Textless Speech-to-Speech Translation on Real Data
Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Juan Pino, Jiatao Gu, Wei-Ning Hsu
Meta AI

Abstract
We present a textless speech-to-speech translation (S2ST) system that can translate speech from one language into another language and can be built without the need of any text data. Different from existing work in the literature, we tackle the challenge in modeling multi-speaker target speech and train the systems with real-world S2ST data. The key to our approach is a self-supervised unit-based speech normalization technique, which fine-tunes a pre-trained speech encoder with paired audios from multiple speakers and a single reference speaker to reduce the variations due to accents, while preserving the lexical content. With only 10 minutes of paired data for speech normalization, we obtain on average 3.2 BLEU gain when training the S2ST model on the VoxPopuli S2ST dataset, compared to a baseline trained on un-normalized speech target. We also incorporate automatically mined S2ST data and show an additional 2.0 BLEU gain. To our knowledge, we are the first to establish a textless S2ST technique that can be trained with real-world data and works for multiple language pairs.

1 Introduction
Speech-to-speech translation (S2ST) technology can help bridge the communication gap between people speaking different languages and enable a natural way of interaction. Conventional S2ST systems (Lavie et al., 1997; Nakamura et al., 2006) usually rely on a cascaded approach by first translating speech in one language to text in another language, either through automatic speech recognition (ASR) followed by machine translation (MT), or an end-to-end speech-to-text translation (S2T) model (Bérard et al., 2016), and then applying text-to-speech (TTS) synthesis to generate the speech output in the target language.

On the other hand, researchers have started exploring direct S2ST (Jia et al., 2019, 2021; Tjandra et al., 2019; Zhang et al., 2020; Kano et al., 2021; Lee et al., 2021), which aims at translating speech in the source language to speech in the target language without the need of text generation as an intermediate step. However, text transcriptions or phoneme annotations of the speech data is often still needed during model training for multitask learning (Jia et al., 2019; Lee et al., 2021) or for learning a decoder that generates intermediate representations (Jia et al., 2021; Kano et al., 2021) to facilitate the generation of speech output.

More than 40% of the languages in the world are without text writing systems. Direct S2ST can serve as the solution to break the communication barriers and enable access to a wider range of knowledge for the speakers of these unwritten languages, while very limited work exist to tackle the challenge of training direct S2ST systems without the use of any text data (Tjandra et al., 2019; Zhang et al., 2020). Moreover, due to the lack of parallel S2ST training data, previous work on direct S2ST mainly rely on TTS to generate synthetic target speech, which is usually clean and high-quality, for model training.

The recent release of the large-scale S2ST data from VoxPopuli (Wang et al., 2021c) has opened up the possibility of conducting S2ST research on real data. In addition, Duquenne et al. (2021) have demonstrated the first proof of concept of direct S2S mining without using ASR or MT systems. The approach may potentially mitigate the data scarcity issue, but the authors had not evaluated the usefulness of such data for S2ST frameworks.

Most recently, Lee et al. (2021) have proposed to take advantage of self-supervised discrete representations (Lakhotia et al., 2021), or discrete units, learned from unlabeled speech data as the target for building a direct S2ST model. The direct S2ST system consists of a speech-to-unit translation (S2UT) model followed by a unit-based vocoder for con-

1https://www.ethnologue.com/
verting discrete representations to speech. Experiments conducted with synthetic target speech data have shown significant improvement for translation between unwritten languages.

In this work, we extend the system in (Lee et al., 2021) and study textless S2ST, i.e. training an S2ST system without the use of any text or phoneme data. We conduct experiments on real S2ST datasets, including VoxPopuli (Wang et al., 2021c) and automatically mined S2ST data (Duquenne et al., 2021). To tackle the challenge of modeling real target speech where there are multiple speakers with various accents, speaking styles and recording conditions, etc., we propose a speech normalization technique that fine-tunes a self-supervised pre-trained model for speech with a limited amount of parallel multiple-to-single speaker speech. Experiments on four language pairs show that when trained with the normalized target speech obtained from a speech normalizer trained with 10-min parallel data, the performance of a textless S2ST model can be improved by 3.2 BLEU points on average compared with a baseline with un-normalized target speech.

The main contributions of this work include:

- We propose a speech normalization technique based on self-supervised discrete units that can remove the variation of paralinguistic aspect of speech from multiple speakers without changing the lexical content. We apply the technique on the target speech of real S2ST data and verify its effectiveness in the context of textless S2ST.

- We empirically demonstrate that with the speech normalization technique, we can further improve a textless S2ST system’s performance by augmenting supervised S2ST data with directly mined S2ST data, demonstrating the usefulness of the latter.

- To the best of our knowledge, we are the first to establish an S2ST technique that can be trained with real-world data and does not require any text data during training, and the technique works for multiple language pairs.

The paper is organized as follows. In the next section, we first summarize related work. We introduce our approach in detail in Section 3. Following detailed description of our experimental setup in Section 4, we present the experimental results and analyses in Section 5. The paper concludes with a discussion of future research directions.

2 Related work

**Direct S2ST** Jia et al. (2019, 2021) propose a sequence-to-sequence model with a speech encoder and a spectrogram decoder that can directly translate speech from one language into another language without generating text translation first. The model can be trained end-to-end, while phoneme data is required in model training. On the other hand, Tjandra et al. (2019); Zhang et al. (2020) build direct S2ST systems for languages without text writing systems by adopting Vector-Quantized Variational Auto-Encoder (VQ-VAE) (van den Oord et al., 2017) to convert target speech into discrete codes and learn a speech-to-code translation model. The systems are trained either without any text data (Tjandra et al., 2019) or with text data from other languages (Zhang et al., 2020). Most recently, Lee et al. (2021) propose a direct S2ST system that predicts self-supervised discrete representations of the target speech and can be trained either with or without text data. In this work, we follow the design in (Lee et al., 2021) and focus on the textless setup.

**S2ST data** VoxPopuli (Wang et al., 2021d) provides the largest S2ST data to-date, totaling 17.3k hours of aligned speech between 15 source languages and 15 target languages from European parliament plenary sessions and the simultaneous interpretations. Zanon Boito et al. (2020) create MaSS by aligning multilingual speech datasets that have the same source content, resulting in a total of 23 hours of S2ST data across eight languages. Overall, there exists few S2ST corpora as the process of creating aligned speech data between two languages usually requires transcribing both data (Tohyama et al., 2004; Bendazzoli et al., 2005) or access to high-quality ASR models (Wang et al., 2021d). On the other hand, Duquenne et al. (2021) propose to directly mine from unlabeled speech data to create S2ST data based on cosine similarity between fixed-size speech representations, resulting in 1.4k hours of mined S2ST data for six language pairs.

Current existing direct S2ST systems are mostly trained with synthetic target speech (Jia et al., 2019; Tjandra et al., 2019; Zhang et al., 2020; Lee et al., 2021; Jia et al., 2021), while the usefulness of the S2ST datasets is often showcased indirectly through a speech retrieval task (Zanon Boito et al., 2020) or human evaluation of the data quality (Duquenne et al., 2021). In this work, we de-
develop an S2ST system that can be trained on real target speech to mitigate the discrepancy between the S2ST system and corpus development process.

**Speech normalization** Speech normalization is widely adopted when building speech synthesis systems to reduce the variation of factors not specified at the input. One manual approach is to use curated clean data from a single speaker with minimal non-textual variation (Wang et al., 2017; Shen et al., 2018; Ren et al., 2019; Ito and Johnson, 2017). However, it is unlikely to collect huge amounts of such data that is needed in building S2ST models.

For automatic methods, silence removal with a voice activity detection (VAD) module is a fundamental approach (Gibiansky et al., 2017; Hayashi et al., 2020; Wang et al., 2021a). Speech enhancement, which removes the acoustic condition variation, is essential when building TTS models with noisy data (Botinhao et al., 2016; Adiga et al., 2019). Speaker normalization through voice conversion, which maps target speech into the same speaker as the source speech in the context of S2ST (Jia et al., 2021), could also be considered as another speech normalization method. In this work, we propose a novel speech normalization method based on self-supervised discrete units. The normalization process maps speech with diverse variation to units with little non-textual variation, which can in principle covers normalization of all the aforementioned aspects.

3 System

We follow the setup in (Lee et al., 2021) to build a sequence-to-sequence speech-to-unit translation (S2UT) model for S2ST and use HuBERT (Hsu et al., 2021) as the self-supervised model for learning discrete representations. Below we describe the proposed speech normalization technique, which is applied on the real target speech, and how to build an S2UT system with the normalized units.

3.1 Self-supervised Unit-based Speech Normalization

**HuBERT and discrete units** The hidden-unit BERT (HuBERT) approach proposed in (Hsu et al., 2021) is an iterative process for self-supervised speech representation learning. In each iteration, the model first performs K-means clustering to generate discrete target labels from its intermediate representations (or the Mel-frequency cepstral coefficient (MFCC) features for the first iteration), and then utilizes the labels to compute a BERT-like (Devlin et al., 2019) loss for model training. After the last iteration of model training, K-means clustering is again performed, and the learned K cluster centroids are used to transform input audio into a sequence of discrete cluster indices. In the end, an utterance \( y \) is represented as \( [z_1, z_2, \ldots, z_T] \), \( z_i \in \{0, 1, \ldots, K - 1\}, \forall 1 \leq i \leq T \), where \( T \) is the number of frames. We refer to the discrete units obtained from the K-means cluster assignment with a pre-trained HuBERT as \textit{orig-unit}.

**Unit-based speech normalization** While discrete units have shown to be able to disentangle lexical content from prosodic or speaker information from the speech generation perspective (Polyak et al., 2021), we observe that \textit{orig-unit} sequences from audios of different speakers speaking the same content can still be quite different due to various accent and other residual variations such as silence and recording conditions. On the other hand, there is less variation in \textit{orig-unit} extracted from...
speech from the same speaker. Figure 1 shows an example of two speakers speaking the same word.

Following the success of self-supervised pre-training and Connectionist Temporal Classification (CTC) finetuning for ASR (Graves et al., 2006; Baevski et al., 2019), we propose a discrete unit-based speech normalization technique by performing CTC finetuning with a self-supervised pre-trained speech encoder using multi-speaker speech as input and discrete units from a single reference speaker as the target. Figure 2 illustrates the process of building the speech normalizer.

First, pairs of audio samples that have the same content but spoken by a random speaker and a reference speaker, respectively, are required. Then, we convert the reference speaker speech into orig-unit with the pre-trained HuBERT model followed by K-means clustering. We further reduce the full orig-unit sequence by removing repeating units (Lakhotia et al., 2021; Lee et al., 2021; Kharitonov et al., 2021; Kreuk et al., 2021). The resulting reduced orig-unit serves as the target in the CTC finetuning stage with the speech from the random speaker as the input.

The process can be viewed as creating “pseudo text” with speech from a single reference speaker, and then training an ASR model with the “pseudo text” as target. The resulting speech normalizer is a discrete unit extractor that converts the input speech to “pseudo text”-like units with CTC decoding, which has less variation due to speaker, accent or prosodic changes. We refer to the unit sequences from the speech normalizer as norm-unit.

### 3.2 Textless S2ST

Below we introduce the main components of the S2ST model (Figure 3).

**Speech encoder** The speech encoder is built by pre-pending a speech downsampling module to a stack of transformer blocks (Vaswani et al., 2017). The downsampling module consists of two 1D-convolutional layers, each with stride 2 and followed by a gated linear unit activation function, resulting in a downsampling factor of 4 (Synnaeve et al., 2019) for the log-mel filterbank input.

**Discrete unit decoder** As the target sequence is discrete, the unit decoder is a stack of transformer blocks as in MT (Vaswani et al., 2017) and is trained with cross-entropy loss with label smoothing. We train the S2UT system by directly using norm-unit sequence as the target. The setup can be viewed as the same as the “reduced” strategy in (Lee et al., 2021), as the speech normalizer is trained on reduced orig-unit sequences as well.

**Auxiliary task** We follow the unwritten language scenario in (Lee et al., 2021) and incorporate an auto-encoding style auxiliary task to help the model converge. An additional cross-attention module and a transformer decoder are added on top of an intermediate layer of the speech encoder for the auxiliary task, which is only used during training and not in inference, and we use reduced orig-unit of the source speech as the auxiliary task target.
Table 1: Number of samples of the data used in training speech normalization models. For Es and Fr, as there is no enough data from VoxPopuli ASR dataset after filtering out the overlap with the S2ST data, we include random samples from the Common Voice 7.0 (CV) (Ardila et al., 2020) dataset (denoted as X%) to create training sets of three different sizes.

|       | duration | En | Es | Fr |
|-------|----------|----|----|----|
| train | 10 mins  | 89 | 97 | 86 |
|       | 1 hr     | 522| 612| 510|
|       | 10 hrs   | 5.1k| 6.7k| 5.9k|
| dev   | -        | 1.2k| 1.5k| 1.5k|

Table 2: Statistics of the data used in S2ST experiments. We train the S2UT models on VoxPopuli (VP) (Wang et al., 2021d) and mined S2ST data (Duquenne et al., 2021) and evaluate on Europarl-ST (EP) (Iranzo-Sánchez et al., 2020) dev and test sets. The source speech from plenary sessions before 2013 are removed from VP to avoid overlap with EP evaluation sets, resulting in different amounts of data between X-Y and Y-X language pairs. (∗: target speech is created with TTS for tracking dev loss during training.)

|       | # samples | source (hrs) | target (hrs) |
|-------|-----------|--------------|--------------|
|       | VP        | mined        |              |
| Es-En | VP        | 159k         | 532.1        | 513.1        |
|       | EP dev    | 314k         | 441.7        | 424.7        |
|       | EP test   | 1.9k         | 5.4          | 5.6*         |
|       |            | 1.8k         | 5.1          | -            |
|       | Fr-En     |              |              |
|       | VP        |              |              |
|       | mined     | 338k         | 447.1        | 469.5        |
|       | EP dev    | 1.5k         | 3.7          | 3.9*         |
|       | EP test   | 1.8k         | 4.7          | -            |
|       |            |              |              |
|       | En-Es     |              |              |
|       | VP        |              |              |
|       | mined     | 314k         | 424.7        | 441.7        |
|       | EP dev    | 1.3k         | 3.0          | 3.0*         |
|       | EP test   | 1.3k         | 2.9          | -            |
|       |            |              |              |
|       | En-Fr     |              |              |
|       | VP        |              |              |
|       | mined     | 338k         | 469.5        | 447.1        |
|       | EP dev    | 1.3k         | 3.0          | 3.0*         |
|       | EP test   | 1.2k         | 2.8          | -            |

4 Experimental Setup

We investigate S2ST for four language pairs: Spanish-English (Es-En), French-English (Fr-En), English-Spanish (En-Es), and English-French (En-Fr). In this section, we first describe the datasets used in the experiments. Then, we introduce a multilingual HuBERT (mHuBERT) model, which serves as the basis for the speech normalizer. Finally, we describe training details of the proposed framework and the baselines. All experiments are conducted using fairseq (Ott et al., 2019; Wang et al., 2020a, 2021b).

4.1 Data

mHuBERT As we focus on modeling target speech in En, Es or Fr, we train an mHuBERT model by combining data from these three languages. We use the data from the 100k subset of the VoxPopuli dataset (Wang et al., 2021c), which contains unlabeled speech data for 23 languages, and we use the 4.5k hrs of unlabeled speech for En, Es and Fr, respectively, totaling 13.5k hours.

Speech normalization We use audio samples from the VoxPopuli ASR dataset (Wang et al., 2021d) as the input multi-speaker speech, and apply text-to-unit (T2U) conversion on the text transcriptions to generate reference target units for training the speech normalizer. The T2U model is the same as a transformer MT model (Vaswani et al., 2017) and is trained with characters as input and reduced orig-unit extracted from audio samples from the TTS dataset (described later) as the target. Though the reference speech data is created synthetically in our study, we believe that collecting a maximum of 10-hr speech from a single speaker is reasonable in the real-world scenario as well (Ito and Johnson, 2017; Park and Mule, 2019).

We build training sets of three different sizes (10 minutes, 1 hour, 10 hour) for each language (Table 1). After removing the audios that exist in the VoxPopuli S2ST dataset described later, we randomly sample from the Common Voice ASR dataset (Ardila et al., 2020) for the settings where there is no enough data (Es 1-hr, Es 10-hr, Fr 10-hr). We also randomly sample 1000 audios from Common Voice dev sets and combine with the filtered VoxPopuli ASR dev sets for model development.
Table 3: Duration of the TTS datasets after VAD pre-processing.

| Language | Dataset      | Duration (hrs) |
|----------|--------------|----------------|
| En       | LJSpeech     | 22.3 0.7       |
| Es       | CSS10        | 20.8 0.2       |
| Fr       | CSS10        | 17.7 0.2       |

S2UT  We use the VoxPopuli S2ST dataset (Wang et al., 2021c) as the supervised S2ST dataset for model training. Take Es-En for example. We combine data from Es source speech to En interpretation with Es interpretation to En source speech together to increase the amount of training data. We evaluate on the dev set and test set from EuroparlST (Iranzo-Sánchez et al., 2020), since it provides text translation for BLEU score computation and is of the same domain as VoxPopuli.

In addition, we investigate incorporating S2ST data automatically mined from LibriVox (Duquenne et al., 2021). The approach is a generalization to the audio domain of distance-based bitext mining in the text domain (Schwenk et al., 2021). The underlying idea is to first learn a joint fixed-size representation for text and audio. In the resulting embedding space, sentences with similar meaning are close, independent of the modality or language. This technique was applied to mine for speech-to-speech alignments in 73k hours of LibriVox audio-books in En, and more than 1.5k hours in Fr and Es. We had early access to this S2S corpus. It will be freely available by the end of 2021.

For each alignment pair, a semantic similarity score is provided which can be used to filter subsets of all the data, focusing either on precision or recall. Table 2 summarizes the statistics of the data used in the experiments for each language pair.

TTS data  We train the unit-based HiFi-GAN vocoder using TTS data for each language. We pre-process the data by applying VAD to remove silence at both ends of the audio. No text data is required during vocoder training. However, we use audios from TTS datasets due to its high quality. In addition, we use the same TTS dataset to train the T2U model for generating reference target units in speech normalizer training and to build the cascaded baselines described in Section 4.3.

4.2 Multilingual HuBERT (mHuBERT)

We build a single mHuBERT model for all three languages using the combination of 13.5k-hr data without applying any language-dependent weights or sampling, since the amount of data is similar between all three languages. A single codebook is used for all three languages, and no language ID information is required during pre-training. The mHuBERT model is pre-trained for three iterations following the recommendation from Hsu et al. (2021) and Lakhota et al. (2021). Each iteration is optimized for 400k steps, and weights are randomly initialized when starting a new iteration.

We examine features extracted from different layers of the third-iteration mHuBERT model, experiment with different $K$ for clustering and find that $K = 1000$ with features from the 11-th layer of the model work the best for our experiments.

4.3 Baselines

S2UT with reduced orig-unit  First, we consider a basic setup by training the S2UT system with reduced orig-unit and without any additional information on the target speech (Lee et al., 2021).

For the second baseline system, we incorporate target speaker information in the model by concatenating a d-vector speaker embedding (Variani et al., 2014) to each frame of the speech encoder output and appending a linear layer to map the concatenated feature vectors to the same dimension as the original encoder output (Jia et al., 2019). The 256-dimensional speaker embedding, which remains fixed during the S2UT model training, is extracted from a speaker verification model pre-trained on VoxCeleb2 (Chung et al., 2018). During inference, we use the speaker embedding averaged from all audios from the TTS dataset of the target language.

S2T+TTS  We transcribe all the S2ST data with open-sourced ASR models (Section 4.5) and train a S2T+TTS system for each language pair. We build 2000 unigram subword units (Kudo, 2018) from the ASR decoded text as the target. For TTS, we explore two approaches: (1) transformer TTS (Li et al., 2019), and (2) text-to-unit (T2U). The transformer TTS model has a text encoder, a spectrogram decoder and a HiFi-GAN vocoder (Kong et al., 2020). The T2U model is the same model used in preparing reference units for speech normalizer training (Section 4.1), and we apply the same unit-based vocoder for the S2UT model for unit-to-speech conversion. Both TTS and T2U are

\footnote{https://librivox.org/api/info}

\footnote{https://github.com/facebookresearch/LASER}
trained with characters as input.

4.4 Textless S2ST training

Speech normalization We finetune the mHu-BERT model for generating norm-unit for En, Es and Fr speech separately, resulting in three language-dependent speech normalizers. We perform CTC finetuning for 25k updates with the transformer parameters fixed for the first 10k steps. We use Adam with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-8}$, and 8k warm-up steps and then exponentially decay the learning rate. We tune the learning rate and masking probabilities on the dev sets and select the best model based on unit error rate (UER) between the model prediction and the reference target units.

S2UT We follow the same model architecture and training procedure in (Lee et al., 2021), except having a larger speech encoder and unit decoder with embedding size 512 and 8 attention heads. We train the models for 600k steps for VoxPopuli S2ST data, and 800k steps for the combination of VoxPopuli and mined data. The model with the best BLEU on the dev set is used for evaluation. All S2UT systems, including the baselines and the system with the proposed norm-unit, are trained with an auxiliary task with reduced orig-unit of the source speech as target and a loss weight of 8.0.

Unit-based vocoder We train one vocoder for each language, respectively. All vocoders are trained with orig-unit sequences as input, since they contain the duration information of natural speech for each unit. We follow the same training procedure as in (Polyak et al., 2021) and train for 500k updates, and the weight on the MSE loss for duration prediction is set to 1.0. The same vocoder is used for generating speech from either orig-unit or norm-unit, as they originate from the same K-means clustering process.

4.5 Evaluation

We use open-sourced ASR models to decode all systems’ speech output. As the ASR output is in lowercase and without digits and punctuation except apostrophes, we normalize the reference text by mapping numbers to spoken forms and re-moving punctuation before computing BLEU using SacreBLEU (Post, 2018).

5 Results

5.1 Textless S2ST

S2ST with supervised data Table 4 summarizes the ASR BLEU scores from systems trained with VoxPopuli S2ST data. We also list the results from applying TTS on the ground truth reference text (8, 9) to demonstrate the impact from ASR errors and potentially low quality speech output on the BLEU score.

First, compared with the basic setup, the baseline with target speaker embedding incorporated can give a 1.2-3 BLEU improvement on three language pairs (1 vs. 2), indicating that there exists variations in orig-unit sequences which are hard to model without extra information from the target speech signals. However, with only 10 minutes of paired multiple-to-single speaker speech data, we obtain norm-unit that helps improve S2UT model performance further by 1.1-1.9 BLEU (2 vs. 3). The translation quality improves as we further increase the amount of parallel data used in training the speech normalizer. In the end, with 10 hours of finetuning data, we obtain an average 4.9 BLEU gain from the four language pairs compared to the basic setup (1 vs. 5). Surprisingly, we see that the S2UT models perform similarly on three language pairs when trained with norm-unit decoded from a speech normalizer trained with 1 hour or 10 hours of data (4 vs. 5).

On the other hand, compared with S2T+TTS systems that uses extra ASR models for converting speech to text for training the translation model (6, 7), our best textless S2ST systems (5) can perform similarly to text-based systems, showing that the approach has the potential to be adopted for building direct S2ST systems for written languages as well to enable faster inference for S2ST.

S2ST with supervised data and mined data

Next, we combine VoxPopuli and the mined S2ST data for model training, and the results are summarized in Table 5. We apply speech normalization trained with 1-hr data when training with mined data, as it provides similar translation performance as a speech normalizer trained with 10-hr data based on previous findings from VoxPopuli-only experiments.

On the Europarl-ST test set, we see consistent
### Table 4: BLEU scores (↑) from systems trained with VoxPopuli S2ST data (Wang et al., 2021d) and evaluated on Europarl-ST (Iranzo-Sánchez et al., 2020) dev and test sets. The best results from S2UT w/ norm-unit are highlighted in bold. (tgt spkemb: target speaker embedding, SN: speech normalization, gt: ground truth, tf: transformer)

| ID | S2UT w/ orig-unit | S2UT w/ norm-unit | S2T + tf TTS | S2T + T2U | gt + tf TTS | gt + T2U |
|----|-------------------|-------------------|--------------|------------|-------------|---------|
| 1  |                   |                   |              |            |             |         |
| 2  |                   |                   |              |            |             |         |
| 3  |                   |                   |              |            |             |         |
| 4  |                   |                   |              |            |             |         |
| 5  |                   |                   |              |            |             |         |
| 6  |                   |                   |              |            |             |         |
| 7  |                   |                   |              |            |             |         |
| 8  |                   |                   |              |            |             |         |
| 9  |                   |                   |              |            |             |         |

Table 5: BLEU scores (↑) from systems trained with the combination of VoxPopuli S2ST data (VP) (Wang et al., 2021d) and mined S2ST data (Duquenne et al., 2021) and evaluated on Europarl-ST (EP) (Iranzo-Sánchez et al., 2020) and CoVoST 2 (CVST) (Wang et al., 2020b) test sets. The state-of-the-art S2T model in (Wang et al., 2021d) is trained on more than 500 hrs of S2T data. The best results from S2UT w/ VP+mined data are highlighted in bold. (tgt spkemb: target speaker embedding, SN: speech normalization, gt: ground truth, tf: transformer)

| ID | S2UT w/ norm-unit | S2T + tf TTS | S2T + T2U | gt + tf TTS | gt + T2U |
|----|-------------------|--------------|------------|-------------|---------|
| 4  |                   |              |            |             |         |
| 10 |                   |              |            |             |         |
| 11 |                   |              |            |             |         |
| 12 |                   |              |            |             |         |
| 13 |                   |              |            |             |         |
| 14 |                   |              |            |             |         |
| 15 |                   |              |            |             |         |
| 16 |                   |              |            |             |         |

5.2 Analysis on the speech normalizer

We analyze norm-unit to understand how the speech normalization process helps improve S2UT performance. First, to verify that the process preserves the original lexical content of the speech, we perform a speech resynthesis study as in (Polyak et al., 2021) by applying ASR on the vocoder output with different versions of discrete units as input and computing word error rate (WER) between the ASR decoded text and ground truth text tran-

trend across the S2UT models trained with norm-unit and the two baselines with orig-unit, where the proposed approach gives on average 3.9 BLEU improvement compared to the basic setup (10 vs. 12), indicating that the speech normalizer trained on VoxPopuli and Common Voice data can also be applied to audios from different domains, e.g. LibriVox. Compared with results from the VoxPopuli only experiments, the addition of mined data with the proposed speech normalization technique achieves an average of 2.0 BLEU gain over four language directions (4 vs. 12).

We also examine model performance on the CoVoST 2 test set (Wang et al., 2020b) and see even larger improvements brought by mined data (10, 11, 12 vs. 4). One possible reason for larger improvements in CoVoST 2 is that the LibriVox domain where the mined data is collected is more similar to the domain of CoVoST 2 than that of Europarl-ST. With target speaker embedding, mined data improves S2ST by 7.1 BLEU averaged over Es-En and Fr-En directions (4 vs. 11). S2UT with norm-unit does not perform as well, and one explanation is that we select the best model based on the Europarl-ST dev set during model training.
Table 6: Speech resynthesis results on the VoxPopuli ASR test set.

|             | En  | Es  | Fr  |
|-------------|-----|-----|-----|
| original audio | 14.2 | 15.5 | 18.5 |
| reduced orig-unit | 22.4 | 22.7 | 24.1 |
| norm-unit (10-min) | 23.5 | 25.3 | 31.7 |
| norm-unit (1-hr)  | 21.2 | 20.5 | 24.6 |
| norm-unit (10-hr) | 22.0 | 25.3 | 24.2 |

Table 7: Unit error rate (UER) between units extracted from 400 pairs of audios from the Common Voice dataset. Each pair contains two speakers reading the same text prompts.

|             | En  | Es  | Fr  |
|-------------|-----|-----|-----|
| reduced orig-unit | 74.4 | 70.6 | 73.5 |
| norm-unit (1-hr)  | 48.2 | 31.6 | 46.4 |

Next, to examine that the speech normalizer reduces variations between unit sequences across speakers, we sample 400 pairs of audios from Common Voice (Ardila et al., 2020) for En, Es and Fr, respectively. Each pair contains two speakers reading the same text prompt. Table 7 shows the unit error rate (UER) between the unit sequences extracted from the paired audios. We see that norm-unit has UER that is on average 58% of the UER of reduced orig-unit, showing that norm-unit has less variations across speakers.

5.3 Analysis of mined data

Each pair of aligned speech in mined data has an associated semantic similarity score. In experiments above, we set the score threshold as 1.06, and use all mined data with scores above it. Given the trade-off between the quality and quantity of mined data, we analyze how the S2ST performance changes with the threshold set in mined data selection. Figure 4 demonstrates BLEU scores on Europarl-ST Es-En test set from systems trained with the combination of VoxPopuli S2ST data and the mined data filtered at various thresholds. The mined data is useful at different thresholds given its gains over the model trained without mined data. As we increase the threshold from 1.06 to 1.07, the performance drops due to less training data.

6 Conclusion

We present a textless S2ST system that can be trained with real target speech data. Key to the success is a self-supervised unit-based speech normalization process, which reduces variations in the target speech while retaining the lexical content. To achieve this, we take advantage of self-supervised discrete representations of a reference speaker speech and perform CTC finetuning with a pre-trained speech encoder. The speech normalizer can be trained with one hour of parallel speech data without the need of any human annotations and works for speech in different recording conditions. We conduct experiments on the VoxPopuli S2ST dataset and the mined speech data to empirically demonstrate its usefulness in improving S2ST system translation quality for the first time. In the future, we plan to investigate more textless approaches to improve model performance such as self-supervised pre-training. All the experiments and the ASR evaluation are conducted with public datasets or open-sourced models. We will release
code in the future.

Acknowledgements

The authors would like to thank Yossi Adi, Adam Polyak and Felix Kreuk for initial discussions on accent normalization.

References

Nagaraj Adiga, Yannis Pantazis, Vassilis Tsiaras, and Yannis Stylianou. 2019. Speech enhancement for noise-robust speech synthesis using wasserstein gan. In INTERSPEECH, pages 1821–1825.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 4218–4222.

Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. 2019. Effectiveness of self-supervised pre-training for speech recognition. arXiv preprint arXiv:1911.03912.

Claudio Bendazzoli, Annalisa Sandrelli, et al. 2005. An approach to corpus-based interpreting studies: developing epic (european parliament interpreting corpus). Proceedings of Challenges of Multidimensional Translation.

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. arXiv preprint arXiv:1612.01744.

Cassia Valentini Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. 2016. Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks. In Interspeech 2016, pages 352–356.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Hongyu Gong, and Holger Schwenk. 2021. Multimodal and multilingual embeddings for large-scale speech mining. Advances in Neural Information Processing Systems, 34.

Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep voice 2: Multi-speaker neural text-to-speech. Advances in neural information processing systems, 30.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning, pages 369–376.

Tomoki Hayashi, Ryuichi Yamamoto, Katsuki Inoue, Takenori Yoshimura, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Yu Zhang, and Xu Tan. 2020. Espnet-tts: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit. In ICASSP 2020:2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 7654–7658. IEEE.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhota, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. arXiv preprint arXiv:2106.07447.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8229–8233. IEEE.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Ye Jia, Michelle Tadmor Ramanovich, Tai Remaze, and Roi Pomerantz. 2021. Translatotron 2: Robust direct speech-to-speech translation. arXiv preprint arXiv:2107.08661.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. Proc. Interspeech 2019, pages 1123–1127.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2021. Transformer-based direct speech-to-speech translation with transcoder. In 2021 IEEE Spoken Language Technology Workshop (SLT), pages 958–965. IEEE.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhota, Tu-Anh Nguyen, Morgane Rivière, Abdelrahman Mohamed, Emmanuel Dupoux, et al. 2021. Text-free prosody-aware generative spoken language modeling. arXiv preprint arXiv:2109.03264.
Jungil Kong, Jaehyoon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33.

Felix Kreuk, Adam Polyak, Jade Copet, Eugene Kharitonov, Tu-Anh Nguyen, Morgane Rivière, Weiming Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, and Yossi Adi. 2021. Textless speech emotion conversion using decomposed and discrete representations. *arXiv preprint arXiv:2111.07402*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. 2021. Generative spoken language modeling from raw audio. *arXiv preprint arXiv:2102.01192*.

Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Galavda, Torsten Zeppenfeld, and Puming Zhan. 1997. JANUS-III: Speech-to-speech translation in multiple languages. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 99–102. IEEE.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, et al. 2021. Direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2107.05604*.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6706–6713.

Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, J-S Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. 2006. The ATR multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):365–376.

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*.

Kyuhyong Park and Thomas Mulc. 2019. Css10: A collection of single speaker speech datasets for 10 languages. *arXiv preprint arXiv:1903.11269*.

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. *arXiv preprint arXiv:2104.00355*.

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.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *arXiv preprint arXiv:1905.09263*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angelo Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *ACL*, page 6490–6500.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Bj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sripam, Vitaliy Liptchinsky, and Ronan Collobert. 2019. End-to-end ASR: from supervised to semi-supervised learning with modern architectures. *arXiv preprint arXiv:1911.08460*.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2019. Speech-to-speech translation between untranscribed unknown languages. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 593–600. IEEE.

Hitomi Tohyama, Shigeki Matsubara, Koichiro Ryu, N Kawaguch, and Yasuyoshi Inagaki. 2004. Ciarr simultaneous interpretation corpus. In *Proc. Oriental C OCOSDA*.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6309–6318.

Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4052–4056. IEEE.
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

Changhan Wang, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Ann Lee, Peng-Jen Chen, Jiatao Gu, and Juan Pino. 2021a. fairseq s∗ 2: A scalable and integrable speech synthesis toolkit. arXiv preprint arXiv:2109.06912.

Changhan Wang, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Ann Lee, Peng-Jen Chen, Jiatao Gu, and Juan Pino. 2021b. fairseq s∗ 2: A scalable and integrable speech synthesis toolkit. arXiv preprint arXiv:2109.06912.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021c. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. arXiv preprint arXiv:2101.00390.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021d. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. 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 993–1003, Online. Association for Computational Linguistics.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020a. fairseq s2t: Fast speech-to-text modeling with fairseq. In Proceedings of the 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (ACL): System Demonstrations.

Changhan Wang, Anne Wu, and Juan Pino. 2020b. CoVoST 2: A massively multilingual speech-to-text translation corpus.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135.

Marcely Zanon Boito, William Havard, Mahault Garnerin, Éric Le Ferrand, and Laurent Besacier. 2020. MaSS: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 6486–6493, Marseille, France. European Language Resources Association.

Chen Zhang, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, and Tie-Yan Liu. 2020. UWSpeech: Speech to speech translation for unwritten languages. arXiv preprint arXiv:2006.07926.
A mHuBERT Training details

Table 8 lists the details for the three iterations of mHuBERT training.

| iteration | target features | K-means |
|-----------|-----------------|---------|
| 1         | MFCC            | 100     |
| 2         | 6-th layer from the first iteration | 500 |
| 3         | 9-th layer from the second iteration | 500 |

Table 8: Setup for the target labels used in mHuBERT training.

B Unit-based Vocoder

Table 9 shows the resynthesis performance of the unit-based vocoder of each language. The WER on the original audio indicates the quality of the open-sourced ASR model we use for evaluation. The WER difference between original audio and orig-unit shows the quality of the vocoder, and the difference between orig-unit and reduced orig-unit shows the further impact brought by the duration prediction module.

| WER (↓) | En | Es | Fr |
|---------|----|----|----|
| original audio | 2.0 | 8.4 | 24.0 |
| orig-unit | 2.8 | 12.0 | 29.3 |
| reduced orig-unit | 3.4 | 11.9 | 31.3 |

Table 9: WER on the TTS dev sets (LJSpeech for En, and CSS10 for Es and Fr) of the audios resynthesized from units.

C Text-to-Unit (T2U)

Table 10 lists the WER of the audios generated by the T2U model, which is used in generating the reference target units for speech normalizer training. As the T2U model is trained with reduced unit sequences as the target, during synthesis, we apply the unit-based vocoder with duration prediction. We can see that T2U with a unit-based vocoder can produce high quality audio and can serve as another option of TTS.

| WER (↓) | En | Es | Fr |
|---------|----|----|----|
| original audio | 2.0 | 8.4 | 24.0 |
| T2U | 4.2 | 9.1 | 24.4 |

Table 10: WER on the TTS dev sets (LJSpeech for En, and CSS10 for Es and Fr).