Textless Speech-to-Speech Translation With Limited Parallel Data

Anonymous ACL submission

Abstract

Existing speech-to-speech translation (S2ST) models fall into two camps: they either leverage text as an intermediate step or require hundreds of hours of parallel speech data. Both approaches are incompatible with textless languages or language pairs with limited parallel data. We present a framework for training textless S2ST models that require just dozens of hours of parallel speech data. We first pretrain a model on large-scale monolingual speech data, finetune it with a small amount of parallel speech data (20-60 hours), and lastly train with unsupervised backtranslation objective. We train and evaluate our models for English-to-German, German-to-English and Marathi-to-English translation on three different domains (European Parliament, Common Voice, and All India Radio) with single-speaker synthesized speech. Evaluated using the ASR-BLEU metric, our models achieve reasonable performance on all three domains, with some being within 1-2 points of our higher-resourced topline.

1 Introduction

Speech-to-speech translation (S2ST) system maps input speech in the source language to output speech in the target language. In many ways, S2ST represents the "holy grail" of translation as it enables natural, real-time, spoken communication. S2ST has a rich history, from cascaded systems combining Automatic Speech Recognition (ASR), Machine Translation (MT), and Text To Speech (TTS) technologies (Nakamura et al., 2006) to recently proposed neural end-to-end systems (Lee et al., 2022a; Seamless Communication et al., 2023) that directly map from input source language speech to output target language speech. S2ST systems (Jia et al., 2019; Lee et al., 2022a,b; Jia et al., 2021; Duquenne et al., 2022; Seamless Communication et al., 2023) have benefited from model and data scaling, leveraging increasing amounts of parallel speech and/or text data across languages. Yet, this is feasible only for a fraction of the world's 7000 languages (Lewis et al., 2016); the majority of world languages have low-resource or no parallel translation data available (Haddow et al., 2022). Furthermore, thousands of languages are primarily spoken without standardized writing systems (about 3000 languages in Ethnologue (Lewis et al., 2016) have no reported writing system), necessitating textless language processing.

Recent work on textless speech translation (Lee et al., 2022b; Kim et al., 2023) requires large amounts of parallel speech data, which is expensive to collect and makes these approaches difficult to adapt for low-resource speech translation. On the other hand, other 'unsupervised S2ST' approaches (Wang et al., 2022a; Fu et al., 2023; Nachmani et al., 2023) do not need any parallel speech data at all, and instead rely on unsupervised cross-lingual learning using large amounts of monolingual speech and text datasets. However, they either train cascaded models that have intermediate text outputs or end-to-end models that use text supervision during training. As a result, they are difficult to adapt for textless languages that are spoken, have non-standard orthographies or poor ASR systems.

In this work, we adapt the unsupervised S2ST pipeline to work in a fully textless manner for the first time. We formulate fully textless S2ST as a unit-to-unit machine translation problem that requires a much more modest amount (dozens of hours) of parallel speech training data. We begin by pretraining an encoder-decoder unit language model over self-supervised speech units using monolingual speech data, followed by finetuning it for S2ST on a low-resource parallel dataset and finally performing unsupervised backtranslation to further improve performance. Figure 1 illustrates our method, comparing it to previous work. Modelling real speech data with speech unit sequences poses challenges, such as inherent unit sequence noise and ambiguity, that are orthogonal

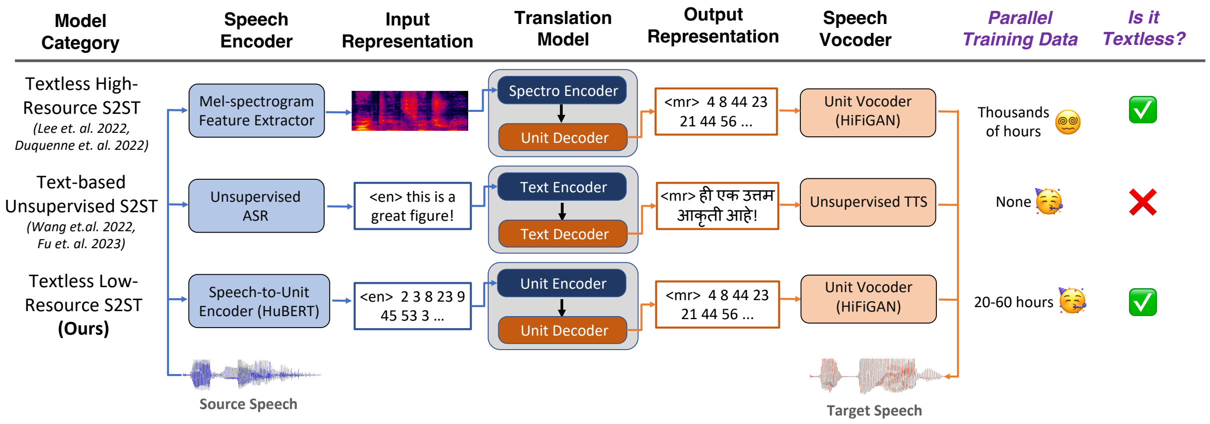


Figure 1: Overview of speech-to-speech translation systems. We compare our formulation to two relevant lines of work. We present the first textless speech-to-speech system that does not require a large parallel training dataset.

to our research questions. Thus, for simplicity, we use single-speaker synthesized speech data to train and evaluate our models, following early S2ST work (Jia et al., 2019).

We train two English ↔ German S2ST models in the European Parliament (Iranzo-Sánchez et al., 2019) and Common Voice (Ardila et al., 2020) domains and two English ↔ Marathi S2ST models in the European Parliament (Iranzo-Sánchez et al., 2019) and All India Radio (Bhogale et al., 2022) domains, and evaluate the en→de, de→en and mr→en translation directions. We find that with just 20 hrs of parallel en→de and de→en data and 60 hrs of parallel en→mr and mr→en data, our models achievable reasonable performance on all three domains, within 1-2 ASR-BLEU of our high-resource supervised topline for the European Parliament domain for the de→en and mr→en directions. We will release code and model weights at the time of publication.

2 Methods

Unsupervised S2ST (Fu et al., 2023) has tackled the problem of text-based low-resource S2ST by representing input and output speech as text sequences and unsupervisedly training a cascaded UASR-UMT-TTS pipeline. To adapt this for textless languages, we represent the input and output speech utterances as discrete, self-supervised unit sequences rather than text sequences. Instead of ASR, we use a speech-to-unit encoder (S2U) and instead of TTS, we use a unit-to-speech vocoder (U2S) largely based on prior work (Hsu et al., 2021; Polyak et al., 2021). To train the translation model, instead of text-based MT, we train a unit encoder-decoder (U2U) S2ST model using our three-step Pretrain-Finetune-Backtranslate approach illustrated in Figure 2 adapted from the unsupervised MT literature (Lample et al., 2018). We now describe each of these components below.

2.1 Speech-to-unit Encoder (S2U)

Past work (Hsu et al., 2021; Chung et al., 2021) has explored learning self-supervised discrete speech representations i.e. units. The learned units preserve much of the input signal's semantic information (Pasad et al., 2021) Critically, text transcriptions are not necessary to discover these units. It is common to train autoregressive language models (Lakhotia et al., 2021; Borsos et al., 2022) over these units, enabling NLP tasks to be performed on spoken language without needing to transcribe speech waveforms into text.

We base our speech-to-unit encoder on HuBERT (Hsu et al., 2021). As proposed by HuBERT, we train a k-means clustering model over embeddings at an intermediate layer, choosing the layer on the basis of the units' PNMI score, a phone-unit mutual information metric. We map each embedding to its nearest k-means cluster center and apply run-length encoding (Lee et al., 2022b). We train a shared English-German k-means model and a separate Marathi one. We also tried XLSR (Conneau et al., 2020) and Indic-wav2vec (Javed et al., 2021), but decided on HuBERT on the basis of its units' high PNMI score. We describe training the clustering model and the evaluation of the speech-to-unit encoder in Section 4.1.

2.2 Unit Encoder-Decoder (U2U)

We train our unit encoder-decoder S2ST model using a Pretrain-Finetune-Backtranslate approach

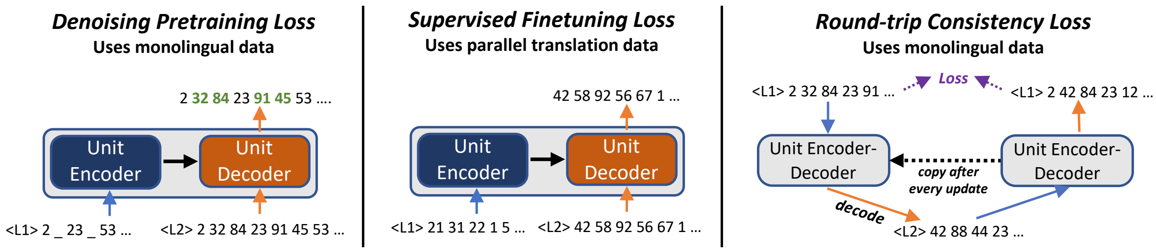


Figure 2: Training a unit-based encoder-decoder model for S2ST. The first **Pretrain** step trains on large-scale monolingual speech data using a denoising pretraining loss. The second **Finetune** step trains on low-resource parallel speech translation data using a supervised finetuning loss. The third **Backtranslate** step trains using the round-trip consistency loss (on monolingual data) and supervised finetuning replay (on parallel data).

(Figure 2). We describe our approach here and provide implementation details in Section 4.2.

Pretrain We initialize with mBART-50 (Liu et al., 2020) (a text encoder-decoder model), reinitializing the input/output embedding layers for our new unit vocab. Unit sequences do not exist in the mBART-50 text token space. However, since units can be treated as text tokens, just with a different vocabulary, we can easily adapt the training pipeline to train on unit sequences rather than text sequences. We pretrain using their denoising objective: given a unit sequence dataset $\mathcal{D}$ and a noising function $g(\cdot)$ (we use one that samples contiguous spans and masks them until a fixed ratio of tokens are masked), the decoder is trained to generate the original sequence X given encoder input $g(X)$, optimizing model weights θ as $\arg\min_\theta \sum_{X \in \mathcal{D}} -\log \Pr(X|g(X);\theta)$.

We train two bilingual unit LMs, one for English-German, and one for English-Marathi. They are trained on unit sequences, derived from monolingual speech corpora in the three languages, generated by the respective S2U encoder (shared for English-German and separate for Marathi). We train one Sentencepiece (Kudo and Richardson, 2018) BPE tokenizer per LM.

Finetune We perform supervised training on the pretrained unit LM using a small parallel S2ST corpus, where the input is a spoken utterance in the source language, and the target is a translated version spoken in the target language. During this finetuning process, we use the standard cross-entropy loss of the decoder generating the target unit sequence, when the ground truth source unit sequence is provided to the encoder.

Backtranslate Finally, we perform unsupervised backtranslation (Lample et al., 2018) on our finetuned model. We follow the standard recipes used in unsupervised text backtranslation, with minor modifications to stabilize training in the speech domain. We briefly describe this procedure which trains the model to reconstruct a unit sequence from a model-generated synthetic translation of the same unit sequence using a round-trip translation consistency loss (visualized in Figure 2). We start with the initial model $\mathcal{M}$ (the 'backward' model) and make a copy of it, calling it $\mathcal{M}'$ (the 'forward' model). Then, for every training step, we run:

1. Get two batches of utterances in the two languages, B_1 and B_2.
2. Use $\mathcal{M}'$ to translate B_1 to translations B_1', and B_2 to translations B_2'; this step is inference only and no gradient updates occur.
3. Given B_1', B_2' as input respectively, compute the decoder cross-entropy loss for the model $\mathcal{M}$ to reconstruct the original utterances B_1, B_2. Using this loss, perform a gradient update on $\mathcal{M}$'s parameters.
4. Copy the updated parameters of $\mathcal{M}$ to $\mathcal{M}'$.

The above corresponds to online backtranslation, where the 'forward' model $\mathcal{M}'$ (generating the synthetic translation) is the same as the 'backward' model $\mathcal{M}$ (used to compute the cross-entropy loss). We also explored offline backtranslation, which updates the forward model every epoch, but did not see much difference in performance. Unlike in unsupervised text backtranslation, the training was unstable in both settings. To resolve this, we mix in some supervised data (used in the finetuning step) with online backtranslation during this last stage, which stabilizes learning and shows gains.

2.3 Unit-to-speech Vocoder (U2S)

We adapt prior work (Polyak et al., 2021) on speech resynthesis from discrete units to build our unit-to-speech vocoder[1]; please refer to this work for

[1] https://github.com/facebookresearch/speech-resynthesis/tree/main/examples/speech_to_speech_translation

Model Name	Languages	Pretrain	Finetune	Backtranslate	Evaluation
$M\mathtt{de}^{\mathtt{EP}}$	de,en	VP (777h) + EP (5381h)	EP-ST (20h)	VP (777h)	EP-ST (9h) en↔de
$M\mathtt{de}^{\mathtt{CV}}$			CVSS (20h)	CV (382h)	CVSS (15h) de→en
$M\mathtt{mr}^{\mathtt{EP}}$	mr,en	VP (529h) + Shr (1000h)	S-EP-ST (60hr)	VP (529h) + Shr (1000h)	S-EP-ST (9h) mr→en
$M\mathtt{mr}^{\mathtt{Shr}}$			S-Shr-ST (60hr)		S-Shr-ST (10h) mr→en

Table 1: Model configurations. For each dataset, we mark their duration in parentheses. Abbreviations: VP = Voxpopuli, EP = Europarl, EP-ST = Europarl-ST, CV = CommonVoice, Shr = Shrutilipi, S-EP-ST = Synth-Europarl-ST, S-Shr-ST = Synth-Shrutilipi-ST.

details of their approach. Given a dataset consisting of speech waveforms and their corresponding unit sequences generated by the S2U encoder, the model trains two submodules; a duration prediction module and a HiFi-GAN (Kong et al., 2020) that converts unit sequences back to speech waveforms. We train separate U2S vocoders for each language (English, German, Marathi).

3 Experimental Setup

3.1 Datasets

Table 1 summarizes datasets used in our work. Durations reported for parallel translation datasets correspond to durations of the source speech. More dataset details are in Table 4 of Appendix A.

English-German For pretraining, we use the union of the transcribed set of Voxpopuli (Wang et al., 2021) and randomly-sampled subsets of the Europarl v3 (Koehn, 2005) train set that we call Europarl-small and Europarl-mid (statistics in Table 4 of Appendix A), collected from European Parliament recordings. For finetuning, we use two datasets: (1) randomly-sampled 20-hr (10-hr per translation direction i.e. en→de and de→en) subset of the Europarl-ST (Iranzo-Sánchez et al., 2019) train set and (2) randomly-sampled 20-hr (10-hr per translation direction) subset of the CVSS (Jia et al., 2022) train set. For the last backtranslation step, we use Voxpopuli and Common Voice 4 (Ardila et al., 2020) data for the round-trip consistency loss. Common Voice and CVSS are collected using the Mozilla Common Voice project and consist of recordings of crowd-sourced workers reading out sentences primarily derived from Wikipedia; they do not belong to the European Parliament domain. For evaluation, we use Europarl-ST (Iranzo-Sánchez et al., 2019) (for both de→en and en→de) and CVSS (Jia et al., 2022) (for de→en) test sets.

English-Marathi For pretraining, we use the union of the Shrutilipi (Bhogale et al., 2022) Marathi dataset, collected from All India Radio broadcasts and the English transcribed Voxpopuli set. We were unable to find domain-matched speech translation datasets for Marathi-English. Thus, we synthetically generate parallel datasets by translating the source language utterance to target language utterance using the Google Translate API[2]. An author of this paper, who speaks both Marathi and English, manually checked a few utterances and found the translations to be of high quality. We construct two such datasets, each consisting of train and test sets: (1) Synth-Europarl-ST: translating the English side of the English-German Europarl-ST train and test sets to Marathi. (2) Synth-Shrutilipi-ST: translating 100-hr and 10-hr subsets of the Shrutilipi dataset to English, creating a train and test set respectively.

For finetuning, we randomly sampled 60-hr (30-hr per translation direction) subsets of the train sets of these two datasets. We empirically found that we need more data in English-Marathi compared to English-German, which we hypothesize is due to greater language and domain dissimilarities. For the backtranslation step, we use the union of Voxpopuli and Shrutilipi datasets for the round-trip consistency loss. For evaluation, we use the test sets of these Synth-Europarl-ST (where Marathi is translated from English), and Synth-Shrutilipi-ST datasets, (where English is translated from Marathi). We only evaluate the mr→en translation direction for both. None of the targets in the test sets of either dataset have been seen during pretraining, making them suitable for use.

3.2 Model Configurations

Table 1 describes training and evaluation datasets for each of our four models. $M\mathtt{de}^{\mathtt{EP}}$ is trained and evaluated entirely within the European Parliament domain: it is pretrained on the union of Vox-

[2]https://cloud.google.com/translate/docs/advanced/batch-translation

populi and Europarl v3, finetuned on Europarl-ST, backtranslated with Voxpopuli, and evaluated on Europarl-ST. $M\mathtt{de}^{\mathsf{CV}}$ uses the same pretraining, but is finetuned on CVSS, backtranslated with Common Voice 4.0, and evaluated on CVSS. Common Voice and CVSS consist of crowd-sourced speech recordings reading aloud sentences primarily derived from Wikipedia, which differ from the European Parliament domain. $M\mathtt{mr}^{\mathsf{EP}}$ and $M\mathtt{mr}^{\mathsf{Shr}}$ are both pretrained and backtranslated with the union of Voxpopuli and Shrutilipi i.e. English European Parliament data and Marathi All India Radio data. $M\mathtt{mr}^{\mathsf{EP}}$ is finetuned and evaluated on the European Parliament domain using Synth-Europarl-ST while $M\mathtt{mr}^{\mathsf{Shr}}$ is finetuned and evaluated on the All India Radio domain using Synth-Shrutilipi-ST. All four models are thus finetuned and evaluated with the same dataset's train and test sets.

3.3 Generating Synthetic Speech Data

We use single-speaker synthesized speech data for both training and evaluation, following early S2ST work (Jia et al., 2019). All of our training datasets have ground truth transcripts; thus, we use TTS models to regenerate the speech from them and use the synthesized speech. We use Coqui-AI's TTS software for English and German.[3] These are VITS (Kim et al., 2021) models, trained on LJSpeech (Ito and Johnson, 2017) and Thorsten (Müller and Kreutz); each have 24 hrs of clean read speech. We use IndicTTS (Kumar et al., 2023) for Marathi; this is a FastPitch (Łańcucki, 2021) model trained on the IndicTTS Database (Baby et al., 2016) that contains around 3 hrs of clean read speech.

4 Model Implementation

4.1 Speech-to-Unit Encoder (S2U)

To choose the speech encoder model and embedding layer, we compare the unit-phoneme PNMI scores of different choices. We decide upon using HuBERT (Hsu et al., 2021) embeddings, with a shared English-German k-means model (with 200 clusters) and a standalone Marathi k-means model (with 100 clusters). We use the 6th HuBERT layer for English and German and the 8th HuBERT layer for Marathi; more details in Appendix D.

[3] We use the `en/ljspeech/vits` model for English and `de/thorsten/vits` model for German. `https://github.com/coqui-ai/TTS`)

4.2 Unit Encoder-Decoder (U2U)

Preprocessing We train one Sentencepiece BPE tokenizer per LM on speech units with a 10000-size vocab, using Voxpopuli for English-German and Voxpopuli plus Shrutilipi for English-Marathi.

Pretrain Both LMs are initialized with `mbart-large-50` (Liu et al., 2020); we reinitialize input/output embedding layers. The noising function g is similar to mBART; until 35% masked tokens, we sample a span length l from a mean-λ Poisson distribution and replace a random contiguous sequence of length l with a MASK token. For English-German model, we pretrain it in several stages with increasing task difficulty. We first train on Voxpopuli for 900k steps with lambda=2. Then, we train on Voxpopuli plus Europarl-small for 5400k steps (2700k with lambda=2 and 2700k with lambda=8). We finally train on Voxpopuli plus Europarl-mid for 2700k steps. For English-Marathi, we train on Voxpopuli plus Shrutilipi with lambda=2 for 900k steps.

For both LMs, the LR scheduler starts with 1e-7, linearly warms up to 1e-5, and then exponentially decays to 1e-6. We train on 4 GPUs. We use batches of 3125 tokens per language for English-German and 6250 tokens per language for English-Marathi, with equal token amounts per language.

Finetune We use label smoothing, dropout of 0.2 and a learning rate of 3e-5. We train for 40 epochs with a total batch size of 3748 tokens on 4 GPUs. We finetune all parameters of the models except for $M\mathtt{de}^{\mathsf{EP}}$, for which we finetune only the last 5 layers of both encoder and decoder as it shows performance gains.

Backtranslate When sampling forward translations, we use nucleus sampling (Holtzman et al., 2019) with top-p value of 0.9 and the temperature of 0.5. We use label smoothing of 0.2, learning rate of 3e-5 and train for 3 epochs with a total batch size of 3748 tokens on 4 GPUs.

4.3 Unit-to-Speech Vocoder (U2S)

A separate vocoder is trained for each language, mapping from the unit vocabulary (size 200 for English-German, size 100 for Marathi) to speech clips at 16kHz, trained on the (speech, unit sequence) pairs generated by the S2U encoder, largely following Polyak et al. (2021). We evaluate S2U+U2S jointly by computing resynthesis WER; details about model and evaluation in Appendix E.

Model	Parallel #hrs	ASR-BLEU ↑ Europarl-ST de→en	en→de	CVSS de→en
Topline models				
Text-based Parallel-Low-Resource S2ST				
ⓐ ASR → MT → TTS (Section 5.2)	20h	23.7	21.3	-
ⓑ UASR → UMT → UTTS (Fu et al., 2023)	0h	-	-	14.7
Textless Parallel-High-Resource S2ST				
ⓒ Bilingual S2S (Duquenne et al., 2022)	≈2600h	16.3	10.1	-
ⓓ Multilingual UTUT (Kim et al., 2023)	650h [4]	15.8	9.8	-
ⓔ Pretrain + Full Finetune (Ours)	110h|180h	12.0	13.4	13.6
Textless Parallel-Low-Resource S2ST				
ⓕ Pretrain + Finetune (Ours)	20h	7.8	6.8	5.8
ⓖ + Backtranslate (Ours)	20h	10.0	8.3	7.7
Ablations				
ⓗ Text mBART + Finetune	20h	1.0	0.3	-
ⓘ + Backtranslate	20h	1.3	0.4	-
ⓙ Pretrain + Backtranslate	0h	0.4	0.1	-
ⓚ Pretrain + Finetune + Backtranslate w/o replay	20h	4.3	4.0	-

Table 2: English-German S2ST evaluation using ASR-BLEU on Europarl-ST (Iranzo-Sánchez et al., 2019) and CVSS (Jia et al., 2022) test sets; higher is better. Topline models use more resources by either needing high-resource parallel data or being text-based (Section 5). The Parallel #hrs column denotes the size of parallel translation training data. In ⓗ it denotes that 110h of EP-ST data and 180h of CVSS data is used to train two separate toplines.

5 Results

5.1 Evaluation Setup

We evaluate the semantics of the speech translation (i.e. whether it preserves the input speech meaning) and leave non-content aspects like naturalness to future work. We use the ASR-BLEU metric following prior work (Lee et al., 2022a,b): the BLEU between the ASR transcript of the hypothesis speech translation and the ground truth text translation. We use SacreBLEU's default parameters. We evaluate the de→en, en→de and mr→en language directions. We opted to not evaluate the en→mr direction due to poor Marathi ASR models that resulted in excessively noisy ASR-BLEU scores. We generate translations from our models using beam search decoding with a beam size of 10. When evaluating on Europarl-ST dataset, we use wav2vec2.0 based ASR models with greedy decoding (Huggingface models `facebook/wav2vec2-large-960h-lv60-self`, `jonatasgrosman/wav2vec2-xls-r-1b-german`) used by prior S2ST work on Europarl-ST (Duquenne et al. (2022); Wang et al. (2022b) and others). When evaluating on CVSS dataset, we use a medium-sized Whisper ASR model used by prior S2ST work on CVSS (Fu et al., 2023). When evaluating Marathi-English translation, we use `facebook/wav2vec2-large-960h-lv60-self`.

5.2 Comparison Systems

We categorize S2ST models based on whether they leverage text as an intermediate step or not (text-based or textless) and how much parallel translation data they use (parallel-high-resource or parallel-low-resource). Our models belong to the textless, parallel-low-resource setting. Due to the lack of baselines in this setting, we instead contrast our models with existing **topline models** trained with more resources, which serve as upper bounds:

Text-based Parallel-Low-Resource S2ST models: ⓐ is a cascaded ASR → MT → TTS system where the MT model is text mBART finetuned on the transcripts of the 20-hr low-resource parallel speech data used by our models. We use the ASR systems used for computing ASR-BLEU (Sec-

		ASR-BLEU ↑	
		EP-ST	Shr-ST
Model	Par. #hrs	mr→en	
Topline models			
Textless Par.-High-Res.			
ⓛ Full FT (Ours)	125|176h	10.9	17.8
Textless Par.-Low-Res.			
ⓜ Pretrain + FT (Ours)	60h	8.3	9.6
ⓝ + BackT (Ours)	60h	9.2	10.0

Table 3: Marathi-English S2ST evaluation using ASR-BLEU on Synth-Europarl-ST and Synth-Shrutilipi-ST test sets; higher is better. The Par. #hrs column denotes the size of parallel training data. In ⓞ it denotes that 125h of Synth-Europarl-ST data and 176h of Synth-Shrutilipi-ST data is used to train two separate toplines.

tion 5.1) and the TTS systems used for generating our data (Section 3.3). ⓑ (Fu et al., 2023) uses a cascaded unsupervised ASR - unsupervised MT - unsupervised TTS model that is trained on large amounts of monolingual speech and text data.

Textless Parallel-High-Resource S2ST models: ⓒ is a bilingual S2ST model trained on a large, mined SpeechMatrix dataset ($\approx$ 2600 hrs of source speech for the en→de and the de→en directions combined) by (Duquenne et al., 2022). ⓓ (Kim et al., 2023) is a multilingual S2ST model trained on 650h of parallel aligned English-German Voxpopuli data, and about 12k hours of parallel aligned data in 18 other X-to-English language pairs. ⓔ and ⓛ are our pretrained unit LMs fine-tuned on more data than our parallel-low-resource models i.e. the Europarl-ST train set (110 hours), the CVSS train set (180 hours), the Synth-Europarl-ST train set (125h) and the Synth-Shrutilipi-ST train set (176h) using the same hyperparameters as our four parallel-low-resource models.

Our **Textless Parallel-Low-Resource S2ST models** consist of four models trained on different domains: $M\mathtt{de}^{\mathtt{EP}}$,$M\mathtt{de}^{\mathtt{CV}}$,$M\mathtt{mr}^{\mathtt{EP}}$ and $M\mathtt{mr}^{\mathtt{Shr}}$ as described in Section 3.2. We evaluate each model with its in-domain evaluation data, i.e., $M\mathtt{de}^{\mathtt{EP}}$ model on Europarl-ST, $M\mathtt{de}^{\mathtt{CV}}$ model on CVSS, $M\mathtt{mr}^{\mathtt{EP}}$ on Synth-Europarl-ST, and the $M\mathtt{mr}^{\mathtt{Shr}}$ model on Synth-Shrutilipi-ST. ⓕ and ⓜ report the model performance after our pretraining and finetuning steps. ⓖ and ⓝ report the model performance after performing backtranslation.

5.3 Main Results

We present results for the English-German pair in Table 2 and the English-Marathi pair in Table 3. We first observe that the text-based parallel-low-resource S2ST topline models (ⓐ-ⓑ) trained with at most 20 hrs of parallel data outperform the best textless S2ST topline models trained with far more parallel speech data (ⓒ-ⓔ). This underscores the inherent task difficulty of learning purely texless S2ST models in the speech domain, even with access to far more training data.

Next, we discuss our textless parallel-low-resource models (rows ⓕ, ⓖ, ⓜ and ⓝ). Rows ⓕ and ⓜ show that our models, given only 20 hr of parallel data (for English-German) and 60 hr of parallel data (for English-Marathi), learn S2ST models with reasonable BLEU scores which consistently improve post-backtranslation in rows ⓖ and ⓝ. Our de→en Europarl-ST and the mr→en Synth-Europarl-ST models are even within 1-2 BLEU of our supervised toplines ⓔ and ⓛ despite being trained on much less data. Another observation is regarding domain effects: the gap between our textless low-resource models and the textless high-resource toplines is smaller for European Parliament domain as compared to the Common Voice and All India Radio domains, likely due to pretrain-finetune domain mismatch (During pretraining, the models only ever see European Parliament domain English data). Finally, a qualitative analysis, based on manually looking at example outputs in Appendix G shows that our models mostly preserve the semantics of the input utterance, but often make grammatical and language modelling mistakes.

Overall, while some of our models show encouraging results in the European Parliament domain, close to supervised toplines, they underperform text-based and textless high-resource toplines.

5.4 Ablations

We perform ablations on the $M\mathtt{de}^{\mathtt{EP}}$ model.

Ablating pretraining Our LM is initialized from the text mBART checkpoint, and then trained on a unit-based denoising objective. Without this pretraining (i.e., finetuning and backtranslating with the base mBART checkpoint), as seen in rows ⓗ and ⓘ, we obtain very low ASR-BLEUs less than 2 points. These results suggest that unit LM pre-

[4] In addition to 650h of parallel German-English data, UTUT is trained on X-to-English translation data from 18 other languages, totalling $\approx$ 12000 hours of parallel data.

training is essential in order to learn good S2ST systems in parallel-low-resource settings.

Ablating finetuning We finetune the pretrained unit LM with te backtranslation round-trip consistency loss without first finetuning with parallel data. The result, ⓙ, shows that this does not work, with near-zero BLEU scores. This suggest some amount of parallel speech is necessary.

Ablating replay in backtranslation We have already seen that adding backtranslation after finetuning boosts performance by 1-2 BLEU; compare row ⓕ to ⓖ or row ⓜ to ⓝ. We replay the supervised low-resource parallel finetuning data during backtranslation to stabilize training. We ablate training with this replay by running the backtranslation step with just the round-trip consistency loss. The result, row ⓚ, shows that the performance worsens compared to the initialization of row ⓕ. With replay, however, we get the results in row ⓖ, which improve upon the initialization.

6 Related Work

6.1 Speech-to-Speech Translation (S2ST)

While cascaded S2ST models (Nakamura et al., 2006; Wahlster, 2000) with intermediate text translations have existed for a long time, end-to-end S2ST models start with Jia et al. (2019), a model that directly translates source language speech waveforms to speech waveforms in the target language. Several S2ST models (Jia et al., 2019, 2021; Lee et al., 2022a; Inaguma et al., 2022) are text-based i.e. they use textual supervision to stabilize training or improve performance, while other S2ST models (Lee et al., 2022b; Li et al., 2022; Kim et al., 2023; Zhu et al., 2023) are textless, usually by representing speech using self-supervised speech units. Most S2ST models require large training datasets of parallel speech translation data.

In order to reduce this dependency on parallel data, unsupervised S2ST systems (Wang et al., 2022b; Fu et al., 2023; Nachmani et al., 2023) that do not use any parallel data at all have been recently proposed. However, none of them are textless; they either train cascaded S2ST models (ASR→MT→TTS) using unsupervised ASR (Liu et al., 2022b), unsupervised MT (Liu et al., 2020) and unsupervised TTS (Liu et al., 2022a), or use text during training (Nachmani et al., 2023). Thus, the crucial cross-lingual translation component is learned over text tokens, limiting applicability to spoken languages. Our textless, parallel-low-resource S2ST model aims to bridge these camps.

6.2 Textless and Unit-Based NLP

While we tackle textless S2ST, textless speech processing has studied in other tasks such as spoken language modeling (Borsos et al., 2022; Lakhotia et al., 2021; Hassid et al., 2024), emotion conversion (Kreuk et al., 2021), image-speech retrieval (Harwath et al., 2016; Peng and Harwath, 2022), spoken question answering (Lin et al., 2022) and speech evaluation (Chen et al., 2022; Besacier et al., 2023). Furthermore, progress in several other speech tasks like TTS (Wang et al., 2023) that involve both speech and text has been achieved by using powerful self-supervised units (semantic units like HuBERT (Hsu et al., 2021) and acoustic units like EnCodec (Défossez et al., 2022)).

7 Conclusion

We present the first textless low-resource speech-to-speech translation system, capable of learning from dozens of hours of parallel translation data, built by pretraining, finetuning, and backtranslating a language model over self-supervised speech unit sequences rather than text. We demonstrate its efficacy on 2 language pairs (English-German and English-Marathi) across 3 different domains. While our models achieve a decent translation performance, close to supervised toplines in some cases, they still underperform models trained on far more data or models that make use of text data, implying that several challenges still remain to make these models highly performant. However, our approach holds great promise for modelling low-resource, primarily spoken languages. We hypothesize, based on similar findings for text machine translation, that scaling our approach to a larger unit LM pretrained on more data will improve performance and may unlock unsupervised textless S2ST akin to unsupervised text MT (Liu et al., 2020). Future work can investigate use of better S2U unit encoders for training better unit LMs, and training unit LMs on a larger set of languages.

Limitations

Textless S2ST models, including ours, still lag in performance behind their text-based counterparts. Therefore, while they work for all languages in theory, they are currently useful only for fully textless languages and should not be used in cases where text data is readily available. Strong open-source pretrained multilingual unit language models are as yet unavailable; as a consequence, the unit LMs we use via our own pretraining have been trained on our limited compute budget and cannot yet benefit from the scale of modern text-based LLMs. Our models are trained and evaluated on synthesized single-speaker data, following early S2ST work. They do not fully generalize to real speech data that has noise and unseen speakers.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. *Preprint*, arXiv:1912.06670.

Arun Baby, Anju Leela Thomas, NL Nishanthi, TTS Consortium, et al. 2016. Resources for indian languages. In *Proceedings of Text, Speech and Dialogue*.

Laurent Besacier, Swen Ribeiro, Olivier Galibert, and Ioan Calapodescu. 2023. A textless metric for speech-to-speech comparison. *Preprint*, arXiv:2210.11835.

Kaushal Santosh Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages. *Preprint*, arXiv:2208.12666.

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

Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. 2022. Blaser: A text-free speech-to-speech translation evaluation metric. *arXiv preprint*.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *Preprint*, arXiv:2108.06209.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *Preprint*, arXiv:2006.13979.

Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswani, Changhan Wang, Juan Pino, Benoît Sagot, and Holger Schwenk. 2022. Speechmatrix: A large-scale mined corpus of multilingual speech-to-speech translations. *Preprint*, arXiv:2211.04508.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *Preprint*, arXiv:2210.13438.

Yu-Kuan Fu, Liang-Hsuan Tseng, Jiatong Shi, Chen-An Li, Tsu-Yuan Hsu, Shinji Watanabe, and Hung yi Lee. 2023. Improving cascaded unsupervised speech translation with denoising back-translation. *Preprint*, arXiv:2305.07455.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

David F. Harwath, A. Torralba, and James R. Glass. 2016. Unsupervised learning of spoken language with visual context. In *NIPS*.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. 2024. Textually pretrained speech language models. *Preprint*, arXiv:2305.13009.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751.

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

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

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2019. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. *arXiv preprint*.

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

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2021. Towards building asr systems for the next billion users. *Preprint*, arXiv:2111.03945.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2021. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. *arXiv preprint*.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022. CVSS corpus and massively multilingual speech-to-speech translation. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 6691–6703.

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. In *Interspeech*.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *Preprint*, arXiv:2106.06103.

Minsu Kim, Jeongsoo Choi, Dahun Kim, and Yong Man Ro. 2023. Many-to-many spoken language translation via unified speech and text representation learning with unit-to-unit translation. *Preprint*, arXiv:2308.01831.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Preprint*, arXiv:2010.05646.

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

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *ArXiv*, abs/1808.06226.

Gokul Karthik Kumar, Praveen S V au2, Pratyush Kumar, Mitesh M. Khapra, and Karthik Nandakumar. 2023. Towards building text-to-speech systems for the next billion users. *Preprint*, arXiv:2211.09536.

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, and Emmanuel Dupoux. 2021. Generative spoken language modeling from raw audio. *CoRR*.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. *ArXiv*, abs/1711.00043.

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

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

M. Paul Lewis, Gary F. Simon, and Charles D. Fennig. 2016. *Ethnologue: Languages of the World, Nineteenth edition*. SIL International. Online version: http://www.ethnologue.com.

Xinjian Li, Ye Jia, and Chung-Cheng Chiu. 2022. Textless direct speech-to-speech translation with discrete speech representation. *Preprint*, arXiv:2211.00115.

Guan-Ting Lin, Yung-Sung Chuang, Ho-Lam Chung, Shu wen Yang, Hsuan-Jui Chen, Shuyan Dong, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Lin shan Lee. 2022. Dual: Discrete spoken unit adaptive learning for textless spoken question answering. *Preprint*, arXiv:2203.04911.

Alexander Liu, Cheng-I Lai, Wei-Ning Hsu, Michael Auli, Alexei Baevski, and James Glass. 2022a. Simple and effective unsupervised speech synthesis. In *INTERSPEECH*.

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

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.

Thorsten Müller and Dominik Kreutz. Thorsten-Voice.

Eliya Nachmani, Alon Levkovitch, Yifan Ding, Chulayuth Asawaroengchai, Heiga Zen, and Michelle Tadmor Ramanovich. 2023. Translatotron 3: Speech to speech translation with monolingual data. *Preprint*, arXiv:2305.17547.

S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto. 2006. The atr multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):365–376.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *ASRU*.

Puyuan Peng and David Harwath. 2022. Fast-slow transformer for visually grounding speech. In *ICASSP*.

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

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *Interspeech 2020*. ISCA.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur, Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. SeamlessM4T—Massively Multilingual & Multimodal Machine Translation. *ArXiv*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Wolfgang Wahlster. 2000. Verbmobil: Foundations of speech-to-speech translation. In *Artificial Intelligence*.

Changhan Wang, Hirofumi Inaguma, Peng-Jen Chen, Ilia Kulikov, Yun Tang, Wei-Ning Hsu, Michael Auli, and Juan Pino. 2022a. Simple and effective unsupervised speech translation. *arXiv preprint*.

Changhan Wang, Hirofumi Inaguma, Peng-Jen Chen, Ilia Kulikov, Yun Tang, Wei-Ning Hsu, Michael Auli, and Juan Pino. 2022b. Simple and Effective Unsupervised Speech Translation. *Preprint*, arXiv:2210.10191.

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

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.

Yongxin Zhu, Zhujin Gao, Xinyuan Zhou, Zhongyi Ye, and Linli Xu. 2023. Diffs2ut: A semantic preserving diffusion model for textless direct speech-to-speech translation. *Preprint*, arXiv:2310.17570.

Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. *Preprint*, arXiv:2006.06873.

A Datasets

We provide a summary of all the datasets used in this paper in Table 4.

B Compute Details

We train all our models on 4 NVIDIA A40s (often using 2 GPUs with gradient accumulation of 2, or 1 GPU with gradient accumulation of 1, which is equivalent to 4 GPUs).

C Length-wise ASR-BLEU Breakdown

In order to investigate how our model performance differs for short, medium and long test examples, for each test dataset (Europarl-ST, CVSS, Synth-EP-ST and Synth-Shruti-ST) we compute the character lengths of every target example and compute the 33rd and 66th percentiles of the length distribution. We call all examples with a length shorter than the 33rd percentile 'short', ones in between the two 'medium', and longer than the 66th percentile 'long'. We then evaluate our best models, row ⓖ (for English-German) and ⓝ (for English-Marathi) from Tables 2 and 3 on each test data

Module	Dataset	Duration	Lang
S2U Encoder: Pretraining	Librispeech	960h	en
S2U Encoder: k-means Clustering	Librispeech, MLS	48h, 48h	en, de
	Shrutilipi	100h	mr
U2U Pretraining	Voxpopuli	529h, 248h	en, de
	Europarl-small	811h, 975h	en, de
	Europarl-mid	2463h, 2918h	en, de
	Shrutilipi	1000h	mr
U2U Finetuning (Toplines)	Europarl-ST	83h,27h	en→de, de→en
	CVSS	91h,88h	en→de, de→en
	Synth-EP-ST	83h,42h	en→mr, mr→en
	Synth-Shr-ST	76h,100h	en→mr, mr→en
U2U Finetuning (Low-Resource)	Europarl-ST	10h,10h	en→de, de→en
	CVSS	10h,10h	en→de, de→en
	Synth-EP-ST	30h,30h	en→mr, mr→en
	Synth-Shr-ST	30h,30h	en→mr, mr→en
U2U Backtranslation	Voxpopuli	529h, 248h	en, de
	Common Voice	294h, 89h	en, de
	Shrutilipi	1000h	mr
U2S Vocoder	Voxpopuli	529h, 248h	en, de
	Shrutilipi	1000h	mr
Evaluation	Europarl-ST	3h,6h	en→de, de→en
	CVSS	15h	de→en
	Synth-EP-ST	9h	mr→en
	Synth-Shr-ST	10h	mr→en

Table 4: Summary of datasets used to develop our system, with datasets used by base pretrained models colored red. Datasets in the U2U Finetune and U2U Evaluation sections are parallel translation datasets, and we report duration statistics for both translation directions separately, the duration being that of the source speech.

Model	**Test Set**	**ASR-BLEU ↑**			
		short	**med**	**long**	**all**
Row g	EP-ST de→en	10.1	10.6	9.5	10.0
Row g	EP-ST en→de	9.6	9.0	7.7	8.3
Row g	CVSS de→en	6.5	8.3	7.7	7.7
Row n	S-EP-ST mr→en	10.9	10.1	8.0	9.2
Row n	S-Shr-ST mr→en	10.9	13.0	8.0	10.0

Table 5: S2ST evaluation using ASR-BLEU, broken down by test set lengths (short, medium, long) as well as the overall ASR-BLEU (all).

subset in Table 5. We see that the model does better on short/medium utterances as compared to long utterances. The performance of the long utterances is within 1 BLEU point of the overall performance.

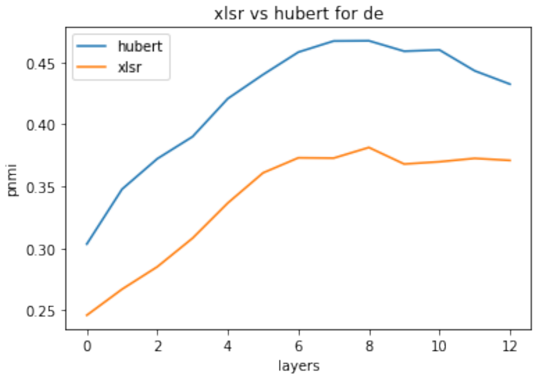
D S2U Encoder Ablations

We decide (a) which speech encoder model to use, (b) whether to learn separate per-language k-means models or a joint k-means model and (c) which encoder layer to take embeddings from, based on the average Pointwise Normalized Mutual Information (PNMI) between unit sequences and phoneme sequences extracted from the same datasets, following Hsu et al. (2021). Our best configuration uses a single Marathi k-means model and a shared English-German k-means model. We find that this works better than training three individual models or a single model, which we hypothesize is due to language similarities.

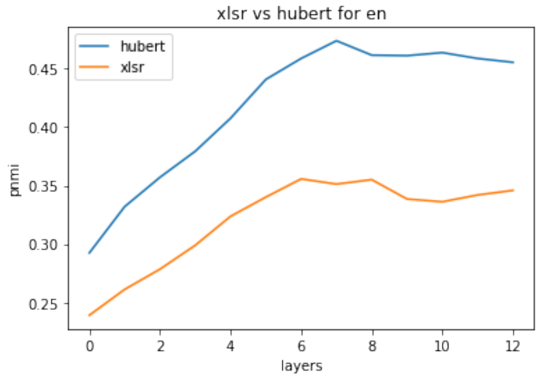
To obtain the phoneme sequences for English and German, we use English and German phonemizers from the Montreal Forced Aligner[5]. For Marathi, we use a Kaldi-based ASR model trained on Shrutilipi data. To train the k-means models, we use ≈ 50 hrs of speech data from each language, obtained from a random subset of Librispeech (Panayotov et al., 2015) for English, MLS (Pratap et al., 2020) for German, and Shrutilipi (Bhogale et al., 2022) for Marathi.

First, we describe our ablations for English-German. We experiment with different base speech models (HuBERT (Hsu et al., 2021) vs. XLSR (Conneau et al., 2020)), layer indices, num-

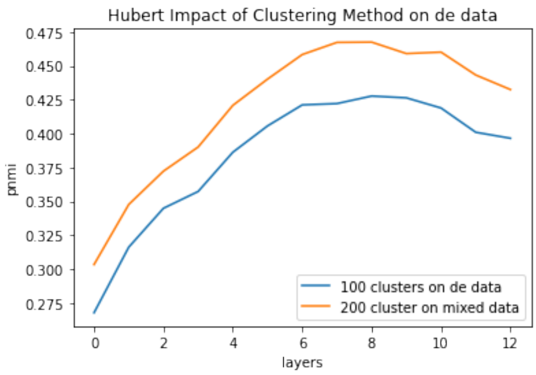
[5] https://montreal-forced-aligner.readthedocs.io/en/latest/



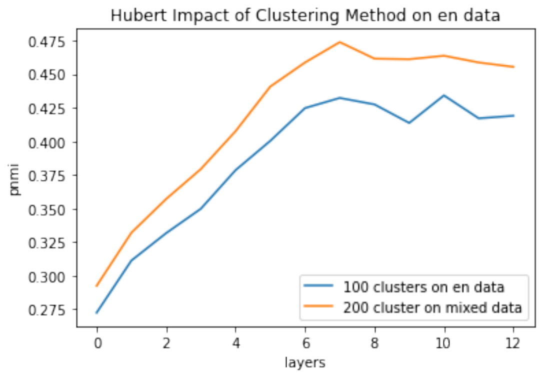
(a) HuBERT vs. XLSR evaluated on German data



(b) HuBERT vs. XLSR evaluated on English data



(c) 100 monolingual vs. 200 mixed units, evaluated on German data



(d) 100 monolingual vs. 200 mixed units, evaluated on English data

Figure 3: PNMI vs. layer index, comparing different clustering settings for English and German. Higher is better.

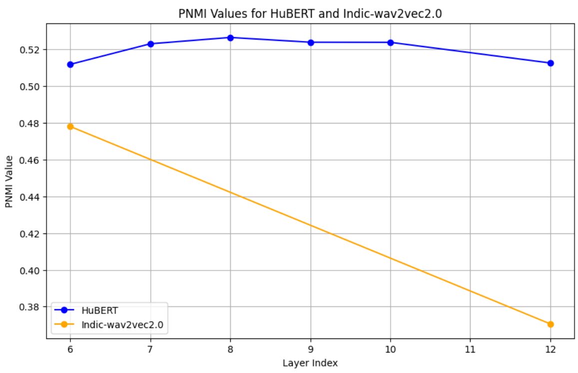


Figure 4: PNMI with HuBERT and Indic wav2vec2.0 evaluated on Shrutilipi, computed for different layer indices, for Marathi. Higher is better.

ber of clusters (100 vs. 200) and types of clusterings (one clustering for both languages jointly v.s. separate clusterings) and choose the configuration that achieves the highest Pointwise Normalized Mutual Information (PNMI). We report PNMI results for these English-German configurations in Figure 3.

For Marathi, we experiment with different base speech models (HuBERT vs Indic-wav2vec2.0 (Javed et al., 2021)) and layer indices. We fix the number of clusters at 100. We choose the configuration that achieves the highest PNMI. We report PNMI results for these Marathi configurations in Figure 4.

E U2S Modelling and Evaluation

Using the unit sequences for the Voxpopuli (English and German) and Shrutilipi (Marathi) datasets, generated from our S2U encoder, we train vocoders to generate the speech from these unit sequences. We train across 4 GPUs with a learning rate of $2e-4$ with a batch size of 128 (for en-de) and 240 (for mr) and train for 60k updates; other hyperparameters follow Polyak et al. (2021). As a sanity check, we evaluate S2U and U2S by computing the resynthesis WER, which measures how well passing a given speech signal through S2U and U2S preserves the content of the input speech signal.

We compute the resynthesis WER as follows: (1) pass input speech to the S2U encoder and generate the unit sequence, (2) pass the generated unit sequence to our U2S vocoder to synthesize

Method	en VP	de VP	en LJS
Ground Truth	4.89	8.44	3.80
(Lee et al., 2022a)	10.56	-	7.69
Ours	8.53	19.46	6.72

Table 6: S2U + U2S resynthesis performance; WER computed between resynthesized speech transcribed by ASR model and ground truth transcripts. Lower WER is better. We also include the ground-truth speech WER as a lower bound. VP = Voxpopuli, LJS = LJSpeech

speech, (3) transcribe the synthesized speech using ASR (4) compute the Word Error Rate between the transcript and the ground truth transcript of the input speech. To account for the errors from ASR, we compute the WER between the ASR transcript of the input speech utterance ('ground-truth' speech) and the ground truth transcript as a lower bound. We use test sets from English and German Voxpopuli (Wang et al., 2021) and English LJSpeech (Ito and Johnson, 2017) with our synthetic single-speaker speech. Table 6 presents these results. We find that the resynthesis WERs are fairly good for English, and worse for German. Based on qualitative analysis of the German input speech (which is already single-speaker synthetic speech) and resynthesized speech (passed through S2U and U2S), we find that the input speech itself makes stress and pronunciation errors, driving up the Ground Truth WER, which further cascades into the model resynthesis WER. We still use this model because it is the best we could build with existing tools.

F Text-based, Parallel-High-Resource S2T/S2ST models

For completeness, we describe existing text-based, parallel-high-resource models in the literature and showcase their results in Table 7. These models date to 2021 and underperform the text-based parallel-low-resource models in our main results (Table 2) but outperform textless parallel-high-resource models. Rows ⓞ-ⓠ are S2T models while ⓡ is an S2ST model. ⓞ (Iranzo-Sánchez et al., 2019) is an ASR-MT cascade model whose MT component is trained on a large-scale text translation dataset OPUS (Tiedemann, 2012). ⓟ and ⓠ are Transformer-based models from Wang et al. (2021) trained on the union of Europarl-ST and CVSS (total duration 226h) with ⓠ being additionally trained on ≈300h of Voxpopuli aligned speech translation data. ⓡ is the Translatotron 2 (Jia et al., 2021), a spectrogram-to-spectrogram encoder-synthesizer model trained with text supervision for the decoder with 120h of German-English data and about 360h of aligned data in 3 other X-to-English language pairs.

G Example Outputs

We present example outputs from our models. First, we showcase 10 cherry-picked examples, 2 examples from each evaluated language pair and domain in Table 8. Our best models, the post-backtranslation models (rows ⓖ and ⓝ in Tables 2 and 3) perform well on these examples. We present the ground-truth transcripts of the source and target utterances, the ASR transcript of the target utterance predicted by the pre-backtranslation finetuned models (rows ⓕ and ⓜ in Tables 2 and 3) and the ASR transcript of the target utterance predicted by our best models, the post-backtranslation models. We can observe that our post-backtranslation models are able to nearly perfectly translate these cherry-picked examples, which can be categorized into examples with (a) no mistakes (rows 1, 5, 7, 9), (b) valid replacements that largely preserve sentence meaning (rows 2, 4, 8) and (c) minor pronunciation errors (rows 6, 10). On the other hand, predictions from the finetuned model are overall worse, categorized into (a) no mistakes (row 1), (b) valid meaning-preserving replacements (row 2), (c) large meaning changes (row 3, 4, 7, 9, 10) and (d) incoherent output (row 5, 6, 8).

We also sample 5 randomly-picked examples, one from each setting to again compare our pre-backtranslation finetuned models and our best post-backtranslation models in Table 9. The examples show that the models are getting several of the words and semantics right, but often mistranslate certain words and make egregious grammatical and language modelling mistakes. We can see that our post-backtranslation model is overall better than the finetuned model for English-German in row (1), (2), worse in row (3), and performs similarly for rows (4) and (5).

Model	Parallel #hrs	ASR-BLEU ↑ Europarl-ST de→en	ASR-BLEU ↑ Europarl-ST en→de	ASR-BLEU ↑ CVSS de→en
ⓞ Cascaded ASR-MT (Iranzo-Sánchez et al., 2019)	N/A	21.3	22.4	-
ⓟ E2E S2T (Wang et al., 2021)	226h	17.5	-	-
ⓠ E2E S2T w/ Voxpop-Aligned (Wang et al., 2021)	≈500h	18.8	-	-
ⓡ Translatotron 2 (Jia et al., 2021)	120h	-	-	19.7

Table 7: English-German translation evaluation using BLEU for topline S2T models (rows ⓞ-ⓠ) and ASR-BLEU for S2ST model, row ⓡ on Europarl-ST (Iranzo-Sánchez et al., 2019) test set; higher is better. The Parallel #hrs column denotes the size of parallel translation training data.

	Source Utterance	Target Utterance (Gold)	Prediction from finetuned model	Prediction from post-backtranslation model
	en→de (Europarl-ST)			
(1)	you can take initiatives	sie können initiativen ergreifen	sie können initiativen ergreifen	sie können initiativen ergreifen
(2)	madam president i <u>supported</u> this report	frau präsidentin ich habe diesen bericht <u>unterstützt</u>	frau präsidentin ich unterstütze diesen bericht	frau präsidentin ich habe diesen bericht <u>gestimmt</u>
	de→en (Europarl-ST)			
(3)	ich denke da sind wir auf dem richtigen weg	i think we are on the right track <u>here</u>	i think we should be aware of this	i think we are on the right track
(4)	ich denke es ist klar dass die bürger und bürgerinnen der europäischen union <u>diese steuer</u> wollen und ich denke dass <u>es eine große verantwortung ist</u>	i think it is clear that the citizens of the european union want <u>this tax</u> and i think <u>we have a great responsibility here</u>	i think that it is clear that the citizens of the european union want to do with these tasks and to do with the european union what it wants to do	i think it is clear that the citizens of the european union want <u>to be taxed</u> and i think <u>it is a major responsibility</u>
	de→en (CVSS)			
(5)	stellst du die musik bitte auf zimmerlautstärke albert rief seine mutter	are you turning the volume down to room volume albert his mother screamed	are you turning the music albert towards its mountain rock	are you turning the volume down to room volume albert his mother screamed
(6)	<u>los</u> angeles liegt an der westküste	<u>los</u> angeles is located on the west coast	loosen hot air line at the west coast	<u>rose</u> angeles is located on the west coast
	mr→en (S-EP-ST)			
(7)	या कारणांमुळे मी या अहवालाच्या बाजूने मत देऊ शकत नाही	for these reasons i cannot vote in favour of this report	for this reason i am in favour of the report	for these reasons i cannot vote in favour of this report
(8)	ते आधीच <u>सुधारित केले गेले आहे</u> परंतु आणखी काम करणे आवश्यक आहे	it has already <u>been modified</u> but more work needs to be done	it is improving barrowness improving but it must be forgotten	it has already <u>made improvements</u> but more work needs to be done
	mr→en (S-Shr-ST)			
(9)	पंचेचाळीस वर्षावरच्या सर्वांनी लसीकरण अवश्य करुन घ्या	all those above forty five years must get vaccinated	more than forty five years of vaccination papers	all those above forty five years must get vaccinated
(10)	ते काल <u>मुंबैइत</u> बातमीदारांशी बोलत होते	he was talking to reporters in <u>mumbai</u> yesterday	he was talking to reporters in mabay to day	he was talking to reporters in <u>mumba</u> yesterday

Table 8: Cherry-picked examples picked for our best S2ST models (the post-backtranslation models), reporting predictions for both finetuned and post-backtranslation models. We manually annotate the differences between the gold utterance and the prediction from the post-backtranslation model, align them to the source utterance and underline the differences.

	Source Utterance	Target Utterance (Gold)	Prediction from finetuned model	Prediction from post-backtranslation model
(1)	***en→de (Europarl-ST)*** goods and cargo have been delayed or not transported at all and businesses both large and small have been affected	waren und güterlieferungen wurden verschoben oder ganz gestoppt und sowohl kleine als auch große unternehmen sind betroffen	kosovo und konsum wurden zerstört oder wurden nicht erwähnt oder angemessen sein können	günstige und kunden wurden im vorle von kmos nicht erwähnt oder noch nicht erwähnt von allen unternehmen großen unternehmen
(2)	***de→en (Europarl-ST)*** wir sollten hier nicht mit zweierlei maß messen	we must not apply double standards here	we should not do so with these matters	we should not be here with the two sides
(3)	***de→en (CVSS)*** ihr schalldeckel trägt herabhängende quasten und ist mit einem pelikan bekrönt	their sounding board has loose hanging tassels and is crowned with a pelican	year study teacher however remaining costs and an ice and hobbies	child dictatorial territorial castes and is managed by a pellikov
(4)	***mr→en (S-EP-ST)*** नैसर्गिक संसाधने आणि निसर्गाचे संरक्षण करण्यासाठी आपल्याला पर्यावरण संरक्षणाच्या क्षेत्रात संवादाची आवश्यकता आहे	we need dialogue in the field of environmental protection in order to conserve natural resources and nature	in order to protect natural resources and defense quality basis we need a clear signal of environmental protection	we need collectively in the area of protection resources for natural resources and jobs
(5)	***mr→en (S-Shr-ST)*** मुंबैइ आणि उपनगरांमध्ये गेल्या काही दिवसांत जोरदार पाऊस झाल्यामुळं सात मुख्य तलावांच्या पाण्यात लक्षणीय वाढ झाल्यानं मुंबैइला पुढील बारा महिने पाणी पुरवठा सुरळीतपणे होऊ शकणार आहे	heavy rains in mumbai and its suburbs in the last few days have significantly increased the water level in the seven main lakes ensuring smooth water supply to mumbai for the next twelve months	in the last few days ero people who have done in mumba mumbai soon reins have done in the last few days in the last few days mumbai	in mumba and opportunities of mumba and mumba who have received water in seventeen t h needs water in the last few days by the water in the mumbai

Table 9: Randomly sampled examples comparing our finetuned and post-backtranslation models.