# Translatotron 3: Speech to Speech Translation with Monolingual Data

**Eliya Nachmani**[1*] **Alon Levkovitch**[1] **Yifan Ding**[2] **Chulayuth Asawaroengchai**[1]
**Heiga Zen** [2] **Michelle Tadmor Ramanovich**[1]
[1]Google Research, [2]Google DeepMind

## Abstract

S2S from monolingual data?

This paper presents *Translatotron 3*, a novel approach to train a direct speech-to-speech translation model from monolingual speech-text datasets only in a fully unsupervised manner. Translatotron 3 combines masked autoencoder, unsupervised embedding mapping, and back-translation to achieve this goal. Experimental results in speech-to-speech translation tasks between Spanish and English show that Translatotron 3 outperforms a baseline cascade system, reporting 18.14 BLEU points improvement on the synthesized Unpaired-Conversational dataset. In contrast to supervised approaches that necessitate real paired data, which is unavailable, or specialized modeling to replicate para-/non-linguistic information, Translatotron 3 showcases its capability to retain para-/non-linguistic such as pauses, speaking rates, and speaker identity. Audio samples can be found in our website `https://google-research.github.io/lingvo-lab/translatotron3`

## 1 Introduction

Recent years have seen significant advancements in the field of direct speech-to-speech translation (S2ST), which aims to translate speech in one language directly into speech in another language Jia et al. [2022a]. One of the major motivations of direct S2ST is the potential to preserve and translate para-/non-linguistic speech characteristics, such as speaking styles, emotions, emphasis, phonation, and vocal bursts. Although these audible characteristics are essential aspects of human verbal communication, they are often lost in traditional cascade speech translation systems, which typically involve a conversion to text Lavie et al. [1997], Wahlster [2013], Nakamura et al. [2006]. Previous direct S2ST models have demonstrated the ability to preserve some para-/non-linguistic characteristics, such as speaker's voice characteristics, via synthesized targets or speaker modeling Jia et al. [2019b, 2022a], Lee et al. [2022]. However, a significant challenge remains: currently there is no publicly available bilingual speech dataset where the target speech has the same para-/non-linguistic characteristics as or imitation of the source speech. As the target speech in the available datasets is often either synthesized or spoken without semantic translation of the para-/non-linguistic characteristics in the source speech, the translated speech by a direct S2ST model trained from these datasets doesn't convey para-/non-linguistic characteristics of the source speech.

Unsupervised machine translation (UMT) is a task to translate texts with no bilingual text datasets. Previous research in UMT has employed techniques such as back-translation Sennrich et al. [2015], where a synthetic translation of the source language is used as "bilingual text datasets" [Artetxe et al., 2018b, Lample et al., 2018a]. Previous S2ST research, such as Jia et al. [2022a], Lee et al. [2021a], primarily used supervised learning that rely on bilingual speech datasets. This dependency introduces two limitations; (1) Supporting low-resource languages is difficult as collecting bilingual

*Correspondence to `eliyn@google.com`

speech datasets including these languages is hard, and (2) due to the lack of bilingual speech datasets with corresponding para-/non-linguistic characteristics in both source and target languages, para-/non-linguistic characteristics in the source speech cannot be transferred to the translated speech.

This paper addresses the problem of unsupervised S2ST, which can eliminate the requirement for bilingual speech datasets. The proposed approach, called *Translatotron 3*, incorporates (1) pre-training the entire model as a masked autoencoder He et al. [2022] with SpecAugment Park et al. [2019], (2) unsupervised embedding mapping based on the multilingual unsupervised embeddings (MUSE) Conneau et al. [2017], and (3) a reconstruction loss based on back-translation Sennrich et al. [2015], to train an encoder-decoder direct S2ST model from Translatotron 2 Jia et al. [2022a] in a fully unsupervised manner. The model has a shared encoder and two decoders, one for each language (source and target). Training consists of two phases. The first phase trains the shared encoder and language-dependent decoders as a masked autoencoder He et al. [2022] using monolingual speech datasets with the Translatotron 2 reconstruction loss plus the unsupervised MUSE loss Conneau et al. [2017]. The use of the MUSE loss allows us to learn an embedding space shared between source and target languages. The second phase further updates the encoder and decoders by transferring the encoded features into the target language and back to the source language via back-translation to compute a reconstruction loss between source speech and back-translated speech. The main contributions of this work are as follows:

- The first fully unsupervised end-to-end model for direct speech to speech translation.
- Our method outperforms by large margin cascade baseline for unsupervised S2ST in two synthesized datasets and on real speech dataset and approach supervised methods for English to Spanish translation on the CVSS dataset.
- Demonstrating the ability of the proposed approach to transfer para-/non-linguistic characteristics such as pauses, speaking rates and speaker identity, from the source speech through experimental validations on a real speech dataset.

This paper is organized as follows. Sections 2 and 3 provide the relevant literature and the background information, respectively. Section 4 describes the proposed approach. Experimental results are given in Section 5. Concluding remarks are given in the final section.

## 2 Related works

### 2.1 Speech-to-Speech Translation

used to be 3 different parts

Conventional S2ST systems such as Lavie et al. [1997], Wahlster [2013], Nakamura et al. [2006] were implemented as a cascade of three separate components: automatic speech recognition (ASR) for the source language, machine translation (MT) for converting source language text to target language text, and text-to-speech synthesis (TTS) for generating speech from the translated text. Jia et al. [2019b] proposed the first end-to-end direct S2ST model called *Translatotron*. It is a single sequence-to-sequence model from source speech to target speech trained in an end-to-end manner using multi-objective tasks. Subsequently, *Translatotron 2* Jia et al. [2022a] was proposed as an improvement on the original Translatotron model,offering better performance, controllability, and robustness. There are various cascaded S2ST systems that utilize learned discrete speech representations as an intermediate representation. For example, Jia et al. [2019b] proposed an S2ST system that first translated the source speech into a discrete representation of the target speech, using a separately trained vector quantized variational autoencoder (VQ-VAE) model van den Oord et al. [2017] for prediction. Zhang et al. [2021] jointly trained a VQ-VAE model with a supervised phoneme recognition objective in different languages. Lee et al. [2021b] used a separately trained vocoder to directly predict the waveform from the discrete representation, without relying on spectrograms. Huang et al. [2022a] used bilateral perturbation with discrete representations separately learned to normalize and enhance the linguistic information contained in the representation. Note that end-to-end training is difficult for these approaches, which may limit their overall performance. Furthermore, these approaches may not be effective in preserving para-/non-linguistic information. There are other approaches based on self-supervision to use untranscribed speech and unspoken text datasets. Tang et al. [2022] used two sub-tasks for pre-training, one for untranscribed speech data and another for unspoken text data. Both of them used masking and reconstruction to learn representations. Dong et al. [2022] used pseudo-labels generated from ASR as input to MT then synthesized translated text using TTS. Finally, it used the generated S2ST dataset to train a model in a supervised manner. There are also approaches based on self-supervised learning techniques to leverage untranscribed speech datasets. Tang et al. [2022] proposed a unified pre-training technique that utilizes two sub-tasks, one for untranscribed speech data and one for unspoken text data. Both sub-tasks involve masking and reconstruction techniques to learn better representations. Similarly, Dong et al. [2022] leveraged pseudo-labels generated from ASR data using MT and TTS to create a generated S2ST dataset for supervised training. There have been various end-to-end S2ST models which aim to improve the performance and efficiency. Kano et al. [2021] introduced an S2ST model with a cascade of three autoregressive decoders, and used pre-trained MT and TTS models as teacher models to facilitate the training of the S2ST model. It requires pre-trained ASR, MT, and TTS models, and multiple training iterations. Wang et al. [2022] proposed an approach that combines teacher models and pseudo-labeling to utilize unlabelled data. Their approach consists of three steps, in which the first step is to adapt a pre-trained speech model to a new language, the second step is to use a cascade of ASR, MT, and TTS to generate pseudo-labels, and the third step is to train an end-to-end model using the pseudo-labeled data. Inaguma et al. [2022] proposed UnitY, a S2ST model using a two-pass architecture. It first generates textual representations using a speech encoder and a text decoder then predicts discrete acoustic units from the textual representations. Finally, it predicts the waveform directly from the discrete units. Wei et al. [2022] proposed an S2ST model that is jointly pre-trained with monolingual speech and bilingual text datasets. Their approach used the HuBERT tokenization Hsu et al. [2021] and masked auto-encoding for speech and text pre-training. After pre-training, the model is fine-tuned using a bilingual speech dataset, such that a mapping between the source language and target language tokens is learned. Finally, the tokens are directly converted into waveform.

### 2.2 Unsupervised Machine Translation

Unsupervised machine translation (UMT) aims to perform MT without the use of bilingual text datasets; Training is done solely using unsupervised, monolingual text datasets. Recent work such as Artetxe et al. [2018b], Lample et al. [2018a] has shown promising results on supervised MT benchmarks using only monolingual corpora. Both approaches were built upon unsupervised cross-lingual embedding mappings, which independently trained word embeddings in two languages then learned a transformation to map them to a shared space through self-learning Artetxe et al. [2017, 2018a] or adversarial training Conneau et al. [2017]. The resulting cross-lingual embeddings were used to initialize a shared encoder for both languages, and the entire system was trained using a combination of masking and back-translation techniques. Another approach to UMT is the use of large language models (LLMs) such as MASS Song et al. [2019] and GPT-3 Brown et al. [2020]. Although they were not trained on supervised, bilingual text datasets, they still achieved the performance close to state-of-the-art on supervised MT benchmarks. These approaches leverage the large pre-trained models and fine-tune them on the monolingual text data using unsupervised learning techniques to learn cross-lingual representations.

use unpaired data via 2 encoder and a cross embedding module.

## 3 Background

**Translatotron 2:** Translatotron 2 is an end-to-end speech-to-speech translation model that is composed of four key components. Specifically, it comprises of a speech encoder, a linguistic decoder, an acoustic synthesizer, and a single attention module that connects these components together. This architecture allows the model to translate speech in one language to speech in another language in a direct manner Jia et al. [2022a]. The model is trained in a supervised fashion, using bilingual speech-text datasets with a loss consisting of a spectrogram reconstruction loss, a total duration loss, and auxiliary phoneme loss. These losses are designed to improve the quality of the generated speech and learn alignments between source speech and target text. The model achieved the state-of-the-art performance on speech-to-speech translation tasks. **Architecture:** The first component is the speech encoder $\mathcal{E}$. The encoder's input is a spectrogram sequence of source speech, $S^s$, and its output is an intermediate representation $\mathcal{E}(S^s)$. The encoder first uses a convolutional layer to sub-sample the input then processes it with a stack of Conformer blocks Gulati et al. [2020]. The second component is an attention module, which is based on multi-head attention Vaswani et al. [2017] and takes $\mathcal{E}(S^s)$ as its input. This module represents the alignment between source spectrogram sequence and target phoneme sequence, since its queries are from the linguistic decoder. It outputs the acoustic information from the source speech, summarized at per-phoneme level. The third component is

4 parts connected by a transformer

3 losses

the linguistic decoder. It is responsible for producing a phoneme sequence corresponding to the target speech. This component employs an autoregressive LSTM stack with teacher-forcing. The last component is the acoustic synthesizer. It is responsible for generating the spectrogram of the translated speech. It takes both intermediate representation in the linguistic-decoder and output from the attention module as its input. It has a per-phoneme duration predictor and an autoregressive LSTM-based decoder from Non-Attentive Tacotron TTS model Shen et al. [2020]. The whole process can be formulated as:

$$S^{t'} = \mathcal{D}\left(\mathcal{E}\left(S^{s}\right)\right), \tag{1}$$

where $S^{t'}$ is the predicted target spectrogram sequence, and $\mathcal{D}$ denotes the entire decoder containing the attention module, the linguistic decoder and the acoustic synthesizer. **Training Objective:** Translatotron 2 is an end-to-end direct S2ST model that is trained in a supervised manner. Its primary loss function used during training is the combination of $L_1$ and $L_2$ losses between predicted target spectrogram sequence $S^{t'}$ and real target spectrogram sequence $S^t$. Given a paired example $\{S^{t'}, S^t\}$, the loss function can be expressed as

T - # of frames
K - # of spec bins


$$\mathcal{L}_{\text{spec}}\left(S^{t'}, S^{t}\right) = \frac{1}{TK}\sum_{i=1}^{T}\sum_{j=1}^{2}\left\|S_i^{t'} - S_i^{t}\right\|_j^j, \tag{2}$$

L1 and L2 loss over entire spectogram

where $S_i^t$ denotes the $i$-th frame of $S^t$, $T$ is the number of frames in $S^t$, $K$ is the number of frequency bins in $S_i^t$, and $\|\cdot\|_j^j$ denotes the $L_j$ distance. The second loss is a duration loss between total number of frames $T$ and the sum of phoneme durations predicted in the acoustic synthesizer. It is given as

GT length

$$\mathcal{L}_{\text{dur}} = \left(T - \boxed{\sum_{i=1}^{p} d_i}\right)^2, \tag{3}$$

sum of phoneme durations (total predicted length)

where $d_i$ is the predicted duration for the $i$-th phoneme. The last loss is the auxiliary phoneme loss. Let $\tilde{P}^t = \{\tilde{P}_1^t, \ldots, \tilde{P}_p^t\}$ be the sequence of predicted probabilities over target phonemes, $P^t = \{P_1^t, \ldots, P_p^t\}$ be the ground-truth target phoneme sequence, and $\mathrm{CE}(\cdot,\cdot)$ be the cross entropy. Then we have

Predicted dist over all phonemes at time t

$$\mathcal{L}_{\text{phn}}\left(\tilde{P}^t, P^t\right) = \frac{1}{P}\sum_{i=1}^{p}\mathrm{CE}\left(\tilde{P}_i^t, P_i^t\right). \tag{4}$$

GT phoneme at time t

The overall loss is given as

$$\mathcal{L} = \mathcal{L}_{\text{spec}}\left(S^{t'}, S^{t}\right) + \lambda_{\text{dur}}\mathcal{L}_{\text{dur}} + \lambda_{\text{phn}}\mathcal{L}_{\text{phn}}\left(\tilde{P}^t, P^t\right), \tag{5}$$

where $\lambda_{\text{dur}}$ and $\lambda_{\text{phn}}$ are the weights associated with the duration and phoneme auxiliary losses, respectively.

### 3.1 MUSE Embedding

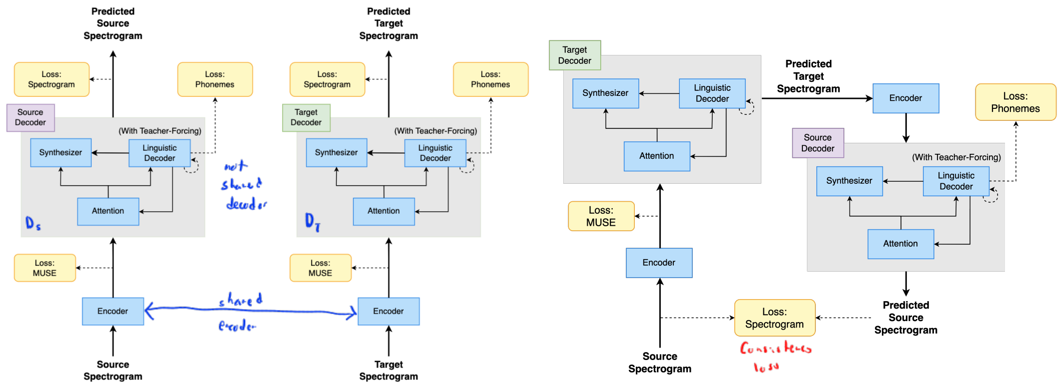
encode text in different languages

Multilingual Unsupervised embeddings (MUSE) Conneau et al. [2017] is a technique to acquire multilingual word embeddings. Assuming that the learning process starts from two sets of word embeddings trained on monolingual corpora, each corresponding to a different language. We can represent the embeddings as two matrices $X \in \mathbb{R}^{c\times M}; Y \in \mathbb{R}^{c\times M}$, where $M$ is the number of representative words and $c$ is the dimension of each embedding. We assume that in the two languages we use the $M$ most frequent words to compute the embeddings and compute them in such a way that they have the same dimension $c$. The goal of the MUSE training procedure is to learn a mapping $W^*$ such that $W^* = \arg\min_{W\in\mathbb{R}^{c\times c}} \|WX - Y\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm. Xing et al. [2015] showed that this technique can be improved by imposing an orthogonality constraint on $W$, thus boiling this objective down to the Procrustes problem as

Find an optimal W, linear transform s.t. the Frobenius norm between the language embeddings is minimized.

$$\begin{gathered} W^* = \underset{W\in\mathbb{R}^{c\times c}}{\arg\min}\|WX - Y\|_F = UV^\intercal, \\ \text{with } U\Sigma V^\intercal = \mathrm{SVD}\left(YX^\intercal\right). \end{gathered} \tag{6}$$

In the unsupervised setting, Eq. (6) is solved indirectly using an adversarial approach. A discriminator is used to classify between elements randomly sampled from $WX$ and $Y$. Then $W$ is updated to




(a) (b)

Figure 1: The two training phases in the proposed approach. (1) Phase 1 uses the reconstruction loss via the auto-encoding path. (2) Phase 2 employs the reconstruction loss via back-translation.

prevent the discriminator from inferring the origin of its input. The discriminator is a simple classifier that is trained to predict if an embedding originated in the source or target language. The loss can be formulated as $\mathcal{L}_w = -\frac{1}{M}\sum_{i=1}^{M}\left\{(\log P_D(source = 0 \mid Wx_i) + \log P_D\left(source = 1 \mid y_i\right)\right\}$, where $P_D(\cdot)$ denotes the discriminator.

max prob that transformed x is target

max prob that $y_i$ is from the source

## 4 Translatotron 3

The proposed approach, Translatotron 3, adopts a novel architecture to allow unsupervised S2ST where there are a shared encoder and separate decoders for the source and target languages. The model is trained using a combination of the unsupervised MUSE embedding loss, reconstruction loss, and S2S back-translation loss. During inference, the shared encoder is utilized to encode the input into a multilingual embedding space, which is subsequently decoded by the target language decoder.

### 4.1 Architecture

The proposed Translatotron 3 employs a shared encoder $\mathcal{E}$ to encode both the source and target languages. The decoder $\mathcal{D}$ is composed of a linguistic decoder, an acoustic synthesizer, and a singular attention module, like Translatotron 2. There are two decoders, one for the source language $\mathcal{D}_s$ and another for the target language $\mathcal{D}_t$. During training, we use monolingual speech-text datasets. It is important to note that these monolingual speech-text datasets are not translations of one another.

#### 4.1.1 Encoder

The encoder $\mathcal{E}$ has the same architecture as the speech encoder in the Translatotron 2 Jia et al. [2022a]. The output of the encoder $\mathcal{E}(S^{in})$ is split into two parts:

$$\mathcal{E}(S^{in}) = \left[\mathcal{E}_m(S^{in}), \mathcal{E}_o(S^{in})\right], \tag{7}$$

encoder output split into MUSE and latter output

where $S^{in}$ can be the source or target language. The first half of the output $\mathcal{E}_m(S^{in})$ is trained to be the MUSE embeddings of the text of the input spectrogram $S^{in}$. This is forced using the MUSE loss that will be explained in Sec.4.2.1. The latter half $\mathcal{E}_o(S^{in})$ is updated without the MUSE loss. It is important to note that the same encoder $\mathcal{E}$ is shared between source and target languages. Furthermore, the MUSE embedding is multilingual in nature. As a result, the encoder is able to learn a multilingual embedding space across source and target languages. This allows a more efficient and effective encoding of the input, as the encoder is able to encode speech in both languages into a common embedding space, rather than maintaining a separate embedding space for each language.

#### 4.1.2 Decoder

Like Translatotron 2, the decoder $\mathcal{D}$ is composed of three distinct components, namely the linguistic decoder, the acoustic synthesizer, and the attention module. To effectively handle the different properties of the source and target languages, it has two separate decoders, $\mathcal{D}^s$ and $\mathcal{D}^t$, for the source and target languages, respectively. The decoder output can be formulate as

$$S^{out} = \mathcal{D}^{out}\left(\mathcal{E}(S^{in})\right), \tag{8}$$

where $S^{in}$ and $S^{out}$ correspond to the input and out spectrogram sequences. Note that both $S^{in}$ and $S^{out}$ may represent either the source or target language, as well as $\mathcal{D}^{out}$.

### 4.2 Training Objective

Phases:
1: Train source/target encoder/decoders
2: Combine models for translation task

Figure 1 illustrates the training methodology of the proposed approach. It consists of two phases: (1) auto-encoding, reconstruction phase and (2) back-translation phase. In the first phase, the network is trained to auto-encode the input to a multilingual embedding space using the MUSE loss and the reconstruction loss. This phase aims to ensure that the network generates meaningful multi-lingual representations. In the second phase, the network is further trained to translate the input spectrogram by utilizing the back-translation loss. To mitigate the issue of catastrophic forgetting and enforcing the latent space to be multilingual, the MUSE loss and the reconstruction loss are also applied in the second phase of the training. To ensure that the encoder learns meaningful properties of the input, rather than simply reconstructing the input, we apply SpecAugment Park et al. [2019] to encoder input at both phases. It has been shown to effectively improve the generalization capabilities of the encoder by augmenting the input data. As SpecAugment masks input over time and frequency axes, the first auto-encoding phase with SpecAugment can be viewed as masked auto-encoder training He et al. [2022].

SpecAugment-
1: Time shifting
2: Randomly mask some frequencies
3: Randomly mask time segments

Orig
Freq mask
time + freq mask

#### 4.2.1 MUSE Loss

To ensure that the encoder $\mathcal{E}$ generates multi-lingual representations that are meaningful for both decoders $\mathcal{D}^s$ and $\mathcal{D}^t$, we employ a MUSE loss during training. The MUSE loss forces the encoder to generate such a representation by using pre-trained MUSE embeddings, as detailed in Sec. 3.1. These embeddings may be computed in a supervised or unsupervised manner. During the training process, given an input transcript with $n$ words, we extract $n$ corresponding MUSE embeddings $E \in \mathbb{R}^{n \times d}$ from the embeddings of the input language. The error between $E$ and the $n$ first output vectors of the encoder $\mathcal{E}$ is then minimized. This can be mathematically represented as

$$\mathcal{L}_{\text{MUSE}}(S^{in}) = \frac{1}{n}\sum_{i=1}^{n}\left\|\mathcal{E}(S^{in})_i - E_i\right\|_2^2, \tag{9}$$

optimize $L_2$ loss between "rich" MUSE embeddings and encoder output

where $S^{in}$ represents the input spectrogram, which may be in the source or target language, $E_i$ is the $d$-dimensional embedding vector for $i$-th word. Note that the encoder is indifferent to the language of the input during inference due to the multilingual nature of the embeddings.

#### 4.2.2 Reconstruction Loss

Figure 1(a) illustrates the reconstruction training phase. In this phase, the network learns to auto-encode both the source and target languages. The reconstruction loss is computed as a linear combination of $\mathcal{L}^{spec}$, $\mathcal{L}^{dur}$ and $\mathcal{L}^{phn}$ for both the source and target languages. The reconstruction can be summarized in the following equations:

$$\begin{aligned}\mathcal{L}_{\text{recon}} = &\ \mathcal{L}^{src}_{\text{spec}}\left(S^{s\prime}, S^s\right) + \mathcal{L}^{src}_{\text{dur}} + \mathcal{L}^{src}_{\text{phn}}\left(\tilde{P}^s, P^s\right) \quad \text{source model} \\ &+ \mathcal{L}^{tgt}_{\text{spec}}\left(S^{t\prime}, S^t\right) + \mathcal{L}^{tgt}_{\text{dur}} + \mathcal{L}^{tgt}_{\text{phn}}\left(\tilde{P}^t, P^t\right), \quad \text{target model}\end{aligned} \tag{10}$$

where $S^s$ and $S^t$ are the source and target spectrograms respectively, $S^{s\prime}$ and $S^{t\prime}$ represents the model spectrogram output predictions, $\tilde{P}^s$ and $\tilde{P}^t$ represent the probability distributions of the phonemes of the source and target languages respectively, and $P^s$ and $P^t$ represent the phoneme sequences of the source and target text respectively. The predicted spectrogram sequences $S^{s\prime}$ and $S^{t\prime}$ are given by

$$S^{s\prime} = \mathcal{D}^s\left(\mathcal{E}(S^s)\right), \quad S^{t\prime} = \mathcal{D}^t\left(\mathcal{E}(S^t)\right). \tag{11}$$

### 4.2.3 Back-Translation Loss

Figure 1(b) illustrates the back-translation training phase. To apply unsupervised training to the sequence-to-sequence task, the back-translation technique is utilized in this phase. The process begins with the encoding of the source input spectrogram, represented as $\mathcal{E}(S^s)$, followed by the use of the target language decoder to produce a pseudo-translation, denoted as $\hat{S}^{t'}$ as

$$\hat{S}^{t'} = \mathcal{D}^t\left(\mathcal{E}(S^s)\right). \tag{12}$$

Subsequently, we proceed to encode the pseudo-translation, utilizing the encoder $\mathcal{E}(\hat{S}^{t'})$. The final step in this process involves decoding the encoded pseudo-translation using the decoder of the source language:

$$\hat{S}^s = \mathcal{D}^s\left(\mathcal{E}(\hat{S}^{t'})\right). \tag{13}$$

Lastly, we apply a loss function to minimize the dissimilarity to the input spectrogram:

$$\mathcal{L}^{src2tgt}_{\text{back-translation}} = \mathcal{L}^{src}_{\text{spec}}\left(\hat{S}^s, S^s\right) + \mathcal{L}^{src}_{\text{dur}} + \mathcal{L}^{src}_{\text{phn}}\left(\tilde{P}^s, P^s\right). \tag{14}$$

We also apply the aforementioned process in the reverse direction, specifically, from the target to the source language. This process entails utilizing the same methodologies and techniques as previously described, with the only difference being the direction of the translation:

$$\mathcal{L}^{tgt2src}_{\text{back-translation}} = \mathcal{L}^{tgt}_{\text{spec}}\left(\hat{S}^t, S^t\right) + \mathcal{L}^{tgt}_{\text{dur}} + \mathcal{L}^{tgt}_{\text{phn}}\left(\tilde{P}^t, P^t\right). \tag{15}$$

The overall back-translation loss is given as

Loss for both directions

$$\mathcal{L}_{\text{back-translation}} = \mathcal{L}^{src2tgt}_{\text{back-translation}} + \mathcal{L}^{tgt2src}_{\text{back-translation}}. \tag{16}$$

Note that $\mathcal{L}^{src}_{\text{spec}}(\hat{S}^s, S^s) + \mathcal{L}^{src}_{\text{spec}}(\hat{S}^t, S^t)$, which is a part of $\mathcal{L}_{\text{back-translation}}$, can be viewed as a cycle consistency loss proposed for unpaired image-to-image translation Zhu et al. [2017].

### 4.2.4 Overall Loss

For the reconstruction phase, the loss is given by

$$\mathcal{L}_{\text{recon-phase}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{MUSE}}(S^s) + \mathcal{L}_{\text{MUSE}}(S^t). \tag{17}$$

In the back-translation phase, the optimization of both the back-translation loss, represented as $\mathcal{L}_{\text{back-translation}}$, and the reconstruction loss, represented as $\mathcal{L}_{\text{recon-phase}}$, is carried out to ensure that the encoder output produces a multilingual embedding. The overall loss is given by

Retain old loss

$$\mathcal{L}_{\text{BT-phase}} = \mathcal{L}_{\text{back-translation}} + \mathcal{L}_{\text{recon-phase}}. \tag{18}$$

## 5 Experiments and Results

To empirically evaluate the performance of the proposed approach, we conducted experiments on the English and Spanish languages using various datasets, including the Common Voice 11 dataset Ardila et al. [2020], as well as two synthesized datasets derived from the Conversational Jia et al. [2019a] and Common Voice 11 datasets. See supplementary material Appendix Tab. 5 for the details of each dataset. Note that all of them were monolingual speech-text datasets. We trained Translatotron 3 models from these synthetic and real speech datasets. A comprehensive table of hyper-parameters is available in supplementary material Appendix A.2. To reconstruct speech from the predicted spectrogram, we used a non-autoregressive neutral vocoder called WaveFit Koizumi et al. [2022]. As it was pre-trained on multi-speaker English-only datasets but generalized well to unseen speakers, we did not fine-tune this model for the proposed model. The model architecture and hyper parameters of WaveFit was basically the same as the original paper Koizumi et al. [2022] except the use the multi-period discriminator from HiFi-GAN Kong et al. [2020] with the same hyper-parameters as the original paper and adjusting sampling frequency-dependent parameters for 16 kHz sampling. Our proposed model was trained using 64 TPUv4 chips Jouppi et al. [2023], which required one week to complete training. During training, the model consisted of 113 million parameters. For inference, we only require a single decoder, and thus the number of parameters is reduced to 80 million. We used a cascaded S2ST system as a baseline. For its MT module, we used a nearest neighbor MT system similar to the one used in Lample et al. [2018b]; after extracting embedding for each word in a source sentence, we computed the similarity score for each embedding in the target language by the dot product between embeddings of the source word and those in the target language then picked up the word in the target language with the highest similarity score. The baseline used a simple technique to translate text between languages, yet it is enough for conveying the general meaning of the translated sentences.

### 5.1 Metrics

**BLEU score:** The translation performance was measured by BLEU on ASR transcriptions from the translated speech (in lowercase, excluding punctuation marks except for apostrophes), compared to the corresponding reference translation text. Because ASR can make errors, such BLEU can be considered as a lower bound of the translation performance. **SQuId MOS:** There has been a growing interest in automatic metrics for speech synthesis, with models such as Patton et al. [2016], Lo et al. [2019], Huang et al. [2022b]. These models predicts a 5-scale subjective mean opinion score (MOS) in naturalness given a speech. SQuId Sellam et al. [2022] is one of the latest multilingual models for MOS prediction based on mSLAM Bapna et al. [2022].**Average cosine similarity of speaker embeddings:** We use the speaker encoder of the PnG NAT Text-To-Speech model Morioka et al. [2022] to compute the speaker embedding of the input and of the translated output. Using cosine similarity we measure the distance between the embeddings and report the average over the whole test set.

Table 1: Performance of S2ST trained with unsupervised MUSE embedding. "SMOS" denotes the 5-scale MOS in naturalness predicted by the SQuId model.

| Dataset | En → Es | | Es → En | |
|---|---|---|---|---|
| | SMOS | BLEU | SMOS | BLEU |
| **U-Conv** | | | | |
| Baseline | 4.05 | 5.58 | 4.03 | 6.13 |
| Proposed | 3.89 | 18.85 | 4.10 | 24.27 |
| **S-CV11** | | | | |
| Baseline | 4.08 | 9.46 | 4.05 | 10.48 |
| Proposed | 4.03 | 13.45 | 4.17 | 14.25 |

Table 2: Ablation analysis. The scores in this table are BLEU.

| **Model** | **En → Es** | **Es → En** |
|---|---|---|
| Proposed | 18.85 | 24.27 |
| −Recon | 0.41 | 2.99 |
| −BackTrans | 1.91 | 3.62 |
| −MuseLoss | 5.44 | 6.22 |
| −SpecAug | 9.23 | 12.88 |

MUSE loss make big difference

### 5.2 Synthesized Speech Data

**Unpaired Conversational Dataset:** The proprietary dataset described in Jia et al. [2019a] was obtained by crowd-sourcing humans to read the both sides of a conversational Spanish-English MT dataset. We synthesized both source and target speech using a multi-lingual multi-speaker Phoneme-and-Grapheme Non-Attentive Tacotron (PnG NAT) TTS model Jia et al. [2021], Shen et al. [2020] and a WaveRNN-based neural vocoder Kalchbrenner et al. [2018]. These TTS and vocoder models were trained from proprietary TTS datasets with more than 1,000 speakers. To prevent that data from being paired, we split the data in half and for one half pick English as the source language and for the other half pick Spanish as the source language. For the training set, we ended up with ∼379K English utterances that were ∼371 hours of speech and ∼379K Spanish utterances that were ∼350 hours of speech. For the test set we used ∼6.5K paired utterances that were ∼5.4 hours in English and ∼4.8 hours in Spanish. Table 1 shows the experimental results. The proposed approach demonstrated substantial improvements over the baseline; +13.27 increase in BLEU for English→Spanish and +18.14 increase in BLEU for Spanish→English. Regarding naturalness of the translated speech, the proposed approach got a slightly better SQuId MOS for Spanish→English but worse SQuId MOS for English→Spanish. Overall, the proposed approach was on par with the TTS baseline, which got ∼4.0 in SQuId MOS. Audio samples are available in our website: `https://google-research.github.io/lingvo-lab/translatotron3`. **Common Voice 11:** We considered two languages from this dataset, English and Spanish, and we randomly sampled ∼50% of English samples to balance the data.

Since the dataset was noisy and diverse, we generated the data using the same TTS model and vocoder that was employed for the Conversational dataset. This was done to ensure consistency and eliminate any potential variations in the dataset that could impact the results of the study. For evaluation, we used the CVSS Jia et al. [2022b] Spanish-English dataset which is a clean and paired subset of the Common Voice Spanish dataset with verified translation to English. Since the target speech was synthesized using state-of-the-art TTS systems, we also synthesized the source speech. The resulting test set had $\sim$13K paired utterances. The proposed approach demonstrated a notable improvement in performance when compared to the baseline. As illustrated in Table 1, the proposed approach achieved approximately $30\%$ improvement in BLEU over the baseline. Specifically, in Spanish to English, the proposed approach exhibited an improvement of $+3.77$ in BLEU over the baseline, and in English to Spanish, it demonstrated an improvement of $+3.99$ in BLEU over the baseline. With respect to the naturalness, the proposed approach got SQuId MOS similar to that of the baseline. The website `https://google-research.github.io/lingvo-lab/translatotron3` include audio samples for further review.

### 5.3 Real Speech Data

For real speech experiment, we used Common Voice 11. Natural (non-synthesized) monolingual speech-text datasets both in English and Spanish were used for training. The evaluation of Spanish-English real speech Translation was conducted using real speech with verified translation from the CoVoST2 test set Wang et al. [2020], which is a subset of the Common Voice 11 test set. However, it should be noted there is no English-Spanish CoVoST2 test set or otherwise any test set from Common Voice 11 with verified translation for English-Spanish and therefore only an English-Spanish evaluation was omitted from this experiment. The proposed approach achieved an $10.67$ in BLEU for the task of Spanish-English Translation which is an improvement of $+0.75$ in BLEU over the baseline which achieves $9.92$ BLEU. Audio samples are available in our website: `https://google-research.github.io/lingvo-lab/translatotron3`.

### 5.4 Comparison to Supervised Approaches

Table 3 shows the performance of the cascade baseline, the proposed approach, and the supervised approaches on the CVSS dataset. For the English to Spanish translation task, the proposed approach demonstrated a level of performance that was comparable to a supervised approach. Notably, the margin by which we fell behind was relatively small, particularly when considering the fact that unsupervised translation is a more challenging task in comparison to supervised translation. On the other hand, for the Spanish to English translation task, the supervised approaches exhibited significantly better performance. While this margin may appear significant, it is worth noting that the difficulty of the task at hand should also be taken into consideration.

### 5.5 Ablation Analysis

To understand the individual impacts of various components within the proposed approach, an ablation study was conducted. This study involved the removal of specific components. (1) disabled SpecAugment ("-SpecAug") Park et al. [2019], (2) removed Back-Translation loss ("-BackTrans") Eq. (16), (3) removed the MUSE embeeding loss Eq. (9) at the encoder output ("-MuseLoss"), and (4) removed the reconstruction loss Eq. (5) ("-Recon"). The results are summarized in Table 2. It can be seen from the table that that each of the aforementioned components contributed to the overall performance improvement of the proposed approach, with the reconstruction loss and the back-translation loss having the most significant impact. Additionally, the inclusion of the MUSE loss in the model also resulted in a notable improvement in performance.

### 5.6 Para-/Non-Linguistic Feature Preservation

Our findings show that the proposed approach can preserve paralinguistic information during translation. The original speaker's voice is preserved to a high degree, as shown in Tab.4. Tab.4 reports the average similarity scores for the test set. For the baseline, we use a TTS system and pick speakers randomly. A Higher score is better, where 0 corresponds to no correlation, 1 corresponds to the same vector, and -1 is an opposing vector. A value of 0.6 shows that there is a good correlation between input and output speaker identities. The baseline values of

Table 3: Comparison of supervised and unsupervised approaches on the CVSS dataset. [1]Results from Jia et al. [2022b], [2]results from Inaguma et al. [2022], and TranSpeech results are based on Huang et al. [2022a].

| Approach | En → Es SMOS | En → Es BLEU | Es → En SMOS | Es → En BLEU |
|---|---|---|---|---|
| **Unsupervised** | | | | |
| Baseline | 4.08 | 9.46 | 4.05 | 10.48 |
| Proposed | 4.03 | 13.45 | 4.17 | 14.25 |
| **Supervised** | | | | |
| T1[1] | - | - | - | 19.8 |
| T2[1] | - | - | 4.61 | 25.4 |
| Cascaded[1] | - | - | 4.64 | 28.8 |
| TranSpeech | - | 14.94 | - | - |
| S2TT+TTS[2] | - | - | - | 18.2 |
| S2UT[2] | - | - | - | 25.9 |
| UnitY[2] | - | - | - | 29.0 |

Table 4: Average cosine similarity between input speaker and output speaker.

| Dataset | En → Es C-Sim | Es → En C-Sim |
|---|---|---|
| **U-Conv** | | |
| Baseline | 0.157 | 0.165 |
| Proposed | 0.627 | 0.651 |
| **S-CV11** | | |
| Baseline | 0.162 | 0.176 |
| Proposed | 0.630 | 0.560 |
| **CV11** | | |
| Baseline | - | 0.176 |
| Proposed | - | 0.341 |

0.16 shows that the speaker embedding vectors are not all aligned in the same direction and that picking random speakers gives no correlation. Moreover, our approach shows promise in maintaining critical elements such as pauses and speaking rates as can be examined in our website: `https://google-research.github.io/lingvo-lab/translatotron3`

## 6 Conclusion

This paper presented Translatotron 3, an unsupervised direct speech-to-speech translation model. It uses the unsupervised embedding word mapping technique and a back-translation training procedure. Unlike the previous approaches, the proposed approach can implicitly preserve some elements of para-/non-linguistic characteristics in the source speech. We demonstrated that the proposed approach improved upon the unsupervised cascade baseline (up to 10.51 increase in BLEU) and approached the performance of supervised systems on the CVSS dataset (by 1.95 gap in BLEU). This suggests that Translatotron 3 is an effective approach for unsupervised S2ST that is able to retain important information from the source speech in the target translation.


## Acknowledgements

The authors would like to thank Nanxin Chen, Yuma Koizumi, Soroosh Mariooryad, RJ Skerry-Ryan, Neil Zeghidour, Christian Frank, Marco Tagliasacchi, Nadav Bar, and the rest of the Google Research team for helpful discussions and previous work on data preparation.


## References


R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common Voice: A massively-multilingual speech corpus. In *Proc. LREC*, 2020.

M. Artetxe, G. Labaka, and E. Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proc. ACL*, pages 451–462, 2017.

M. Artetxe, G. Labaka, and E. Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proc. ACL*, 2018a.

M. Artetxe, G. Labaka, E. Agirre, and K. Cho. Unsupervised neural machine translation. In *Proc. ICLR*, 2018b.

A. Bapna, C. Cherry, Y. Zhang, Y. Jia, M. Johnson, Y. Cheng, S. Khanuja, J. Riesa, and A. Conneau. mSLAM: Massively multilingual joint pre-training for speech and text. *arXiv:2202.01374*, 2022.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Proc. NeurIPS*, 33:1877–1901, 2020.

A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data. *arXiv:1710.04087*, 2017.

Q. Dong, F. Yue, T. Ko, M. Wang, Q. Bai, and Y. Zhang. Leveraging pseudo-labeled data to improve direct speech-to-speech translation. *arXiv:2205.08993*, 2022.

A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv:2005.08100*, 2020.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proc. CVPR*, pages 16000–16009, 2022.

W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. ASLP*, 29:3451–3460, 2021.

R. Huang, Z. Zhao, J. Liu, H. Liu, Y. Ren, L. Zhang, and J. He. TranSpeech: Speech-to-speech translation with bilateral perturbation. *arXiv:2205.12523*, 2022a.

W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda. LDNet: Unified listener dependent modeling in MOS prediction for synthetic speech. In *Proc. ICASSP*, pages 896–900, 2022b.

H. Inaguma, S. Popuri, I. Kulikov, P.-J. Chen, C. Wang, Y.-A. Chung, Y. Tang, A. Lee, S. Watanabe, and J. Pino. UnitY: Two-pass direct speech-to-speech translation with discrete units. *arXiv:2212.08055*, 2022.

Y. Jia, M. Johnson, W. Macherey, R. J. Weiss, Y. Cao, C.-C. Chiu, N. Ari, S. Laurenzo, and Y. Wu. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *Proc. ICASSP*, pages 7180–7184, 2019a.

Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu. Direct speech-to-speech translation with a sequence-to-sequence model. *Proc. Interspeech*, pages 1123–1127, 2019b.

Y. Jia, H. Zen, J. Shen, Y. Zhang, and Y. Wu. PnG BERT: Augmented BERT on phonemes and graphemes for neural TTS. In *Proc. Interspeech*, 2021.

Y. Jia, M. T. Ramanovich, T. Remez, and R. Pomerantz. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *Proc. ICML*, pages 10120–10134, 2022a.

Y. Jia, M. T. Ramanovich, Q. Wang, and H. Zen. CVSS corpus and massively multilingual speech-to-speech translation. *arXiv:2201.03713*, 2022b.

N. P. Jouppi, G. Kurian, S. Li, P. Ma, R. Nagarajan, L. Nai, N. Patil, S. Subramanian, A. Swing, B. Towles, et al. Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings. *arXiv preprint arXiv:2304.01433*, 2023.

N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu. Efficient neural audio synthesis. In *Proc. ICML*, 2018.

T. Kano, S. Sakti, and S. Nakamura. Transformer-based direct speech-to-speech translation with transcoder. In *Proc. IEEE SLT*, pages 958–965, 2021.

Y. Koizumi, K. Yatabe, H. Zen, and M. Bacchiani. WaveFit: An iterative and non-autoregressive neural vocoder based on fixed-point iteration. *arXiv:2210.01029*, 2022.

J. Kong, J. Kim, and J. Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Proc. NeurIPS*, 33:17022–17033, 2020.

G. Lample, A. Conneau, L. Denoyer, and M. Ranzato. Unsupervised machine translation using monolingual corpora only. In *Proc. ICLR*, 2018a.

G. Lample, A. Conneau, M. Ranzato, L. Denoyer, and H. Je'gou. Word translation without parallel data. In *Proc. ICLR*, 2018b.

A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, and P. Zhan. JANUS-III: Speech-to-speech translation in multiple languages. In *Proc. ICASSP*, volume 1, pages 99–102, 1997.

A. Lee, P.-J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang, et al. Direct speech-to-speech translation with discrete units. *arXiv:2107.05604*, 2021a.

A. Lee, H. Gong, P.-A. Duquenne, H. Schwenk, P.-J. Chen, C. Wang, S. Popuri, J. Pino, J. Gu, and W.-N. Hsu. Textless speech-to-speech translation on real data. *arXiv:2112.08352*, 2021b.

A. Lee, P.-J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang, P. Juan, and W.-N. Hsu. Direct speech-to-speech translation with discrete units. In *Proc. ACL*, pages 3327–3339, 2022.

C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang. MOSNet: Deep learning-based objective assessment for voice conversion. *Proc. Interspeech*, pages 1541–1545, 2019.

N. Morioka, H. Zen, N. Chen, Y. Zhang, and Y. Ding. Residual adapters for few-shot text-to-speech speaker adaptation. *arXiv preprint arXiv:2210.15868*, 2022.

S. Nakamura, K. Markov, H. Nakaiwa, G.-i. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto. The ATR multilingual speech-to-speech translation system. *IEEE Trans. ASLP*, 14(2):365–376, 2006.

D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. *Proc. Interspeech*, pages 2613–2617, 2019.

B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley. Automos: Learning a non-intrusive assessor of naturalness-of-speech. *arXiv:1611.09207*, 2016.

T. Sellam, A. Bapna, J. Camp, D. Mackinnon, A. P. Parikh, and J. Riesa. SQuId: Measuring speech naturalness in many languages. *arXiv:2210.06324*, 2022.

R. Sennrich, B. Haddow, and A. Birch. Improving neural machine translation models with monolingual data. *arXiv:1511.06709*, 2015.

J. Shen, Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu. Non-Attentive Tacotron: Robust and controllable neural TTS synthesis including unsupervised duration modeling. *arXiv:2010.04301*, 2020.

K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu. MASS: Masked sequence to sequence pre-training for language generation. In *Proc. ICML*, pages 5926–5936, 2019.

Y. Tang, H. Gong, N. Dong, C. Wang, W.-N. Hsu, J. Gu, A. Baevski, X. Li, A. Mohamed, M. Auli, et al. Unified speech-text pre-training for speech translation and recognition. In *Proc. ACL*, pages 1488–1499, 2022.

A. van den Oord, O. Vinyals, et al. Neural discrete representation learning. *Proc. NeurIPS*, 30, 2017.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Proc. NeurIPS*, 30, 2017.

W. Wahlster. *Verbmobil: Foundations of speech-to-speech translation*. Springer Science & Business Media, 2013.

C. Wang, A. Wu, and J. Pino. CoVoST 2 and massively multilingual speech-to-text translation. *arXiv:2007.10310*, 2020.

C. Wang, H. Inaguma, P.-J. Chen, I. Kulikov, Y. Tang, W.-N. Hsu, M. Auli, and J. Pino. Simple and effective unsupervised speech translation. *arXiv:2210.10191*, 2022.

K. Wei, L. Zhou, Z. Zhang, L. Chen, S. Liu, L. He, J. Li, and F. Wei. Joint pre-training with speech and bilingual text for direct speech to speech translation. *arXiv:2210.17027*, 2022.

C. Xing, D. Wang, C. Liu, and Y. Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proc. NAACL*, pages 1006–1011, 2015.

C. Zhang, X. Tan, Y. Ren, T. Qin, K. Zhang, and T.-Y. Liu. Uwspeech: Speech to speech translation for unwritten languages. In *Proc. AAAI*, volume 35, pages 14319–14327, 2021.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. ICCV*, pages 2223–2232, 2017.

# A Appendix

## A.1 Monolingual Datasets

Table 5: Monolingual datasets used.

| | U-Conversational Jia et al. [2019a] | S-CV11 Ardila et al. [2020] | CV11 Ardila et al. [2020] |
|---|---|---|---|
| Languages | Es $\leftrightarrow$ En | Es $\leftrightarrow$ En | Es $\leftrightarrow$ En |
| Domain | Synthesized | Synthesized | Read, short-form |
| Sample rate (down-sampled) | 24 kHz (16 kHz) | 24 kHz (16 kHz) | 48 kHz (16 kHz) |
| Spanish hours (utterances) | 350 (379K) | 340 (230k) | 413 (230k) |
| English hours (utterances) | 371 (379k) | 427 (440k) | 708 (440k) |
| Synthesized by | PnG NAT + WaveFit | PnG NAT + WaveFit | Real Speech |

### A.2 Table of hyper-parameters

Table 6: Model hyper-parameters used in the experiments. ("$\times n$": $n$ layers)

| | Conversational | S-CommonVoice 11 | CommonVoice 11 |
|---|---|---|---|
| *Input & Output* | | | |
| Sample rate (Hz) | 16,000 | 16,000 | 16,000 |
| Mel channels | 128 | 128 | 128 |
| Mel lower band (Hz) | 20 | 20 | 20 |
| Mel upper band (Hz) | 8,000 | 8,000 | 8,000 |
| Frame size (ms) | 50.0 | 50.0 | 50.0 |
| Frame step (ms) | 12.5 | 12.5 | 12.5 |
| *SpecAugment* | | | |
| Freq blocks | 2 | 2 | 2 |
| Time blocks | 10 | 10 | 10 |
| Freq block max length ratio | 0.33 | 0.33 | 0.33 |
| Time block max length ratio | 0.05 | 0.05 | 0.05 |
| *Encoder* | | | |
| Conformer dims | $144 \times 16$ | $144 \times 16$ | $144 \times 16$ |
| Attention heads | 4 | 4 | 4 |
| Conv kernal size | 32 | 32 | 32 |
| Subsample factor | 4 | 4 | 4 |
| *Attention (source & target)* | | | |
| Output & Hidden dim | 512 | 512 | 512 |
| Attention heads | 8 | 8 | 8 |
| Dropout prob | 0.2 | 0.2 | 0.2 |
| *Decoder (source & target)* | | | |
| Transformer (dim $\times$ layers) | $512 \times 4$ | $512 \times 4$ | $512 \times 4$ |
| Hidden dims | $512 \times 4$ | $512 \times 4$ | $512 \times 4$ |
| Dropout prob | 0.3 | 0.3 | 0.3 |
| Phoneme embedding dim | 256 | 256 | 256 |
| Label smoothing uncertainty | 0.1 | 0.1 | 0.1 |
| Loss weight | 1.0 | 1.0 | 1.0 |
| *Duration predictor (source & target)* | | | |
| Bi-LSTM (dim $\times$ layers) | $128 \times 2$ | $128 \times 2$ | $128 \times 2$ |
| Loss weight | 1.0 | 1.0 | 10.0 |
| *Synthesizer (source & target)* | | | |
| LSTM dims | $1{,}024 \times 2$ | $1{,}024 \times 2$ | $1{,}024 \times 2$ |
| LSTM zoneout prob | 0.1 | 0.1 | 0.1 |
| Pre-net dims | $128 \times 2$ | $128 \times 2$ | $128 \times 2$ |
| Pre-net dropout prob | 0.5 | 0.5 | 0.5 |
| Post-net (kernel, channels) $\times$ layers | $(5, 512) \times 4 + (5, 128)$ | $(5, 512) \times 4 + (5, 128)$ | $(5, 512) \times 4 + (5, 128)$ |
| Loss weight | 1.0 | 1.0 | 1.0 |
| *WaveFit vocoder* | | | |
| Iterations | 5 | 5 | 5 |
| UBlock upsampling factors | [5, 5, 2, 2, 2] | [5, 5, 2, 2, 2] | [5, 5, 2, 2, 2] |
| STFT loss resolutions | 3 | 3 | 3 |
| Hann win size, frame shift, FFT size res 1 | [160, 32, 512] | [160, 32, 512] | [160, 32, 512] |
| Hann win size, frame shift, FFT size res 2 | [400, 80, 1024] | [400, 80, 1024] | [400, 80, 1024] |
| Hann win size, frame shift, FFT size res 3 | [800, 160, 2048] | [800, 160, 2048] | [800, 160, 2048] |
| Multi-period discriminator | Kong et al. [2020] | Kong et al. [2020] | Kong et al. [2020] |
| Multi-period discriminator loss weight | 1.0 | 1.0 | 1.0 |
| *Training* | | | |
| Optimizer | Adam | Adam | Adam |
| Learning rate schedule | Vaswani et al. [2017] | Vaswani et al. [2017] | Vaswani et al. [2017] |
| Learning rate (peak) | $1.3 \times 10^{-3}$ | $1.3 \times 10^{-3}$ | $1.3 \times 10^{-3}$ |
| Warm-up steps | 20K | 40K | 20K |
| Batch size | 512 | 512 | 512 |
| $L^2$ regularization weight | $10^{-6}$ | $10^{-6}$ | $10^{-6}$ |
| MUSE loss weight | 100000.0 | 1000.0 | 1.0 |