Simple and Effective Unsupervised Speech Translation

Anonymous ACL submission

Abstract

The amount of labeled data to train models for speech tasks is limited for most languages, however, the data scarcity is exacerbated for speech translation which requires labeled data covering two different languages. To address this issue, we study a simple and effective approach to build speech translation systems without labeled data by leveraging recent advances in unsupervised speech recognition, machine translation and speech synthesis, either in a pipeline approach, or to generate pseudo-labels for training end-to-end speech translation models. Furthermore, we present an unsupervised domain adaptation technique for pre-trained speech models which improves the performance of downstream unsupervised speech recognition, especially for low-resource settings. Experiments show that unsupervised speech-to-text translation outperforms the previous unsupervised state of the art by 3.2 BLEU on the Libri-Trans benchmark, on CoVoST 2, our best systems outperform the best supervised end-to-end models (without pre-training) from only two years ago by an average of 5.0 BLEU over five X-En directions. We also report competitive results on MuST-C and CVSS benchmarks.

1 Introduction

Training supervised speech systems requires large amounts of labeled data which is often not available for all but a small fraction of the over 7,000 languages spoken around the world (Lewis et al., 2022). Despite much recent effort in creating speech translation corpora (Di Gangi et al., 2019a; Wang et al., 2021b), only a few dozen language directions are covered. The lack of labeled training data is even more acute for speech translation because it requires aligned labeled data in two languages which increases the effort to create such datasets. This poses the question of whether speech translation systems can be built using less labeled data or no labeled data at all.

Recent work on unsupervised speech recognition has achieved performance that can enable useful systems using no labeled data (Yeh et al., 2019; Liu et al., 2018; Chen et al., 2019; Baevski et al., 2021; Liu et al., 2022a), enabled in large part by the advances in self-supervised speech representation learning (Schneider et al., 2019; Baevski et al., 2020). These techniques were also used to build unsupervised text-to-speech systems (Liu et al., 2022b). Similarly, unsupervised text-to-text machine translation has shown great promise for certain language directions (Conneau et al., 2018; Lample et al., 2018; Artetxe et al., 2018).

In this paper, we study a method to build end-to-end unsupervised speech-to-text and speech-to-speech translation systems trained on synthetic training data obtained by cascading existing unsupervised techniques: we first transcribe speech utterances in the source language using unsupervised speech recognition (Baevski et al., 2021; Liu et al., 2022a), then translate the resulting transcription using unsupervised machine translation (Lample et al., 2018; Artetxe et al., 2018; Liu et al., 2020), and finally synthesize the translation into a target language speech utterance using unsupervised speech synthesis (Liu et al., 2022b). We also consider applying the pipeline directly at inference time. Our approach benefits from the use of self-supervised speech models (Baevski et al., 2020; Liu et al., 2020) and to further improve performance, we present a technique to adapt existing self-supervised models to the target domain.

2 Background

Unsupervised speech recognition. Liu et al. (2018) presents some of the earliest work on unsupervised phoneme recognition and their work applies adversarial training. Wav2vec-U (Baevski et al., 2021) effectively applied self-supervised speech representations, introduced a new evaluation metric and compared to state-of-the-art super-

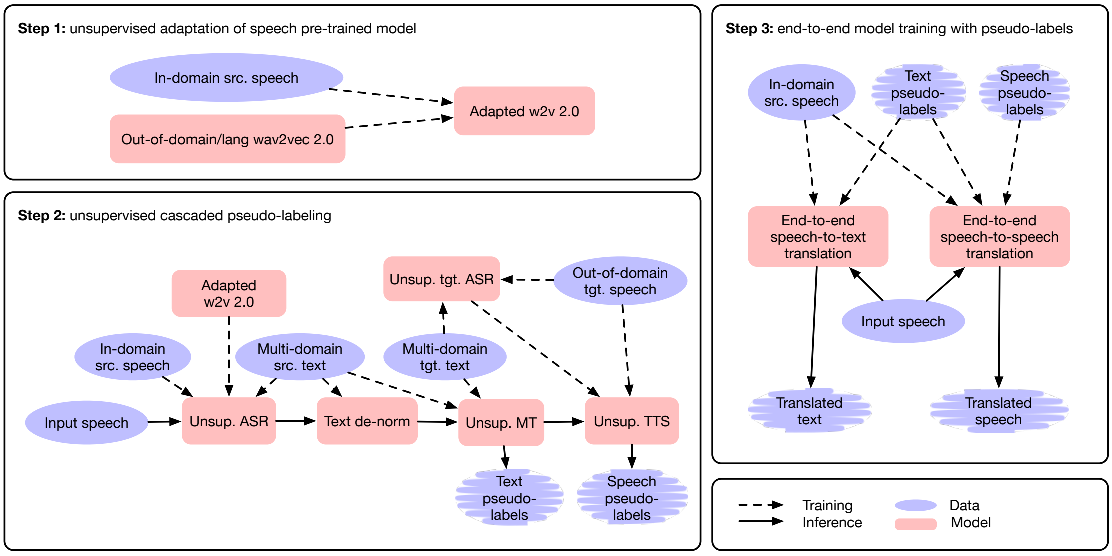

Figure 1: Overview of the proposed approach to unsupervised speech-to-text translation (S2TT) and speech-to-speech translation (S2ST). We first adapt speech pre-trained model (wav2vec 2.0) for the input language and domain of interest, and then cascade unsupervised speech recognition (ASR), unsupervised text de-normalization, unsupervised machine translation (MT) and unsupervised speech synthesis (TTS) models to produce pseudo-labels for end-to-end S2TT and S2ST model training. Our models rely only on unlabeled speech data and unpaired text data without the need of any human annotation.

vised systems trained on large amounts of labeled data. Wav2vec-U 2.0 (Liu et al., 2022a) simplifies audio-side pre-processing and improves accuracy through better architecture as well as better training objective. Lin et al. (2022) shows that out-of-domain speech pre-training or out-of-domain text data hurts the training robustness of Wav2vec-U models, especially under low-resource settings.

Unsupervised speech synthesis. Recent work has demonstrated unsupervised speech synthesis systems to be able to achieve comparable performance to supervised systems (Liu et al., 2022b; Ni et al., 2022). The systems are trained on data resulting from labeling speech audio data with unsupervised speech recognition models and training text-to-speech models on the resulting models.

Unsupervised machine translation. Lample et al. (2018) and Artetxe et al. (2018) built the first fully unsupervised machine translation (MT) systems by exploiting cross-lingual similarity of representations in multilingual sequence-to-sequence models, as well as back-translation for further refinements of the initial models. mBART (Liu et al., 2020) used a similar model architecture and training process to build unsupervised MT models, but it utilized a larger-scale multilingual text corpus (Conneau et al., 2020) and an updated noising strategy for pre-training with denoising autoencoder objective.

End-to-end speech translation. End-to-end sequence-to-sequence modeling has witnessed increased applications in speech-to-text translation (Duong et al., 2016; Bérard et al., 2016; Weiss et al., 2017; Bansal et al., 2017; Vila et al., 2018; Di Gangi et al., 2019b; Ren et al., 2020; Li et al., 2021) and speech-to-speech translation (Jia et al., 2019; Kano et al., 2021; Jia et al., 2022a). Compared to cascaded systems, end-to-end speech translation models have simpler pipeline and lower inference latency. It is shown that recent end-to-end speech-to-text translation (S2TT) models perform comparably to the cascaded counterparts on the well-established MuST-C benchmark (Bentivogli et al., 2021). Given the scarcity of speech translation corpora, there are recent attempts on building end-to-end S2TT models under low-resource settings (Bansal et al., 2018, 2019; Cheng et al., 2021) or unsupervised settings (Chung et al., 2019).

3 Methods

Figure 1 provides an overview of our proposed approach to unsupervised speech-to-text translation (S2TT) and speech-to-speech translation (S2ST). We leverage a cascade of unsupervised models to produce pseudo-labels for end-to-end S2TT and S2ST model training. To mitigate language and domain mismatch in speech pre-training (wav2vec 2.0), we finetune wav2vec 2.0 models using unlabeled in-domain speech data, and then use the adapted models to build downstream speech recognition models.

3.1 Unsupervised Cascaded Pseudo-Labeling

We cascade unsupervised speech recognition (ASR), unsupervised text de-normalization (TDN) and unsupervised machine translation (MT) models to produce pseudo-labels for S2TT. For S2ST, we additionally apply unsupervised speech synthesis (TTS) models to MT model outputs to obtain synthesized target speech.

Unsupervised ASR. We adopt wav2vec-U 2.0 (Liu et al., 2022a), which learns a mapping from self-supervised speech representations to phonemes via adversarial training and decodes phonemes into words via a weighted finite state transducer (Mohri, 1997). To improve adversarial training stability and suppress overfitting in the low-resource settings, we add Gaussian noise to the frozen input features X

$$X' = X + \mathcal{N}(0, \sigma^2)$$

as well as R-Drop regularization (Wu et al., 2021) to the logit outputs of the generator

$$\begin{aligned}\mathcal{L}_{rdp} = &\frac{1}{2}\mathcal{D}_{KL}(\mathcal{G}_1(X') \,||\, \mathcal{G}_2(X')) \\ &+ \frac{1}{2}\mathcal{D}_{KL}(\mathcal{G}_2(X') \,||\, \mathcal{G}_1(X'))\end{aligned}$$

where $\mathcal{G}_1$ and $\mathcal{G}_2$ are two generator instances with different dropout masks, and $\mathcal{D}_{KL}$ is the Kullback-Leibler (KL) divergence. We add weighted $\alpha\mathcal{L}_{rdp}$ to the wav2vec-U 2.0 objective function, where α is a hyper-parameter. After adversarial learning, we follow Baevski et al. (2021) to perform self-training with a Hidden Markov Model (HMM), and fine-tune the adapted wav2vec 2.0 model again with the CTC objective on the HMM labels. We denote the final ASR model as "w2vu2-CTC".

Unsupervised MT. We adopt mBART (Liu et al., 2020), which has a Transformer architecture (Vaswani et al., 2017) with model parameters shared across all training languages. It first obtains initial cross-lingual alignments for all languages via a denoising autoencoder objective (Vincent et al., 2010), and then refines the alignments for one specific language pair via bidirectional online back-translation on that pair of languages. We denote this model as "mBART-OBT".

Unsupervised TDN. ASR models decode normalized spoken-form texts, which have no case or punctuation (except hyphen and apostrophe). MT models, however, encode unnormalized written-form texts that have case and punctuation. This discrepancy leads to quality degradation when we cascade the two models directly for pseudo-labeling. To mitigate the mismatch, we de-normalize ASR model outputs into their unnormalized written form before feeding them into MT models. The text de-normalizer is a mBART model pre-trained with denoising autoencoder objective and fine-tuned with paired data of raw text (output) and its normalized version (input).

Unsupervised TTS. We follow Liu et al. (2022b) to produce phoneme labels for unlabeled speech data with wav2vec-U 2.0, and then train an autoregressive Transformer TTS model (Li et al., 2019) on the pseudo-labeled data. For wav2vec-U 2.0, we perform HMM-based self-training and fine-tune pre-trained wav2vec 2.0 model with HMM phoneme labels. To alleviate under-generation and over-generation issues in autoregressive models, we add R-Drop style consistency loss

$$\mathcal{L}_c = ||\mathcal{P}_1^{EOS}(X) - \mathcal{P}_2^{EOS}(X)||_1$$

to the objective function (weighted by a hyperparameter α) for better end-of-sentence (EOS) predictions, where $\mathcal{P}_1^{EOS}$ and $\mathcal{P}_2^{EOS}$ are two EOS predictions on the same input X with different dropout masks.

3.2 Unsupervised Adaptation of wav2vec 2.0 Pre-trained Models

Next, we present a method to improve performance when the domain of the data used for self-supervised pre-training differs from the downstream task domain which is often the case for low-resource languages. Specifically, we adapt out-of-domain or out-of-language wav2vec 2.0 models to the domain and language of interest by fine-tuning the entire wav2vec 2.0 models on discrete labels obtained from unlabeled in-domain data using the CTC objective (Graves et al., 2006).

To obtain discrete labels, we first collect all the wav2vec 2.0 speech representations for the training data, and perform k-means clustering to identify K clusters. Then for each utterance, we label each of its T speech representation frames $\mathbf{x}_t$ by the corresponding cluster ids $y_t \in \{1, ..., K\}$, where $t \in \{1, ..., T\}$. Finally, we merge identical consecutive y_t to obtain the final labels $y'_{t'}$, where $t' \in \{1, ..., T'\}$ and $T' \leq T$.

After unsupervised fine-tuning with discrete labels, we discard the output projection layer used for the CTC objective, and use the resulting wav2vec 2.0 trunk instead of the original wav2vec 2.0 model in the downstream tasks. The adapted models are used to extract speech representations for wav2vec-U 2.0 models, as well as pre-train encoders of the CTC models in wav2vec-U self-training.

3.3 End-to-end Model Training with Pseudo-labels

After obtaining pseudo-labels from the cascade of unsupervised models, we train end-to-end S2TT and S2TT models with supervised objectives on these pseudo-labels. For end-to-end S2TT, we adopt the model architecture in Li et al. (2021), which we denote as "w2v2-mBART". We pre-train its encoder by the unsupervised ASR model, w2vu2-CTC, and pre-train its decoder by the unsupervised MT model, mBART-OBT. For end-to-end S2ST, we adopt a variant of Translatotron 2 (Jia et al., 2022a), Spec-T2, which adds an additional encoder in between Translatotron 2's two decoders, and replace Translatotron 2's second decoder by an autoregressive Transformer decoder (Li et al., 2019). Similar to w2v2-mBART, we pre-train Spec-T2's first encoder and first decoder by w2vu2-CTC and mBART-OBT, respectively.

4 Experimental Setup

We evaluate our translation models on 5 directions into English (Fr-En, Es-En, Ru-En, Et-En and Lv-En) and 3 directions out of English (En-Es, En-Ru and En-Fr). The 5 non-English languages are from 4 different Indo-European language family sub-groups: Romance (Fr and Es), Slavic (Ru), Uralic (Et) and Baltic (Lv). For the X-En directions, we evaluate S2TT models on CoVoST 2 (Wang et al., 2021b) and evaluate S2ST models on CVSS-C (Jia et al., 2022b), which adds synthetic target speech to CoVoST 2 with a single canonical speaker voice. For the En-X directions, we only evaluate S2TT models. We use MuST-C (Di Gangi et al., 2019a) for En-Es and En-Ru, as well as Libri-Trans (Kocabiyikoglu et al., 2018) for En-Fr. For Libri-Trans, we follow Chung et al. (2019) to combine validation set and test set for evaluation.

Speech pre-training. We use robust wav2vec 2.0 (Hsu et al., 2021) for English speech, which is trained on datasets from multiple domains. For non-English speech, we adapt open-source VoxPopuli[1] (Wang et al., 2021a) models by CTC fine-tuning with 1024 discrete labels (Fr, Es and Ru) or 128 discrete labels (Et and Lv). We use monolingual VoxPopuli models for Fr and Es, and multilingual models of similar languages for Ru, Et and Lv (Slavic, Uralic and Baltic languages, respectively). We extract speech representations from the 15-th layer of the original wav2vec 2.0 models for computing discrete labels.

Speech recognition. For wav2vec-U 2.0 models, we extract speech representations from the 19-th (15-th) layer of the adapted (original) wav2vec 2.0 models. We increase the dropout on the batch normalized input features to 0.2. We set $\sigma = 0.1$ for input Gaussian noise and $\alpha = 1.0$ for R-Drop regularization. For wav2vec-U 2.0 loss weights, we set $\eta = 3$ and choose λ, γ and δ from 1.0 / 1.5, 1.5 / 2.5 and 0.3 / 0.5, respectively. For text data, we use open web crawled corpus, CC-100 (Conneau et al., 2020), which is created with little curation and has large language coverage. For supervised baselines, we fine-tune adapted wav2vec 2.0 models with CTC objective on labeled data, which we denote as "w2v2-CTC".

Machine translation. We use CC-100 (Conneau et al., 2020) to train bilingual mBART *large* models for each language pair. For bidirectional online back-translation, we use the same CC100 data and follow Liu et al. (2020) to apply 99% vocabulary masking for the first 500 updates. For supervised baselines, we fine-tune mBART models with labeled data, which we denote as "mBART-FT".

Speech synthesis. We train Transformer models (with $\mathcal{L}_c$ weight $\alpha = 1.0$) on CVSS-C target speech from the It-En direction to avoid content overlaps with the selected 5 directions. For grapheme-to-phoneme conversion, we employ g2pE (Park, 2019) for English texts and Phonemizer (Bernard, 2015) with espeak-ng[2] backend for texts in other languages. We resample audios to 22,050Hz and extract log-Mel spectrogram with FFT size 1024, window length 1024 and hop length 256.

End-to-end speech translation. For bilingual S2TT, we pre-train its encoder/decoder with w2vu2-CTC/mBART-OBT for unsupervised mod-

[1]https://github.com/facebookresearch/voxpopuli
[2]https://github.com/espeak-ng/espeak-ng

Duration (hrs)	Fr-En 264	Es-En 113	Ru-En 16	Et-En 3	Lv-En 2	Avg.
Bilingual setup						
Supervised learning + pre-training						
End-to-end (w2v2-mBART)	35.7	36.2	39.4	5.7	13.5	26.1
Supervised learning						
End-to-end (S2T Transformer; Wang et al. 2020)	26.3	23.0	14.8	0.1	2.5	13.3
Unsupervised learning						
Cascaded (ASR→TDN→MT)	24.4	23.4	27.8	8.5	7.6	18.3
End-to-end (w2v2-mBART)	24.2	24.0	25.6	3.9	2.8	16.1
Multilingual setup						
Supervised learning + pre-training						
End-to-end (w2v2-mBART), 21 langs.→En (Babu et al., 2021)	32.9	34.1	26.4	3.5	6.0	20.6
Supervised learning						
End-to-end (S2T Transformer), 21 langs.→En (Wang et al., 2020)	26.9	26.3	9.6	0.4	0.6	12.8
Unsupervised learning						
End-to-end (w2v2-mBART), {Fr,Es,Ru,Et,Lv} → En	24.3	24.0	22.8	3.1	1.0	15.0

Table 1: Bilingual and multilingual X-En **speech-to-text translation** results: test BLEU on CoVoST 2. Et-En and Lv-En are low-resource with only 3h and 2h of training data, respectively. End-to-end modeling on these two directions suffers from overfitting.

Duration (hrs)	En-Es 504	En-Ru 489	En-Fr 100
Supervised learning + pre-training			
End-to-end (w2v2-mBART)	32.4	20.0	23.1
Supervised learning			
End-to-end (S2T Transformer)	27.2†	15.3†	11.4
Unsupervised learning			
Chung et al. (2019)‡	N/A	N/A	12.2
Cascaded (ASR→TDN→MT)	22.0	10.0	15.4
End-to-end (w2v2-mBART)	23.8	9.8	15.3

Table 2: Bilingual En-X **speech-to-text translation** results: test BLEU on MuST-C (En-Es and En-Ru) and Libri-Trans (En-Fr). Our best system outperforms previous state of the art (Chung et al., 2019) on Libri-Trans by 3.7 BLEU. † Wang et al. (2020). ‡ We report the $\mathrm{S}_{\mathrm{libri}}$-$\mathrm{T}_{\mathrm{libri}}$ + $\mathrm{LM}_{\mathrm{wiki}}$ + $\mathrm{DAE}_{\mathrm{wiki}}$ configuration with the best result selected supervisedly out of 10 runs.

els, or with w2v2-CTC/mBART-FT for supervised models that leverage pre-training. To alleviate overfitting in low-resource settings (Ru-En, Et-En and Lv-En), we duplicate training examples and equip them with 2 different pseudo-labels from mBART-OBT beam search decoding. For multilingual S2TT and S2ST, we pre-train speech encoder with XLS-R 0.3B (Babu et al., 2021), and pre-train text decoder with mBART-OBT from the En-Fr direction.

Checkpoint selection and averaging. For unsupervised ASR, we adopt the unsupervised metric in Baevski et al. (2021) and average the best 2 checkpoints in the same run. For unsupervised MT and unsupervised TTS, we average the last 5 checkpoints. For end-to-end S2TT/S2ST, we sort checkpoints by losses on the pseudo-labeled validation set and average the best 5 checkpoints.

Automatic evaluation of speech outputs. Following a common practice, we first transcribe English speech outputs from the TTS or S2ST model with an open-source English ASR model[3], and then calculate WER or BLEU on the ASR transcription for automatic evaluation scores.

5 Results

5.1 X-En Speech-to-Text Translation

For X-En S2TT, we consider models trained for a single language direction (bilingual) and models covering multiple directions (multilingual). Results are reported on five translation directions into English of the CoVoST 2 benchmark and we focus on end-to-end systems but we also consider a cascade of unsupervised models. Supervised models are purely trained on labeled data without pre-training, while as supervised models with pre-training use wav2vec and mBART models, unsupervised models also use pre-trained models but no labeled data.

Table 1 shows that unsupervised end-to-end models outperform the supervised baselines by 5.0 BLEU on average over the five translation directions of the bilingual setup. The supervised models represent the best supervised end-to-end models from two years ago. These improvements are due

[3]https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec ("Wav2Vec 2.0 Large (LV-60) + Self Training")

Source duration (hrs)	Fr-En 264	Es-En 113	Ru-En 16	Et-En 3	Lv-En 2	Avg.
Supervised learning + pre-training						
End-to-end (Spec-T2), {Fr,Es,Ru,Et,Lv} → En	31.8	32.3	32.9	5.2	7.5	21.9
Supervised learning						
End-to-end (Spec-T2), {Fr,Es,Ru,Et,Lv} → En	27.4	27.7	25.4	4.1	2.5	17.4
Unsupervised learning						
Cascaded (ASR→TDN→MT→TTS), bilingual	21.6	21.2	25.3	7.2	7.7	16.6
End-to-end (Spec-T2), {Fr,Es,Ru,Et,Lv} → En	21.2	20.1	19.9	3.2	2.8	13.4

Table 3: Multilingual X-En **speech-to-speech translation** results: test BLEU on CVSS-C. Our multilingual model is trained on a subset of 5 directions out of the 21 available directions. Appendix A.1 presents a comparison of our supervised model to Jia et al. (2022b) in the 21-direction setting, which performs roughly similarly.

wav2vec 2.0 features	Domain	Hours	Multi-lingual	Seen lang.	Fine-tuning	Fr 264h	Es 113h	Ru 16h	Et 3h	Lv 2h
VoxPopuli (Wang et al., 2021a)	out	21K-89K	*	*	none	26.7	21.4	> 60	> 60	> 60
					unsup.	21.4	18.3	25.6	22.4	27.8
XLS-R (Babu et al., 2021)	in+out	436K	✓	✓	none	26.1	21.9	32.8	> 60	> 60
					unsup.	23.4	19.0	28.3	26.4	> 60
Robust wav2vec 2.0 (Hsu et al., 2021)	out	63K			none	> 60	29.3	> 60	> 60	> 60
					unsup.	31.5	22.7	35.2	35.1	> 60

Table 4: Different wav2vec 2.0 features for non-English unsupervised ASR (wav2vec-U 2.0) training: validation PER on CoVoST 2 with Viterbi decoding. All models use the wav2vec 2.0 *large* configuration. We unsupervisedly finetune wav2vec 2.0 models to the language and domain of interest. "$*$": Monolingual models for Fr and Es; multilingual models of similar languages for Ru, Et and Lv (trained on the Slavic, Uralic and Baltic languages in VoxPopuli, respectively).

to advances in unsupervised modeling as well as self-supervised pre-training. The supervised models with pre-training perform generally far above the unsupervised models and shows that there is potential to improve unsupervised speech translation in the future.

The cascaded unsupervised setup performs better than the end-to-end approach for directions with little synthetic training data such as Ru-En, Et-En and Lv-En. This is because end-to-end models are trained on datasets comprising as little as two hours of synthetic speech translation data on which they overfit. Cascaded unsupervised models do not suffer under this issue because they exploit more text for unsupervised machine translation (Table 7).

Supervised learning with pre-training for the bilingual setup performs better than the multilingual setup because only a single translation direction needs to be modeled and because the mBART model was pre-trained on 50 languages while as only a single language is being used in the X-En setup.

5.2 En-X Speech-to-Text Translation

For bilingual En-X S2TT, we compare our unsupervised models to the previous state of the art (Chung et al., 2019) on Libri-Trans (En-Fr) and we also evaluate them on the MuST-C benchmark for En-Es and En-Ru directions. Table 2 shows the test BLEU of our models and the baselines on both benchmarks. On Libri-Trans, our best system outperforms the previous state of the art, an alignment-based cascaded system, by 3.2 BLEU (Chung et al., 2019). On MuST-C, our models also achieve competitive results in this high-resource setting of around 500 hours of training data, with 3.4 BLEU and 5.5 BLEU behind the supervised baselines on En-Es and En-Ru, respectively.

5.3 X-En Speech-to-Speech Translation

To train a multilingual X-En speech-to-speech translation model, we combine pseudo-labeled bilingual data for multiple translation directions and use the Spec-T2 architecture, a variant of Translatotron 2. We build supervised Spec-T2 baselines with and without pre-training and evaluate on the CVSS-C benchmark. Table 3 shows that

	Fr	Es	Ru	Et	Lv	En	Avg.
Duration (hrs)	264	113	16	3	2	504	
Supervised learning + pre-training							
w2v2-CTC	15.7	7.0	7.1	11.1	5.9	6.3	8.9
Supervised learning							
Transformer†	18.3	16.0	31.4	65.7	51.8	12.1	32.6
Unsupervised learning							
w2vu2-CTC	23.2	10.3	15.7	17.6	14.8	12.7	15.7

Table 5: Speech recognition results: test WER on CoVoST 2 and MuST-C (En-Es). Semi-supervised and unsupervised models are decoded with 4-gram language model. † Wang et al. (2020).

	CVSS	Libri-Trans	MuST-C
JS Divergence	0.207	0.376	0.369
Supervised learning			
Transformer	12.8	15.0	16.8
Unsupervised learning			
Transformer	15.2	17.1	20.1

Table 6: Speech synthesis results: validation WER for re-synthesis on CVSS-C, Libri-Trans and MuST-C. To quantify training-inference time domain similarity, we follow Lin et al. (2022) to compute Jensen–Shannon divergence ("JSD") on 4-gram phoneme distributions. Low JSD suggests high similarity.

the best unsupervised system is on average only 0.8 BLEU below the supervised baseline. We believe that the unsupervised approach is less effective for speech-to-speech translation compared to speech-to-translation because of the increased error accumulation in the synthetic data creation process due to the addition of the unsupervised speech synthesis component to which we input unsupervised translation output which in turn is based on unsupervised speech recognition transcriptions. Similarly to speech-to-text translation, the cascaded unsupervised model performs better than the end to end approach and this is most prominent for low-resource directions.

5.4 Speech Pre-training

We evaluate the effectiveness of the unsupervised adaptation technique of wav2vec 2.0 models (§3.1) on the five non-English languages, which have less training data than English. We train wav2vec-U 2.0 models on CoVoST 2 with features extracted from three different wav2vec 2.0 models and their adapted versions: 1) Out-of-domain models, "VoxPopuli" (Wang et al., 2021a), that are trained with data in the same language (for Fr and Es) or similar languages (for Ru, Et and Lv) from the same language family subgroup; 2) a massively multilingual model for 128 languages, "XLS-R" (Babu et al., 2021), whose training data contains CoVoST 2; 3) a multi-domain English model, "robust wav2vec 2.0" (Hsu et al., 2021), where the target languages are unseen. We report validation PER on Viterbi predictions in Table 4. Speech pre-training on mismatched domains or languages ("VoxPopuli" and "robust wav2vec 2.0") leads to training convergence failure on three low-resource languages (Ru, Et and Lv). The two languages with the least amount of data, Et and Lv, even fail with in-domain multilingual pre-training. Unsupervised adaptation significantly improves training convergence and model performance for all the 3 scenarios of speech pre-training. In an example worst case scenario, Et-En wav2vec-U 2.0 model is successfully trained with only 3 hours of Et speech data and features from an adapted out-of-language out-of-domain wav2vec 2.0 model ("robust wav2vec 2.0").

5.5 Speech Recognition

Next, we evaluate the performance of unsupervised speech recognition in our setting. We decode our pre-trained supervised baselines ("w2v2-CTC") and unsupervised models ("w2vu2-CTC") with 4-gram language model. They are compared with previous un-pre-trained supervised baselines (Wang et al., 2020) on CoVoST 2 and MuST-C (for En), whose results (test WER) can be found in Table 5. We see that our unsupervised end-to-end models outperform un-pre-trained supervised baselines on all six languages with an average 16.9 WER reduction over the supervised one. Unsupervised ASR works best for languages with little labeled data due to the use of pre-trained features and advances in unsupervised algorithms.

5.6 Speech Synthesis

In our unsupervised setting, the target speech data does not share the same domain as the source one. This realistic setting leads to training-inference time domain mismatch on TTS models. We evaluate the effects of this mismatch by a re-synthesis task on 3 different datasets: CVSS-C (from It-En), Libri-Trans and MuST-C. We synthesize speech using validation texts and report WER on the ASR transcription of the synthesized speech. To quantize domain similarity, we follow Lin et al. (2022) to compute Jensen–Shannon divergence ("JSD") on 4-gram phoneme distributions, where low JSD sug-

	Fr-En	Es-En	Ru-En	Et-En	Lv-En	En-Es	En-Ru	En-Fr	
2.1B En text, non-En text	428M	379M	849M	46M	68M	379M	849M	428M	Avg.
Bitext	207K	79K	12K	1.8K	2.3K	259K	259K	47K	
Supervised learning + pre-training									
mBART-FT	46.7	46.0	48.4	23.3	29.6	38.7	23.1	21.5	34.6
Supervised learning									
Transformer	37.9[†]	36.3[†]	19.8[†]	0.3[†]	0.2[†]	33.8	15.8	17.9	20.3
Unsupervised learning									
mBART-OBT	40.1	43.8	48.6	19.0	25.0	38.5	22.2	22.1	32.4

Table 7: Machine translation results: test BLEU on CoVoST 2 (X-En), MuST-C (En-Es and En-Ru) and Libri-Trans (En-Fr). We finetune mBART model with bitext data for supervised learning and with unpaired pre-training data for unsupervised learning. [†] Wang et al. (2020).

	Fr-En	Es-En	Ru-En	Et-En	Lv-En	En-Es	En-Ru	En-Fr	Avg.
BLEU on raw text									
ASR→TDN→MT	24.4	23.4	27.8	8.5	7.6	22.0	10.0	15.4	17.4
Remove TDN	17.2	18.3	20.7	5.7	7.8	17.2	8.9	10.4	13.3
BLEU on normalized text (case and punctuation removed)									
ASR→TDN→MT	25.0	23.9	28.7	7.9	9.5	23.7	9.4	15.5	18.0
Remove TDN	23.1	24.1	26.9	7.2	9.4	23.1	9.4	15.1	17.3

Table 8: Effectiveness of text de-normalization in the unsupervised pipeline evaluated in terms of speech-to-text translation on CoVoST 2 (X-En), MuST-C (En-Es and En-Ru) and Libri-Trans (En-Fr). We report test BLEU on either raw text or normalized text. TDN not only recovers case and punctuation, but also leads to better translation of content.

gests high similarity. Table 6 shows the results. We see that both supervised and unsupervised models have higher WER on less similar domains (Libri-Trans and MuST-C).

5.7 Machine Translation

We evaluate our unsupervised models ("mBART-OBT") on the CoVoST 2, MuST-C and Libri-Trans benchmarks with test BLEU. For comparison, we also build supervised Transformer baselines ("Transformer") and supervised mBART baselines ("mBART-FT"). Results are shown in Table 7. We observe that our unsupervised models outperform supervised baselines by 12.1 BLEU on average over the eight considered translation directions. They are behind supervised baselines by only 2.2 BLEU on average. In contrast to supervised baselines that leverage in-domain paired data, the unsupervised models use unpaired CC100 data which is web data.

5.8 Text De-normalization

We verify the effectiveness of text de-normalization (TDN) by ablating it in the unsupervised cascaded pipeline. In Table 8, we show test BLEU calculated on either raw text (BLEU_{raw}) or normalized text ($\text{BLEU}_{\text{norm}}$) for the ablation. We see that TDN improves BLEU_{raw} greatly by 4.1 on average over all the directions. From the improvements on $\text{BLEU}_{\text{norm}}$, we conclude that TDN not only recovers case and punctuation, but also improves translation of the content.

6 Conclusion

In this paper, we present a simple and effective approach to unsupervised speech-to-text translation (S2TT) and speech-to-speech translation (S2ST). Our S2TT systems outperform the previous state of the art on Libri-Trans by 3.2 BLEU as well as the best supervised end-to-end models (without pre-training) on CoVoST 2 from only two years ago by an average of 5.0 BLEU over five translation directions into English. Our S2TT and S2ST systems also perform competitively on the MuST-C and CVSS-C benchmarks.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. Unsupervised speech recognition. *Advances in Neural Information Processing Systems*, 34:27826–27839.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proc. of NeurIPS*.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Low-resource speech-to-text translation. *Proc. Interspeech 2018*, pages 1298–1302.

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 *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 58–68.

Sameer Bansal, Herman Kamper, Adam Lopez, and Sharon Goldwater. 2017. Towards speech-to-text translation without speech recognition. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 474–479.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *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 2873–2887.

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

Mathieu Bernard. 2015. Phonemizer. https://github.com/bootphon/phonemizer.

Kuan-Yu Chen, Che-Ping Tsai, Da-Rong Liu, Hung-Yi Lee, and Lin shan Lee. 2019. Completely unsupervised speech recognition by a generative adversarial network harmonized with iteratively refined hidden markov models. In *Proc. of Interspeech*.

Yao-Fei Cheng, Hung-Shin Lee, and Hsin-Min Wang. 2021. AlloST: Low-Resource Speech Translation Without Source Transcription. In *Proc. Interspeech 2021*, pages 2252–2256.

Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. 2019. Towards unsupervised speech-to-text translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7170–7174. IEEE.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. *Proc. of ICLR*.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019a. Must-c: a multilingual speech translation corpus. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2012–2017. Association for Computational Linguistics.

Mattia A Di Gangi, Matteo Negri, and Marco Turchi. 2019b. Adapting transformer to end-to-end spoken language translation. In *INTERSPEECH 2019*, pages 1133–1137. International Speech Communication Association (ISCA).

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.

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.

Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. 2021. Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training. In *Proc. Interspeech 2021*, pages 721–725.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2022a. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10120–10134. PMLR.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022b. Cvss corpus and massively multilingual speech-to-speech translation. *arXiv preprint arXiv:2201.03713*.

Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Z. Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. In *INTERSPEECH*.

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.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

M. Paul Lewis, Gary F. Simon, and Charles D. Fennig. 2022. Ethnologue: Languages of the world, 25th edition. Online version: http://www.ethnologue.com.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6706–6713.

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.

Guan-Ting Lin, Chan-Jan Hsu, Da-Rong Liu, Hung-Yi Lee, and Yu Tsao. 2022. Analyzing the robustness of unsupervised speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8202–8206. IEEE.

Alexander H Liu, Wei-Ning Hsu, Michael Auli, and Alexei Baevski. 2022a. Towards end-to-end unsupervised speech recognition. *arXiv preprint arXiv:2204.02492*.

Alexander H Liu, Cheng-I Jeff Lai, Wei-Ning Hsu, Michael Auli, Alexei Baevskiv, and James Glass. 2022b. Simple and effective unsupervised speech synthesis. *arXiv preprint arXiv:2204.02524*.

Da-Rong Liu, Kuan-Yu Chen, Hung-Yi Lee, and Lin shan Lee. 2018. Completely unsupervised phoneme recognition by adversarially learning mapping relationships from audio embeddings. *Proc. of Interspeech*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Mehryar Mohri. 1997. Finite-state transducers in language and speech processing. *Computational linguistics*, 23(2):269–311.

Junrui Ni, Liming Wang, Heting Gao, Kaizhi Qian, Yang Zhang, Shiyu Chang, and Mark Hasegawa-Johnson. 2022. Unsupervised text-to-speech synthesis by unsupervised automatic speech recognition. *arXiv preprint arXiv:2203.15796*.

Jongseok Park, Kyubyong & Kim. 2019. g2pe. https://github.com/Kyubyong/g2p.

Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. Simulspeech: End-to-end simultaneous speech to text translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3787–3796.

Setffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proc. of Interspeech*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Laura Cross Vila, Carlos Escolano, José AR Fonollosa, and Marta R Costa-Jussa. 2018. End-to-end speech translation with the transformer. In *IberSPEECH*, pages 60–63.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12).

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021a. 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.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021b. Covost 2 and massively multilingual speech translation. In *Interspeech*, pages 2247–2251.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. *Proc. Interspeech 2017*, pages 2625–2629.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Chih-Kuan Yeh, Jianshu Chen, Chengzhu Yu, and Dong Yu. 2019. Unsupervised speech recognition via segmental empirical output distribution matching. In *Proc. of ICLR*.

A Appendix

A.1 Comparison of our CVSS-C supervised baseline to previous work

X-En direction	Fr	Es	Ru	Et	Lv	Avg.
Evaluated by a proprietary ASR						
Jia et al. (2022b)	32.4	33.4	23.2	3.2	2.8	19.0
Evaluated by an open-source ASR						
Ours	33.8	34.6	29.4	3.1	3.2	20.8

Table 9: Multilingual supervised baselines on CVSS-C for translating 21 languages into English. We report test BLEU on ASR transcription of the translated speech.

For evaluation of CVSS-C models, we use an open-source English ASR model[4] to transcribe translated speech for BLEU calculation. The previous work (Jia et al., 2022b), however, used transcription from a proprietary ASR model which we do not have access to. As a result, BLEU numbers reported for our model and the previous work are not directly comparable, but the small difference suggests that the two models perform roughly similarly.

A.2 Data Overview for Supervised Learning and Unsupervised Learning

	Fr-En	Es-En	Ru-En	Et-En	Lv-En
Supervised learning					
Src. paired speech	264	113	16	3	2
Src. paired text	207K	79K	12K	1.8K	2.3K
Tgt. paired speech	174	70	13	3	1
Tgt. paired text	207K	79K	12K	1.8K	2.3K
Unsupervised learning					
Src. speech	23K	21K	89K	43K	28K
Src. text	428M	379M	849M	46M	68M
Tgt. speech	29	29	29	29	29
Tgt. text	2.1B	2.1B	2.1B	2.1B	2.1B

	En-Es	En-Ru	En-Fr
Supervised learning			
Src. paired speech	504	489	100
Src. paired text	259K	259K	47K
Tgt. paired text	259K	259K	47K
Unsupervised learning			
Src. speech	63K	63K	63K
Src. text	2.1B	2.1B	2.1B
Tgt. text	379M	849M	428M

Table 10: Overview of the speech data (hours) and text data (sentences) used in supervised learning and unsupervised learning.

[4]https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec ("Wav2Vec 2.0 Large (LV-60) + Self Training")

Table 10 provides an overview for the speech and text data used in supervised learning and unsupervised learning.