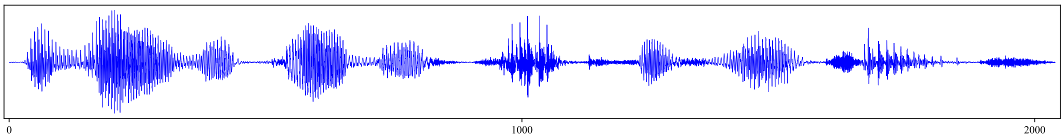
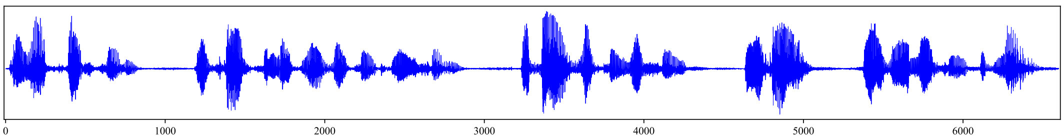
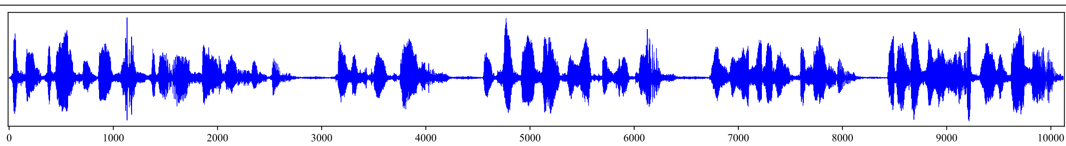
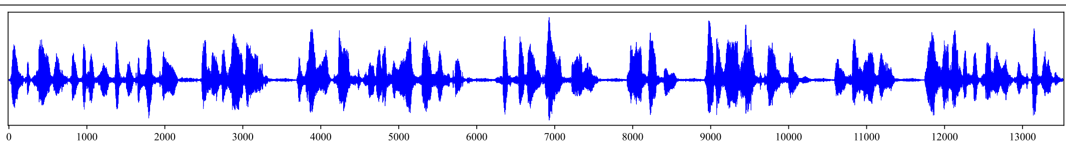
BT4ST: German text→English speech	
Case 1	
Target (De)	Die Art und Weise, wie wir diese Steuern zahlen, wird sich verändern. (*There is going to be a change in how we pay these taxes.*)
Unit	71 82 73 70 14 76 94 32 64 60 70 14 76 11 64 74 27 47 59 33 94 32 64 1 66 82 11 45 64 65 6 15 92 57 31 59 33 91 43 74 2 89 6 15 7 23 53 62 29 28 60 70 14 76 53 97 19 65 74 27 21 95 59 87 94 32 64 81 57 21 95 41 20
Speech (En)	
ASR Output	The way we pay these taxes will change.
Case 2	
Target (De)	Auch diese Frage soll letztlich Aufschluss darüber geben, welche Voraussetzungen es für die Entstehung von Leben gibt. (*This question should also provide information regarding the preconditions for the origins of life.*)
Unit	71 86 38 44 80 26 87 91 43 74 2 78 33 14 76 68 9 43 6 95 92 21 95 23 42 44 80 81 83 84 57 96 55 39 67 54 57 93 86 53 62 29 28 37 24 51 19 74 2 31 59 23 16 50 87 53 9 74 2 90 35 11 64 1 66 47 11 45 64 74 27 89 59 23 44 80 18 27 78 33 90 35 69 65 29 95 23 42 80 81 83 84 57 96 55 39 67 54 57 93 86 68 73 16 66 47 87 91 17 19 70 2 70 14 68 9 74 2 27 89 59 23 44 80 18 66 31 53 65 6 95 23 42 44 80 18 6 15 49 41 84 57 96 55 67 54 57 93 82 87 0 30 37 24 61 46 79 81 83 2 96 55 67 54 57 93 3 48 46 30 99 82 11 64 53 73 16 50 52 30 1 21 95 23 44 80 18 6 15 49 7 23 73 16 66 77 90 35 24 13 58 32 81 65 3 41 20
Speech (En)	
ASR Output	And that question is ultimately the conclusion about what conditions there are for the emergence of life.
Case 3	
Target (De)	Doch die von ihr getesteten Geräte sind auch für die Straßenplaner interessant, denn sie arbeiten nicht mit GPS und liefern nur begrenzte Informationen, die regelmäßig per Modem hochgeladen werden. (*But the devices it is testing appeal to highway planners because they don't use GPS and deliver a limited amount of information, uploaded periodically by modem.*)
Unit	71 47 76 9 74 2 82 11 45 64 29 28 92 31 23 73 16 77 24 13 58 32 65 6 15 7 23 62 29 28 92 82 87 94 32 64 74 27 31 59 33 91 43 6 49 92 31 23 62 1 85 5 30 37 51 19 65 6 49 7 87 97 19 37 86 53 44 80 18 21 95 52 25 62 6 49 92 31 87 42 88 81 83 84 57 96 55 39 67 54 57 93 3 52 30 99 82 62 6 49 92 21 52 25 45 64 74 2 27 47 59 33 90 35 13 91 38 44 80 85 5 79 29 6 49 41 84 57 96 55 39 67 54 57 93 47 11 45 64 74 27 89 59 33 68 9 29 28 92 82 87 94 32 64 1 66 31 87 97 51 19 2 66 31 23 69 70 14 46 30 74 2 78 14 76 62 1 66 21 95 45 64 74 27 47 59 45 64 87 91 43 6 15 49 41 84 57 96 55 39 67 54 57 86 38 44 80 18 66 31 23 73 90 35 53 16 50 77 53 1 85 53 1 85 53 44 80 18 65 3 52 30 16 50 87 94 32 64 65 95 23 42 80 81 83 84 57 96 55 67 54 57 93 82 87 9 85 5 30 70 52 25 9 32 1 66 89 98 53 90 35 5 30 90 35 11 64 37 68 43 74 2 47 90 35 97 1 85 23 62 1 66 27 47 24 13 58 16 50 24 61 9 85 42 16 81 20
Speech (En)	
ASR Output	But these devices they tested are also interesting for the street planners because they don't do work with GPS and the limited information that are regularly uploaded by modem.
Case 4	
Target (De)	Diese Kartelle sind dumm, wenn sie meinen, sie könnten sich unter dem Radar hinweg durchgraben, sagte die Generalstaatsanwältin des US Distrikts Southern California, Laura Duffy, bei einer Pressekonferenz vor einem Lagerhaus in San Diego, wo das eine Ende des Tunnels entdeckt worden war. (*These cartels are stupid if they think that they can dig through under the radar, said the US Attorney for the District of Southern California, Laura Diffy, at a press conference held in front of a warehouse in San Diego, where the end of the tunnel was discovered.*)
Unit	71 82 11 45 64 29 28 92 89 23 62 74 27 31 59 33 17 51 19 29 28 5 30 65 6 49 92 31 87 69 74 27 47 76 53 1 85 53 65 3 77 98 69 65 3 77 87 53 88 74 2 89 98 69 74 27 89 23 62 1 66 31 87 53 32 1 66 2 3 59 52 25 69 81 84 96 55 39 67 54 57 93 86 68 44 80 18 85 5 30 99 82 73 70 52 25 94 32 64 1 85 87 24 61 46 79 81 83 84 57 96 55 39 67 54 57 93 82 62 1 66 21 95 87 38 44 80 60 52 25 19 74 27 47 33 52 24 61 43 6 15 49 7 23 62 74 27 31 59 69 1 85 5 30 25 73 16 99 82 11 98 69 14 76 87 9 43 6 15 7 87 94 32 64 1 66 31 53 62 6 49 92 21 52 25 42 74 2 31 41 84 57 96 55 39 67 54 57 93 6 15 7 87 9 1 66 6 15 49 7 87 68 99 82 5 30 44 80 18 27 89 59 33 91 17 19 73 65 3 48 46 30 44 80 98 53 42 81 83 84 57 96 55 39 67 54 57 93 90 35 48 46 30 25 73 1 66 31 87 68 43 3 77 11 64 81 83 84 57 96 55 39 67 54 57 93 86 91 9 85 73 70 14 76 94 0 30 75 91 17 43 6 15 49 92 89 33 24 61 68 44 80 18 65 3 52 25 42 44 80 18 2 6 15 49 41 84 57 96 55 39 67 54 57 93 86 53 44 80 18 6 15 49 7 87 38 44 80 18 85 11 64 53 94 32 1 66 89 87 97 19 81 83 84 57 96 55 39 67 54 57 93 70 14 76 0 30 70 14 68 44 80 85 87 38 44 80 18 85 23 73 16 99 82 73 62 74 27 31 59 33 68 44 80 85 19 70 14 76 62 29 28 92 31 23 62 6 49 92 89 87 68 16 77 5 79 1 85 10 83 20
Speech (En)	
ASR Output	These cattles are stupid if you think you could dig through under the radar, the general prostitutor of the USA district, said Southern California, Laura Duffy, at a warehouse conference in San Diego, where one end of the tunnel was discovered.

Table 7: Examples of synthetic En→De ST data.

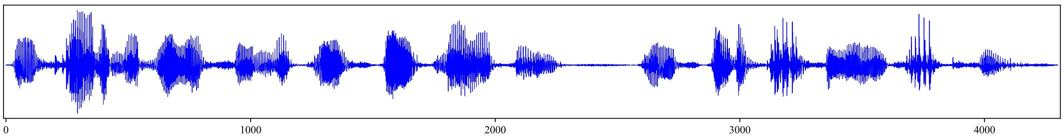
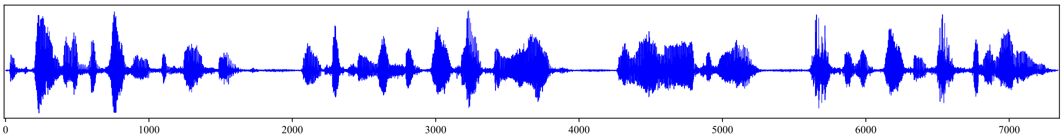
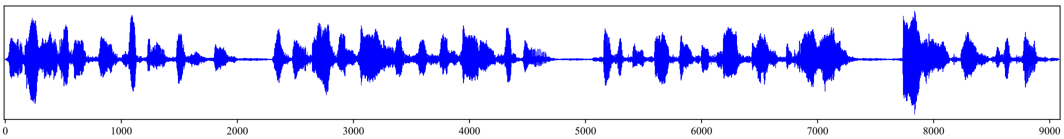
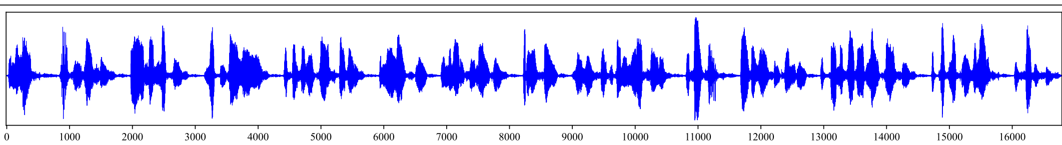
BT4ST: French text→English speech	
Case 1	
Target (Fr)	Ces automobilistes paieront bientôt à État des frais kilométriques au lieu des taxes sur essence. (*Those drivers will soon pay the mileage fees instead of gas taxes to the state.*)
Unit	71 82 11 45 64 29 28 37 24 61 9 85 73 16 50 14 19 16 66 47 11 53 90 35 62 6 15 49 7 69 44 80 60 70 14 19 74 27 57 47 59 33 94 32 64 65 6 15 49 92 57 31 87 94 32 64 74 2 66 50 24 13 58 17 19 65 3 77 11 45 64 81 84 57 96 55 39 67 54 57 86 53 44 80 18 6 15 49 92 57 31 87 9 85 73 16 77 66 27 89 87 91 43 6 15 7 23 19 90 35 11 64 44 80 18 27 31 59 33 91 43 2 89 6 15 7 11 64 81 29 6 15 41 20
Speech (En)	
ASR Output	These automobilists soon will pay state mile fee instead of gasoline taxis.
Case 2	
Target (Fr)	Un scandale sur la présence de viande de cheval dans des plats cuisinés a éclaté en Europe au début de année, à la suite de tests effectués en Irlande. (*A scandal on the presence of horse meat in prepared meals had broken out in Europe at the beginning of the year, following tests carried out in Ireland.*)
Unit	71 86 62 6 15 49 92 89 87 91 38 44 80 18 85 19 90 35 73 16 99 82 73 74 27 47 52 25 9 29 28 23 44 80 18 6 15 7 23 73 16 77 75 33 48 51 19 65 6 49 92 50 11 45 64 2 27 31 41 84 57 96 55 39 67 54 57 93 86 53 44 80 18 27 89 59 53 74 2 21 95 23 44 80 18 66 31 53 65 6 95 23 53 62 29 28 92 27 47 52 25 97 65 74 27 89 59 33 91 17 43 2 31 23 53 44 80 18 98 69 48 46 30 25 42 16 74 2 47 41 83 84 57 96 55 39 67 54 57 93 86 53 44 80 82 11 64 0 87 9 90 35 11 64 87 94 32 64 1 21 95 23 73 16 77 98 45 64 0 42 79 81 83 84 57 96 55 39 67 54 57 93 86 91 43 3 2 31 23 73 60 70 14 76 62 29 28 92 27 47 52 25 97 65 74 27 89 23 44 80 18 27 31 59 87 91 43 6 15 49 92 31 23 53 1 85 53 44 80 26 24 13 58 90 35 42 80 18 81 31 41 20
Speech (En)	
ASR Output	A scandal of the presence of horse meat in kitchen dishes pro-cut in Europe in the early age of year after it was pro-contested in Ireland.
Case 3	
Target (Fr)	La seule autre compagnie imposant de tels frais est la compagnie hongroise Wizz Air, a déclaré le consultant auprès de compagnies aériennes Jay Sorensen, qui suit de près les frais en supplément. (*The only other airline with such a fee is Hungary Wizz Air, said airline consultant Jay Sorensen, who closely tracks add-on fees.*)
Unit	71 82 11 45 64 37 51 90 35 11 64 37 68 99 82 5 30 65 3 52 25 11 45 64 29 28 92 27 89 59 33 68 16 74 2 47 23 44 80 85 11 64 65 6 15 7 87 68 9 74 21 95 23 62 29 28 49 3 77 11 45 64 29 28 41 84 57 96 55 39 67 54 57 93 86 53 62 29 28 92 82 11 45 64 37 86 94 0 30 75 33 68 44 80 18 66 89 98 87 0 30 25 11 32 53 44 80 60 70 14 76 53 62 29 28 60 70 14 76 53 62 29 28 75 33 68 44 80 18 66 89 98 87 0 30 25 11 32 53 44 80 18 27 89 59 33 68 16 18 2 47 23 42 44 80 85 11 64 81 83 84 57 96 55 39 67 54 57 93 6 15 7 87 9 1 66 82 73 74 27 89 59 23 44 80 18 6 15 49 7 24 51 19 74 2 31 23 53 88 18 27 31 59 23 69 1 66 21 95 87 94 32 64 65 6 15 49 7 60 48 46 30 25 44 80 18 6 15 7 23 44 80 37 94 0 30 90 35 24 13 58 32 44 80 18 29 28 15 41 84 57 96 55 39 67 54 57 93 3 24 61 51 19 90 35 97 14 76 88 18 74 27 78 33 90 35 97 65 6 49 92 31 23 99 82 73 74 27 78 33 24 61 43 6 15 49 92 31 41 20
Speech (En)	
ASR Output	The only other freeze companies such as fees is the Air Hungarian with Hungarian company said the consulting to Jay Sorensen Airlines following close to the cost.
Case 4	
Target (Fr)	Ces dernières accusations se basent entre autres sur des lettres rédigées par avocat des frères Sainte-Croix, Me Émile Perrin, dans les années 1990, mais aussi par les recherches faites dans les archives à ce sujet par le frère Wilson Kennedy, un ancien frère de Sainte-Croix qui a dénoncé publiquement les sévices. (*The latter accusations are partly based on letters written by the lawyer of the brothers of the Holy Cross, Mr Emile Perrin QC, in the 1990s, as well as through research carried out in the archives on this subject by Brother Wilson Kennedy, a former brother of the Holy Cross who has publicly denounced the abuses.*)
Unit	71 86 53 44 80 82 73 90 35 24 61 43 6 15 49 92 31 41 84 96 55 39 67 54 57 93 86 91 43 74 2 89 98 69 29 28 87 94 32 65 6 95 23 42 80 18 29 28 49 41 84 96 55 67 54 57 93 86 24 61 46 30 25 73 16 50 24 68 80 18 90 35 87 9 43 74 2 31 59 23 42 29 28 49 41 84 96 55 67 54 57 93 70 52 25 53 74 2 23 80 18 66 47 24 13 58 9 90 35 48 46 76 58 32 0 42 29 28 49 41 84 96 55 67 54 57 93 86 73 16 99 82 73 75 33 97 90 35 11 64 74 27 78 52 24 61 43 6 49 92 47 52 24 68 99 82 5 42 29 28 49 41 84 96 55 67 54 57 93 86 73 16 50 11 45 64 53 90 35 94 32 64 74 2 6 49 92 50 11 64 81 84 96 55 67 54 57 93 86 53 44 80 82 73 44 80 26 24 13 58 44 18 27 31 59 45 64 80 26 24 13 58 44 18 27 31 59 11 64 29 28 49 41 84 96 55 67 54 57 93 47 76 9 85 60 48 51 19 65 6 49 7 87 97 69 81 84 96 55 67 54 57 93 3 99 60 52 25 69 60 70 52 25 11 45 64 65 6 49 7 87 42 9 74 2 21 95 92 31 87 68 44 80 85 53 44 80 82 11 64 87 24 61 43 74 2 89 59 33 24 13 58 42 16 77 3 41 84 96 55 67 54 57 93 86 73 16 66 47 87 91 17 43 2 66 82 87 91 43 2 31 41 84 96 55 67 54 57 93 47 24 13 58 9 99 82 73 70 14 76 97 19 65 6 49 7 23 44 80 18 27 89 59 33 38 44 80 85 42 62 1 85 11 64 81 84 96 55 67 54 57 93 86 73 3 48 51 16 50 76 73 16 66 47 52 24 68 99 82 5 30 73 16 77 90 35 97 69 65 74 27 78 52 24 61 43 6 49 7 23 42 29 28 49 41 84 96 55 67 54 57 93 82 73 74 2 27 47 33 68 16 66 47 90 35 53 74 2 89 90 35 11 64 1 66 31 23 73 44 80 26 87 91 17 44 18 2 6 49 92 31 41 84 96 55 67 54 57 93 82 62 6 49 7 87 9 16 77 62 6 49 7 42 29 28 49 41 20
Speech (En)	
ASR Output	In the last. Accusations are among letters written by lawyers of the Holy Cross Brothers ameliates me in the 1990s, but also through research done in the archive about that by the Wilson-Kennedy, a former Brother of Low Crosses that publicly denounced the services.

Table 8: Examples of synthetic En→Fr ST data.

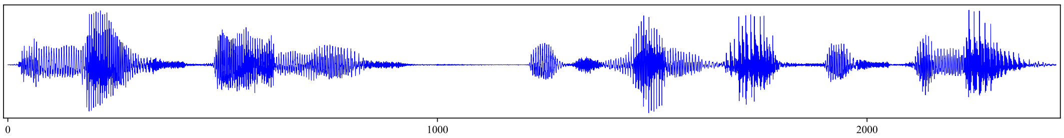
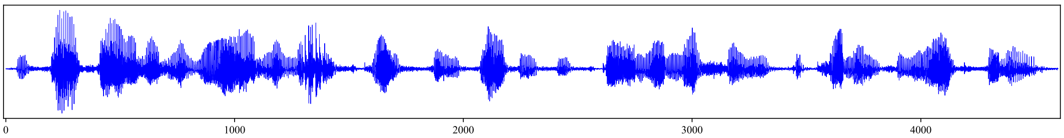
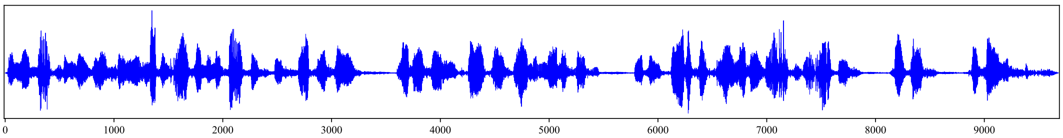
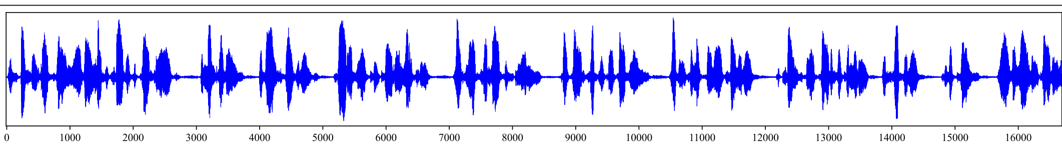
BT4ST: Spanish text→English speech	
Case 1	
Target (Es)	En la mayoría de las familias, cada uno desayuna solo. (*In a majority of families, everyone has breakfast separately.*)
Unit	71 86 53 44 80 18 50 24 97 65 6 49 92 3 77 87 94 38 17 16 50 90 35 11 64 29 28 49 41 84 57 96 55 67 54 57 40 57 93 86 45 64 74 21 95 60 70 14 68 44 80 18 66 47 52 25 91 43 74 2 3 77 23 62 6 49 92 31 23 73 19 90 35 24 68 97 81 20
Speech (En)	
ASR Output	In most families, each one breakfast alone.
Case 2	
Target (Es)	Os sentáis al volante en la costa oeste, en San Francisco, y vuestra misión es llegar los primeros a Nueva York. (*You get in the car on the west coast, in San Francisco, and your task is to be the first one to reach New York.*)
Unit	71 93 98 45 65 6 15 7 87 91 9 74 2 31 87 91 17 68 44 80 85 68 44 80 82 73 70 14 76 45 64 53 17 19 35 68 44 80 82 73 70 14 76 87 91 43 6 15 49 92 57 89 59 33 87 97 65 6 15 49 92 57 31 23 53 44 80 18 6 15 7 87 38 43 16 3 77 23 44 18 6 15 7 23 62 6 15 49 92 57 89 87 97 14 76 44 80 98 5 30 16 50 87 53 65 6 95 23 42 44 80 85 53 62 29 6 15 92 57 31 23 73 74 27 57 89 59 33 68 16 50 77 53 44 80 18 27 31 59 23 44 80 26 11 98 0 5 46 30 74 2 89 6 15 7 53 1 85 11 64 81 83 20
Speech (En)	
ASR Output	You sat down on the wheel on the west coast in San Francisco and your mission is to come into New York City.
Case 3	
Target (Es)	En la redacción seguramente tendremos una discusión sobre esto durante varios días, sobre si durante la programación del juego ninguno pensó siquiera un poco o si los autores de verdad nos toman por una panda de bobos. (*For several days, our editorial staff kept discussing whether when designing the game no one thought for a moment or whether its authors really think we are such morons.*)
Unit	71 86 53 44 80 60 52 25 11 64 1 66 31 87 91 43 74 2 21 95 23 44 80 60 70 14 76 45 64 60 70 14 76 53 97 19 65 6 95 5 30 90 35 11 64 75 91 9 16 77 23 44 80 37 24 46 30 1 66 89 98 53 16 50 77 42 44 80 85 73 16 66 47 87 91 17 43 74 2 82 53 62 6 49 60 3 52 30 65 6 49 7 87 9 16 77 52 25 19 1 66 31 87 94 32 64 29 28 41 84 57 96 55 39 67 54 57 93 70 14 76 9 99 82 5 30 1 66 31 87 5 30 25 88 18 66 27 47 59 33 90 35 94 32 64 74 27 47 52 25 97 16 66 78 52 25 38 16 50 77 88 18 26 87 68 44 80 85 73 16 99 82 87 42 16 50 81 84 57 96 55 39 67 54 57 93 45 64 16 77 23 44 80 18 6 3 7 24 61 9 85 73 16 66 47 87 53 9 74 2 37 48 46 30 70 14 76 9 99 82 5 30 99 82 11 64 87 91 43 74 2 21 95 60 14 19 37 24 61 43 99 3 82 5 30 29 28 49 41 84 57 96 55 67 54 57 40 57 31 59 94 32 74 2 89 87 68 43 6 49 41 84 96 55 67 54 57 93 3 52 30 73 16 66 47 87 94 38 44 80 18 85 73 16 65 6 49 7 69 81 29 28 49 41 20
Speech (En)	
ASR Output	In reduction, we will surely have an argument about this for several days, whether during play programming none of them even thought of it, or whether the actual authors take us for a band of zoos.
Case 4	
Target (Es)	El avión especial, en el que tuvieron que viajar a Podgorica, entre otros los vicepresidentes Dalibor Kucera y Rajchl, no ha salido esta mañana por problemas técnicos en Praga, teniendo que buscarse una alternativa para transportar a parte del comité ejecutivo al lugar donde tendrá lugar el partido de clasificación. (*An extra flight that was to take, among others, the Vice Presidents Dalibor Kučera and Rajchl to Podgorica did not leave Prague in the morning due to a technical failure. An alternative was sought for in order to transport part of the executive board to the scene of the barraged rematch.*)
Unit	71 82 62 6 49 92 47 87 9 43 6 95 23 19 37 86 0 30 74 2 47 59 33 90 35 94 32 64 44 80 60 70 14 46 30 99 82 87 94 32 75 91 9 1 66 31 23 62 74 27 21 59 52 25 91 17 16 77 19 74 27 31 23 73 74 27 47 33 24 61 43 1 66 78 48 46 30 25 42 43 74 2 89 41 83 84 57 96 55 39 67 54 57 93 86 73 16 50 24 68 88 18 37 68 99 82 5 79 29 28 49 41 84 96 55 67 54 57 93 82 73 16 66 77 24 13 58 32 65 6 49 92 47 52 25 9 29 28 23 1 85 53 42 44 80 18 2 6 15 49 41 84 96 55 67 54 57 93 66 31 87 91 17 19 90 35 73 16 66 47 48 46 30 74 2 27 78 33 69 65 74 27 89 59 98 0 30 25 37 86 38 44 80 18 70 52 25 91 43 74 2 21 95 60 48 19 81 83 84 57 96 55 39 67 54 57 93 75 91 9 29 28 92 26 24 61 43 74 2 89 59 33 68 16 50 87 91 17 43 74 2 82 62 6 49 92 50 48 46 30 80 85 42 88 81 83 84 57 96 55 67 54 57 93 3 52 30 16 74 27 47 52 24 61 43 1 66 2 27 31 59 87 9 43 74 2 26 73 74 27 89 78 19 74 27 47 52 24 61 16 66 47 90 35 42 16 50 18 29 28 49 41 84 96 55 39 67 54 57 93 75 91 9 16 77 88 18 27 31 59 23 73 90 35 76 74 2 3 52 30 44 80 26 24 51 19 65 74 27 31 59 33 52 30 44 80 85 53 1 85 53 42 16 77 3 41 84 96 55 67 54 40 57 31 23 62 74 27 21 59 52 25 38 44 80 18 6 49 92 47 48 46 30 74 2 27 47 33 24 46 30 1 85 73 16 99 82 73 74 27 89 59 23 73 16 50 53 1 85 11 64 81 84 96 55 67 54 57 93 86 53 1 66 89 29 28 7 87 9 43 74 2 89 98 53 1 85 53 42 16 77 3 41 84 96 55 67 54 40 57 93 31 59 23 99 82 73 74 27 47 59 33 90 35 94 32 64 65 6 49 41 84 96 55 67 54 40 57 93 70 14 76 0 30 99 82 73 70 52 25 94 32 64 85 88 18 27 47 33 24 46 30 1 85 11 64 60 70 14
Speech (En)	
ASR Output	The special airplane where they had to travel to Pagoric. Among others, the vice-presidents Dalibor, Kukir, and Rachel has not come out this morning for broad technical problems, having to look for an alternative to transport part of the committee executive to the place where the rating party was going to be held.

Table 9: Examples of synthetic En→Es ST data.

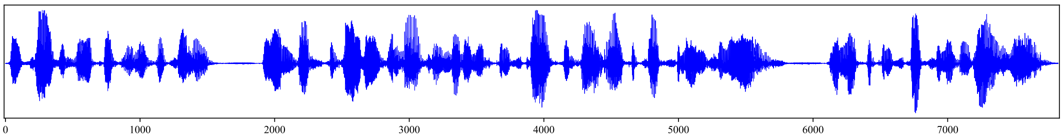
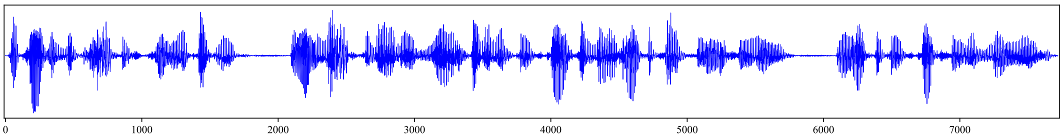
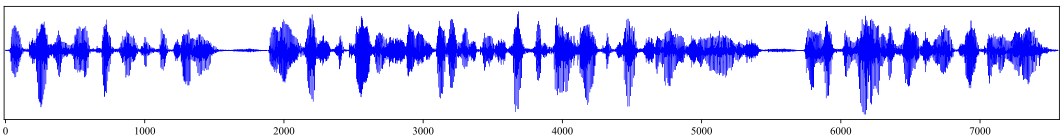
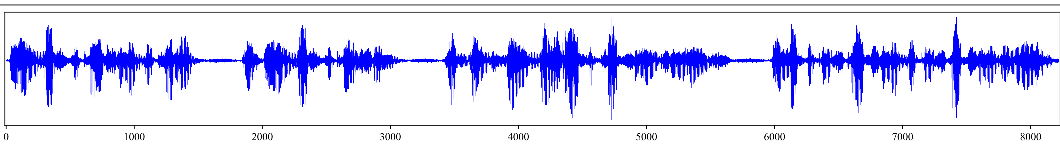
Diverse **BT4ST: One German text→Multiple English speeches**	
Target (De)	Wegen der aufwendigen Ausstattung des Tunnels gehen die Ermittler davon aus, dass er von Architekten und Ingenieuren konstruiert wurde und dass der Bau rund ein Jahr in Anspruch nahm. (*Due to the elaborate configuration of the tunnel, investigators are working on the assumption that it was build by architects and engineers and that the construction took around one year.*)
Output 1	
Unit	71 47 11 45 64 74 27 89 78 33 24 61 68 9 29 28 23 73 16 99 82 11 64 53 73 74 27 78 14 76 53 16 74 2 50 42 44 80 85 73 16 99 82 73 62 74 27 31 59 33 68 44 80 85 19 81 83 84 57 96 55 39 67 54 57 93 82 11 45 64 53 44 80 18 66 77 87 9 43 6 49 92 31 23 73 1 66 89 94 32 64 85 5 30 29 28 23 73 16 50 87 91 9 1 21 95 23 42 44 80 18 82 87 9 85 53 60 70 14 76 62 29 28 92 57 89 23 44 80 18 6 49 92 21 52 24 61 43 2 31 23 53 1 66 47 24 13 58 37 24 46 30 74 2 89 23 62 74 27 31 59 87 91 9 43 74 2 6 49 7 23 44 80 85 87 38 44 80 18 21 95 23 44 80 26 11 45 64 0 79 29 28 49 41 84 57 96 55 39 67 54 57 86 38 44 80 82 87 9 74 2 82 73 74 27 89 23 44 80 18 6 49 92 52 24 68 43 74 2 95 23 44 80 85 73 16 77 73 52 70 52 25 91 17 44 18 85 11 98 45 64 0 79 81 83 20
Speech (En)	
ASR Output	Because of the equipment of the tunnel. The investigators imagined that it was constructed by architects and engineers, and that the construction of around a year.
Output 2	
Unit	71 47 45 74 27 89 59 33 68 9 29 28 23 73 16 99 66 82 73 74 27 47 59 33 91 43 6 15 7 23 62 1 21 95 92 27 31 59 33 68 44 80 85 19 90 35 73 74 27 78 14 76 53 9 74 2 50 42 44 80 18 2 31 41 84 57 96 55 39 67 54 57 93 82 11 45 64 53 44 80 18 66 77 87 91 43 6 15 49 92 31 23 1 66 89 87 94 32 64 85 5 30 29 28 60 70 52 25 94 32 64 29 28 92 82 87 9 85 53 60 70 14 76 62 29 28 92 57 89 23 44 80 18 6 49 92 21 52 24 61 43 74 2 31 23 53 1 66 47 24 13 58 37 24 46 30 74 2 89 23 62 74 27 31 59 87 91 43 74 2 6 15 7 23 44 80 26 87 38 44 80 18 21 95 23 44 80 26 11 0 79 29 28 41 84 57 96 55 39 67 54 57 93 86 38 44 80 82 87 9 74 2 82 73 74 27 89 23 44 80 18 6 49 92 21 52 24 68 43 74 2 65 95 23 44 80 60 70 14 76 62 29 28 60 70 14 68 44 80 18 98 0 79 81 83 20
Speech (En)	
ASR Output	Because of the passage tunnel equipment, the investigators raised that it was constructed by architects and engineers, and that the construction was one year.
Output 3	
Unit	71 47 11 45 64 74 27 89 59 33 68 9 29 28 23 73 16 99 82 11 64 53 73 74 27 78 14 76 53 9 74 2 50 42 44 18 2 31 23 73 16 99 82 73 62 74 27 31 59 33 68 44 80 85 19 81 83 84 57 96 55 39 67 54 57 93 82 11 45 64 53 44 80 18 66 77 87 9 43 6 49 92 31 23 73 74 27 89 59 94 32 64 85 5 30 65 6 49 7 69 16 50 18 29 28 92 82 87 9 85 53 60 70 14 76 62 29 28 92 89 23 44 80 18 6 49 92 52 24 68 43 74 2 31 23 53 1 66 47 24 13 58 37 24 46 30 74 2 89 23 62 74 27 31 59 87 91 43 74 2 6 49 7 23 44 80 85 87 38 44 80 18 21 95 23 44 80 26 11 0 79 29 28 49 41 84 57 96 55 39 67 54 57 86 38 44 80 18 82 87 9 74 2 82 73 70 52 25 91 17 68 80 18 85 23 73 16 66 77 87 38 44 18 21 95 60 90 35 11 64 74 27 31 59 94 32 74 27 89 23 53 44 80 85 73 11 98 45 0 79 81 83 20
Speech (En)	
ASR Output	Because of the equipment of the tunnel. The investigator assumes that it was constructed by architects and engineers and that the round eventually taken a year.
Output 4	
Unit	71 82 11 45 64 53 44 80 18 66 77 87 9 43 6 49 92 31 23 73 74 27 89 59 94 32 64 65 6 95 23 44 80 85 73 16 99 82 73 62 74 27 31 59 33 68 80 85 19 81 83 84 57 96 55 39 67 54 57 93 86 53 62 29 28 92 82 11 45 64 53 44 80 18 66 77 87 9 43 6 49 92 31 23 73 74 27 89 59 94 32 64 65 6 95 23 42 80 81 83 84 57 96 55 39 67 54 57 93 75 9 29 28 66 47 53 44 80 18 66 31 23 62 29 28 7 87 24 13 58 44 80 18 66 47 24 13 58 37 24 46 30 74 2 89 23 62 74 2 27 31 59 87 91 9 43 74 2 6 49 7 23 44 80 85 53 44 80 18 21 95 23 44 80 26 11 0 79 65 6 49 41 84 57 96 55 39 67 54 57 86 38 44 80 18 82 87 9 74 2 82 73 74 27 89 23 44 80 18 6 49 92 21 52 25 24 68 43 74 2 65 95 23 44 80 85 73 16 99 82 73 74 27 89 23 44 80 18 6 49 92 52 24 68 43 74 2 65 95 23 44 80 85 73 16 77 73 11 98 45 64 0 79 81 83 20
Speech (En)	
ASR Output	The investigation of the tunnel is the investigation has been designed by architects and engineers and that the construction of the construction of a year.

Table 10: Examples of synthetic En→De ST data generated by ***Diverse*** **BT4ST** method.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The data and pre-trained model we used are only for research purposes without violating the license.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Ethics Statement

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We use MuST-C and WMT datasets without modification, which are all widely used public datasets. We don't think this needs much discussion.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4

C ☑ Did you run computational experiments?

Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section4, Section 5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4, Section 5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.