Translatotron 2: High-quality direct speech-to-speech translation with voice preservation

Ye Jia [1] **Michelle Tadmor Ramanovich** [1] **Tal Remez** [1] **Roi Pomerantz** [1]

Abstract

We present *Translatotron 2*, a neural direct speech-to-speech translation model that can be trained end-to-end. Translatotron 2 consists of a speech encoder, a linguistic decoder, an acoustic synthesizer, and a single attention module that connects them together. Experimental results on three datasets consistently show that Translatotron 2 outperforms the original Translatotron by a large margin on both translation quality (up to +15.5 BLEU) and speech generation quality, and approaches the same of cascade systems. In addition, we propose a simple method for preserving speakers' voices from the source speech to the translation speech in a different language. Unlike existing approaches, the proposed method is able to preserve each speaker's voice on speaker turns without requiring for speaker segmentation. Furthermore, compared to existing approaches, it better preserves speaker's privacy and mitigates potential misuse of voice cloning for creating spoofing audio artifacts.

1. Introduction

Speech-to-speech translation (S2ST) is highly beneficial for breaking down communication barriers between people not sharing a common language. Conventional automatic S2ST systems are composed of a cascade of three components: automatic speech recognition (ASR), text-to-text machine translation (MT), and text-to-speech (TTS) synthesis (Lavie et al., 1997; Wahlster, 2000; Nakamura et al., 2006). In the past few years, direct speech-to-text translation (ST) is rapidly emerging, and has outperformed the cascade of ASR and MT (Weiss et al., 2017; Jia et al., 2019a; Di Gangi et al., 2019; McCarthy et al., 2020; Ansari et al., 2020; Wang et al., 2021b; Anastasopoulos et al., 2021), which makes the cascade of ST and TTS as S2ST feasible (Jia et al., 2019b).

Recently, works on S2ST without relying on intermediate text representation are emerging, such as end-to-end direct S2ST (Jia et al., 2019b; Kano et al., 2021) and cascade S2ST based on discrete speech representation (Tjandra et al., 2019; Zhang et al., 2021; Lee et al., 2022; 2021a; Ma et al., 2021). Compared to text-centric cascaded systems, although such approaches require parallel S2ST data, which is scarce, they have the potential advantages of: 1) Preserving paralinguistic and non-linguistic information during translation, such as speaker's voice (Jia et al., 2019b), emotion and prosody; 2) Supporting languages without written form, or being able to be trained without transcription of speech (Tjandra et al., 2019; Zhang et al., 2021; Lee et al., 2022; 2021a); 3) Reduced computational requirements and lower inference latency (Lee et al., 2022); 4) Avoiding error compounding across sub-systems (Jia et al., 2022); 5) Easier on handling contents that do not need to be translated, such as names and proper nouns (Jia et al., 2019b).

Among these works, Translatotron (Jia et al., 2019b) is the first model that is able to directly translate speech in one language to speech in another language. It obtained reasonable translation quality and high naturalness in the predicted translation speech, and is able to preserve speakers' voices during the speech translation. However, the translation quality from Translatotron still underperforms cascade baselines by a large margin, and the translation speech it produces suffers from over-generation issues, such as babbling and long pause. Such weaknesses make this model not yet practical for production. Nevertheless, it remains the state-of-the-art of end-to-end direct S2ST.

In this paper, we first tackle the performance gap between end-to-end direct S2ST and cascade S2ST. We propose *Translatotron 2*, a novel direct S2ST model that is able to be trained end-to-end. We conduct experiments on three S2ST datasets, including multilingual S2ST. The results consistently suggest that Translatotron 2 significantly outperforms Translatotron in terms of both translation quality (up to +15.5 BLEU) and speech generation quality, and approaches the same of cascade S2ST. When a simple data augmentation *ConcatAug* is used, the translation quality gap on the Fisher Spanish-English corpus (Post et al., 2013) is

[1]Google Research. Correspondence to: Ye Jia <jiaye@google.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

reduced from 16.4 to 0.4 BLEU. These results are the first time that end-to-end direct S2ST approaches cascade S2ST.

In addition, we propose a simple method for preserving speakers' voices during S2ST without relying on any speaker representation (ID or embedding). The proposed method enables Translatotron 2 to preserve each speaker's voice on speaker turns without requiring for speaker separation, which is the first of its kind. Furthermore, compared to existing approaches of voice preservation, the proposed method better preserves speaker's privacy (Pathak & Raj, 2012) and mitigates potential misuse of voice cloning for creating spoofing audio artifacts.

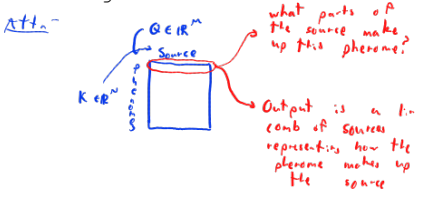
Audio samples from Translatotron 2 are available online.[1]

2. Related works

S2ST Until very recently, automatic S2ST systems are typically composed of a cascade of ASR, MT, and TTS components (Lavie et al., 1997; Wahlster, 2000; Nakamura et al., 2006; ITU, 2016). Translatotron (Jia et al., 2019b) is the first direct S2ST model, which is a sequence-to-sequence model trained end-to-end in a multi-objective task. It has shown reasonable translation quality and speech naturalness, but still underperformed a baseline of ST → TTS cascade by a large margin. It also demonstrated the capacity of preserving speakers' voices during the translation, by leveraging a speaker encoder separately trained in a speaker verification task (Wan et al., 2018; Jia et al., 2018).

A few recent works proposed cascade S2ST systems using learned discrete speech representation as the intermediate representation instead of text or phoneme. Tjandra et al. (2019) introduced such an S2ST system that first translated the source speech into a discrete representation of the target speech which was predicted from a separately trained VQ-VAE (Oord et al., 2017), then constructed the target speech spectrogram from the discrete representation using the VQ-VAE decoder. Zhang et al. (2021) additionally trained the VQ-VAE jointly with a supervised phoneme recognition objective in different languages. Lee et al. (2022; 2021a) used a separately trained vocoder to directly predict waveform from the discrete representation without relying on spectrogram; for the best performance, this vocoder included a duration predictor, akin to a generative TTS model. All these works require multiple components being trained in multiple steps, but are not able to be trained end-to-end. Another potential limitation of such an approach is that it may not be effective in preserving paralinguistic and non-linguistic information. Oppositely, it can be desired that such variation be removed in the discrete representation (Lee et al., 2021a).

[1] https://google-research.github.io/lingvo-lab/translatotron2/



Kano et al. (2021) introduced an end-to-end 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 end-to-end model. It requires pre-trained ASR, MT, and TTS models, and the end-to-end model itself has to be trained in multiple steps.

While most of these works conducted experiments using synthetic datasets with translation speech in a clean single speaker's voice, Jia et al. (2019b); Lee et al. (2021a) conducted experiments using multi-speaker human recordings.

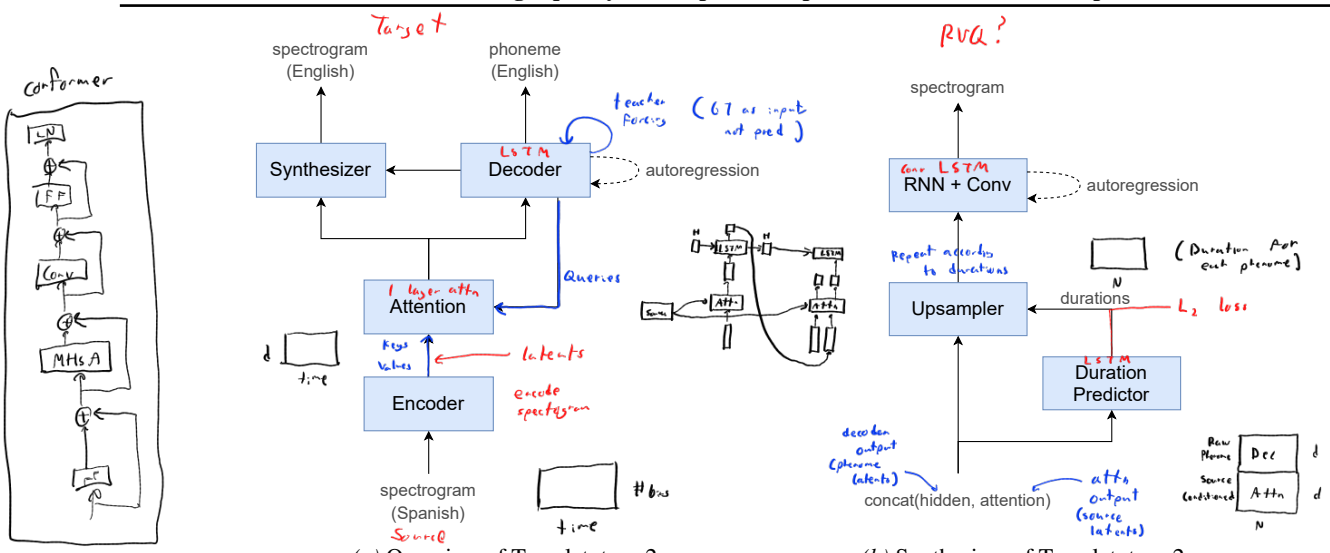
Although these recent works generated speech translation in novel ways without relying on TTS subsystems, only a few of them (Jia et al., 2019b; Lee et al., 2022) have evaluated the perceptual quality (e.g. naturalness) of the produced speech translation, which is critical to S2ST (Wagner et al., 2019; Salesky et al., 2021), with the rest focused only on the translation quality.

TTS Translatotron uses a decoder similar to the Tacotron 2 TTS model (Shen et al., 2018), which is an attention-based autoregressive decoder. Due to the flexibility of the attention mechanism, they both suffer from robustness issues such as over-generation. Recent TTS models such as FastSpeech (Ren et al., 2019; 2021) and Non-Attentive Tacotron (NAT) (Shen et al., 2020) demonstrated that replacing the attention module with a duration-based upsampler yields more robust synthesized speech, as quantitatively evaluated at a large scale in Shen et al. (2020). The synthesizer component in this work resembles these works.

Voice conversion and anti-spoofing The performance of voice conversion has progressed rapidly in the recent years, and is reaching a quality that is hard for automatic speaker verification (ASV) systems to detect (Yi et al., 2020). ASVspoof 2019 (Todisco et al., 2019; Wang et al., 2020) found that it was challenging to detect spoof audios generated from a zero-shot voice cloning TTS model (Jia et al., 2018), which was followed by the original Translatotron for preserving speakers' voices during S2ST. Such progress poses concerns on related techniques being misused for creating spoofing artifacts. We propose a new voice preservation method for S2ST with the motivation of avoiding such potential misuse.

3. Translatotron 2

We designed the architecture of Translatotron 2 to address three performance bottlenecks existing in the original Translatotron: 1) The utilization of the auxiliary textual supervision during training is suboptimal, namely, the attention alignment learned by the auxiliary ST task does not directly contribute to the main S2ST task; 2) The challenge posed by modeling the translation alignment between two very



(a) Overview of Translatotron 2. *(b)* Synthesizer of Translatotron 2.

Figure 1: A Translatotron 2 model that translates Spanish speech into English speech.

long spectrogram sequences using the attention mechanism; 3) Attention-based speech generation is known to suffer from robustness issues such as over-generation and under-generation (Shen et al., 2020; Ren et al., 2019; He et al., 2019; Zheng et al., 2019; Battenberg et al., 2020).

We addressed these bottlenecks by designing a novel S2ST model architecture composed of a speech encoder, a linguistic decoder, an acoustic synthesizer, and a single attention module connecting them together (Figure 1a). The model is jointly trained with a speech-to-speech translation objective and a speech-to-phoneme translation objective.

The following subsections describe each component of Translatotron 2. Note that as shown in the ablation studies in Sec. 5.5.1, while the specific architectural choices of these components help the performance of Translatotron 2, the primary improvement comes from the high-level architecture rather than the choice of each individual component.

3.1. Speech encoder

The encoder of Translatotron 2 takes the mel-spectrogram of the source speech as the input, and produces a hidden representation which encodes both linguistic and acoustic information from the source speech. We use Conformer (Gulati et al., 2020) as the architecture of the encoder. It first subsamples the input mel-spectrogram with a convolutional layer, and then processes it with a stack of Conformer blocks. Each Conformer block is composed of a feed-forward layer, a self-attention layer, a convolution layer, and a second feed-forward layer. SpecAugment (Park et al., 2019) is applied at the training time as data augmentation.

3.2. Linguistic decoder

The autoregressive decoder is responsible for producing linguistic information in the translation speech. It takes the encoder output through the attention module, and predicts a phoneme sequence corresponding to the translation speech. We use an LSTM stack (Hochreiter & Schmidhuber, 1997) as the decoder, assisted with regularization including Zoneout (Krueger et al., 2017) and label smoothing (Szegedy et al., 2016). The combination of the encoder, the decoder, and the attention module is similar to a typical ST model, except that it predicts phonemes instead of subword tokens.

3.3. Acoustic synthesizer

The synthesizer is responsible for acoustic generation of the translation speech. It takes the intermediate output from the decoder (before a final projection and softmax for phoneme prediction), as well as the context output from the attention as its input, and generates a mel-spectrogram corresponding to the translation speech. It is similar to the decoders in typical neural TTS models. The predicted mel-spectrogram can be converted to waveform using an estimation algorithm such as Griffin & Lim (1984) or a neural vocoder such as WaveRNN (Kalchbrenner et al., 2018).

We use the duration-based autoregressive synthesizer from the NAT (Shen et al., 2020) TTS model (Figure 1b). It first predicts durations for each elements in the input sequence, then temporally upsamples the input sequence based on the predicted durations. After that, an LSTM stack is used for generating the target spectrogram without altering the sequence length. A final residual convolutional block further refines the generated spectrogram. Unlike in NAT, we do not supervise the duration prediction on per-phoneme duration

labels, to avoid additional requirement on the training data. Instead, an L^2 loss on the total predicted duration of the entire utterance is used (similar to the "naïve approach" of unsupervised duration modeling in Shen et al. (2020)).

3.4. A single attention

It is critical that Translatotron 2 utilizes a single attention module for both the linguistic decoder and the acoustic synthesizer. This attention models both linguistic and acoustic alignments between the source and the target speeches. A multi-head attention (Vaswani et al., 2017) is used.

The queries to this attention are from the linguistic decoder. As a result, unlike in the original Translatotron, this attention does not directly model the translation alignment between two very long spectrogram sequences. Instead, it models the alignment between a source spectrogram sequence and a shorter target phoneme sequence, which is significantly easier to learn.

In the meantime, the attention provides acoustic information from the source speech to the synthesizer, summarized at per-phoneme level. Such summarized acoustic information is not only usually sufficient for speech generation but also eases the duration prediction per-phoneme because it is of the same granularity. Because a single attention is used, the linguistic and acoustic information seen by the synthesizer is synchronized temporally. Such synchronization enables Translatotron 2 to preserve paralinguistic and non-linguistic information at fine granularity, such as preserving each speaker's voice on speaker turns (Sec. 4.2).

Although the synthesizer takes attention output as part of its input, the attention is not driven (i.e. queried) by the synthesizer. As a result, while it benefits from the attention on obtaining aligned acoustic information from the source speech, it does not suffer from robustness issues as in typical attention-based speech synthesis models.

4. Voice preserving

The original Translatotron (Jia et al., 2019b) demonstrated the capacity of preserving source speakers' voices in the translation speech, by conditioning its synthesizer on a speaker embedding generated from a separately trained speaker encoder. In fact, it is capable of generating the translation speech in a different speaker's voice, as long as a clip of the target speaker's recording is used as the reference audio to the speaker encoder, or the embedding of the target speaker is directly available. While this is impressively powerful, it can potentially be misused for generating spoofing audio with arbitrary content, posing a concern for production deployment.

To mitigate such risks, we propose a new approach for preserving speaker's voice during S2ST, so that the trained models are restricted to preserving the source speaker's voice, but not able to generate speech in a different speaker's voice. In addition, this approach enables S2ST models to preserve each speaker's voice on input speech with speaker turns, without requiring for speaker segmentation.

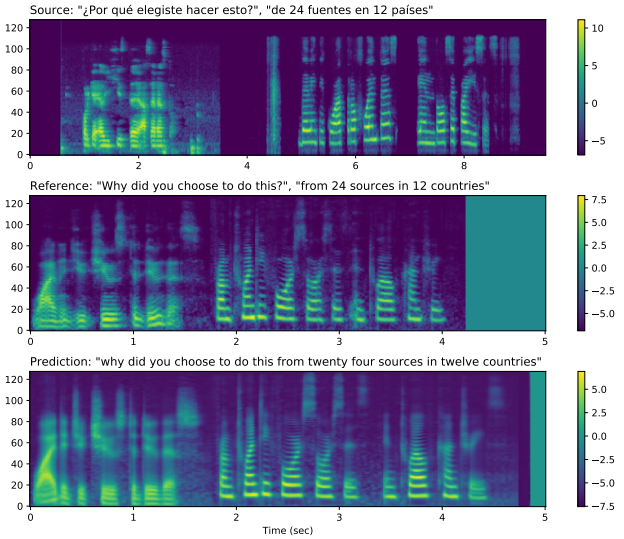


Figure 2: Sample mel-spectrograms on input with speaker turns. The input speech includes an utterance from a male speaker followed by another utterance from a female speaker. Translatotron 2 preserves the voices of each speaker in the translation speech.

4.1. Training-time voice transferring

In our approach, the key to restrict S2ST models to voice preservation but not arbitrary voice cloning (from a different speaker) is to move the powerful voice transferring to only happen at the training time (or the training data preparation time) but not the inference time. In contrast, it happens at both the training time and the inference time in the original Translatotron.

To preserve speakers' voices across translation, we train S2ST models on parallel utterances with the same speaker's voice on both sides. Such a dataset with human recordings on both sides is extremely difficult to collect, because it requires a large number of fluent bilingual speakers. Instead, we use a TTS model capable of cross-lingual voice cloning to synthesize such training targets.

We modified the PnG NAT (Jia et al., 2021; Shen et al., 2020) TTS model by incorporating a separately trained speaker encoder (Wan et al., 2018) in the same way as in Jia et al. (2018), and trained it on the LibriTTS corpus (Zen et al., 2019). The resulting TTS model is capable of zero-shot voice transferring, but synthesizes in a better quality and

Table 1: Datasets for experiments with translation speech in a single-speaker's voice.

	Conversational (Jia et al., 2019a)	Fisher Es-En (Post et al., 2013)	CoVoST 2 (Wang et al., 2021a)
Languages	es→en	es→en	es, fr, de, ca → en
Domain	Read, short-form	Telephone conversation	Read, short-form
Source sample rate	16-48 kHz	8 kHz	48 kHz
Utterance pairs	979k	120k	321k
Source hours	1,400	127	476
Target hours	619	96	296
Target synthesized by	Tacotron 2 + Griffin-Lim	Parallel WaveNet	PnG NAT + WaveRNN

more robust than Jia et al. (2018).[2] We used this model to synthesize translation speech in the source speaker's voice as the training targets in our experiments. Other TTS models capable of cross-lingual voice modeling, such as Zhang et al. (2019); Chen et al. (2019); Xin et al. (2021), could also be utilized.

4.2. Speaker turns

Because the single attention module provides linguistic and acoustic information temporally synchronized (Sec. 3.4), Translatotron 2 is theoretically capable of voice preservation in complicated scenarios such as speaker turns. However, proper training data with speaker turns is required to demonstrate such capacity, which is difficult to obtain. We propose a simple data augmentation to enable such training.

ConcatAug To enable direct S2ST models to preserve each speaker's voice for input with speaker turns, we augmented the training data by randomly sampling pairs of training examples and concatenating the source speech, the target speech, and the target phoneme sequences to construct new training examples. The resulting new examples contain two speakers' voices in both the source and the target speech, which enables the model to learn on examples with speaker turns. See Figure 2 for an example of such concatenation and the prediction from Translatotron 2 on it.

Such augmentation does not only enable the model to learn voice retention on speaker turns, but also increases the diversity of the speech content as well as the complexity of the acoustic conditions in the training examples, which may further improve the translation quality of the model, especially on small datasets (Sec. 5.1). Narayanan et al. (2019) uses a similar augmentation but in a more complicated fashion, for improving ASR performance on multi-speaker inputs.

5. Experiments

We conducted experiments on three datasets, including two Spanish→English datasets and a multilingual→English dataset. All datasets use TTS synthesized target speech in 24 kHz sample rate. The phonemes used at training time were converted from the transcripts using a proprietary G2P system. See Table 1 for the details of each dataset. We evaluated the translation quality, naturalness and robustness of the produced translation speech, as well as speaker similarity for voice preservation. All models were implemented using the Lingvo framework (Shen et al., 2019). A comprehensive table of hyper-parameters is available in Appendix A.

5.1. Translation quality

To evaluate the translation quality, we used the same two datasets as in Jia et al. (2019b), both of which have translation speech in a single female speaker's voice. Following Jia et al. (2019b), the translation quality is measured by BLEU on ASR transcription from the translation speech (in lowercase, excluding punctuation marks except for apostrophes), compared to reference translation text. Because ASR makes errors, such BLEU can be thought of as a lower bound of the translation quality. We used an ASR model from Park et al. (2020), trained on LibriSpeech (Panayotov et al., 2015) and LibriLight (Kahn et al., 2020) corpora. For a fair comparison, we retrained the baseline Translatotron models and evaluated them using the same ASR model. The same ST→TTS cascade S2ST baselines from Jia et al. (2019b) were used and re-evaluated, which were composed of strong ST models and a Tacotron 2 TTS model. The predicted mel-spectrogram is converted to waveform using the Griffin-Lim algorithm for all models.

As shown in Table 2, the translation quality from Translatotron 2 outperformed the original Translatotron by +15.5 BLEU on Fisher Es-En and +5.2 BLEU on Conversational. Applying ConcatAug further improved the performance on the smaller Fisher Es-En dataset by +0.5 BLEU. These improvements narrowed down the performance gap between end-to-end direct S2ST and cascade S2ST from 16.4 / 8.4 down to 0.4 / 3.7 BLEU on the two datasets respectively.

5.2. Speech naturalness

The naturalness of the predicted translation speech is evaluated by subjective listening test, reporting 5-scale mean opin-

[2] A detailed description of this zero-shot voice transferring TTS model is available in our follow-up work (Jia et al., 2022).

Table 2: Performance of S2ST in a single speaker's voice. BLEU were computed with 1 reference for the Conversational test set, and with 4 references for the Fisher test set.

	Conversational			Fisher Es-En		
	BLEU	MOS	UDR (%)	BLEU	MOS	UDR (%)
End-to-end direct S2ST:						
Translatotron 2	**55.6**	**4.21 ± 0.06**	0.16	42.4	**3.98 ± 0.08**	**0.07**
+ ConcatAug	55.1	4.19 ± 0.06	**0.13**	**42.9**	3.79 ± 0.09	0.14
Translatotron	50.4	4.15 ± 0.07	0.69	26.9	3.70 ± 0.08	0.48
Cascade (ST → TTS)	58.8	4.31 ± 0.06	0.21	43.3	4.04 ± 0.08	0.13
Reference (synthetic)	81.9	3.37 ± 0.09	0.43	88.6	3.95 ± 0.07	0.07
Discrete representation-based cascade S2ST:						
Zhang et al. (2021) (trained w/o text)	-	-	-	9.4	-	-
Lee et al. (2022) (trained w/ text)	-	-	-	39.9	3.41 ± 0.14	-

ion scores (MOS) with 95% confidence interval on 1,000 randomly sampled predictions. A WaveRNN-based neural vocoder was used for converting the mel-spectrograms predicted from S2ST models to waveforms.

As shown in Table 2, the naturalness of the translation speech predicted from Translatotron 2 is significantly better than from the original Translatotron, and is on-par with or very close to the cascade systems which used one of the state-of-the-art TTS models, Tacotron 2, for synthesizing translation speech from text.

Consistent with Jia et al. (2019b), despite that the training targets in the Conversational dataset is synthesized with a lower quality Griffin-Lim vocoder, the trained S2ST model is able to produce translation speech in significantly higher naturalness when a higher quality neural vocoder is used at inference time.

5.3. Speech robustness

We specifically evaluated the robustness issue of over-generation in the predicted translation speech, such as babbling or long pause, measured by unaligned duration ratio (UDR) (Shen et al., 2020) with a 1-second threshold.[3] The ASR transcription from the translation speech is used for alignment, using a confidence islands-based forced alignment model (Chiu et al., 2018).

As can be seen from Table 2, the UDR from Translatotron 2 is about 7 and 4 times lower than from Translatotron on the Fisher Es-En and Conversational datasets, respectively. It is even about 3 times lower than the training targets from the Conversational set, while is about the same as the training targets from Fisher Es-En. This can be explained by the fact that the training targets in the Conversational set were synthesized by the Tacotron 2 TTS model, which by itself suffered from over-generation, while the same in Fisher Es-En were synthesized by a more robust Parallel WaveNet (Oord et al., 2018) TTS model (see Table 1). The results suggest that Translatotron 2 drastically improves robustness than Translatotron, and is also robust to a small ratio of disfluency in the training targets.

[3]Under-generation (i.e. WDR from Shen et al. (2020)) does not apply because of the nature of translation. Related errors are reflected in the BLEU evaluation.

Table 3: S2ST performance with voice preservation on Conversational dataset. Speaker similarity MOS is evaluated between Spanish source speech and English translation speech. (Numbers not directly comparable to Table 2 because of dataset differences.)

	BLEU	Naturalness	Similarity
Proposed:			
Translatotron 2	**57.3**	**3.24 ± 0.08**	**2.33 ± 0.08**
+ ConcatAug	56.8	2.94 ± 0.08	2.12 ± 0.07
Translatotron	48.5	2.55 ± 0.09	2.30 ± 0.07
+ ConcatAug	51.3	2.76 ± 0.09	2.19 ± 0.07
Reference (synthetic)	81.3	3.40 ± 0.08	2.55 ± 0.07
Jia et al. (2019b):			
Translatotron	36.2	3.15 ± 0.08	1.85 ± 0.06
Reference (human)	59.9	4.10 ± 0.06	-

5.4. Voice preservation

To evaluate the ability of preserving speakers' voices while translating their speeches from one language to another, we augmented the Conversational dataset by synthesizing target speech using a voice-transferring TTS model as described in Sec. 4.1. Examples with source speech shorter than 1 second were discarded for the stability of voice transferring. The result dataset contains parallel utterances with similar voices on both sides. S2ST models were trained on this dataset without any explicit conditioning on speaker embeddings or IDs (i.e. no speaker encoder for the original Translatotron). Following Jia et al. (2019b), we reduced the pre-net dimension of the synthesizer to 16 to encourage it to infer voice information from the encoder output instead of from the teacher-forcing inputs.

Table 4: Voice preservation performance on speaker turns. Speaker similarity MOS between the leading/trailing 1.6-second segment from the English translation speech and the entire 1st/2nd source speaker's Spanish speech is reported. (↑ / ↓: higher/lower values are better.)

	1st source speaker		2nd source speaker	
	Leading seg. ↑	Trailing seg. ↓	Leading seg. ↓	Trailing seg. ↑
Translatotron 2	2.22 ± 0.07	2.15 ± 0.07	2.04 ± 0.07	2.00 ± 0.07
+ ConcatAug	$\mathbf{2.44 \pm 0.07}$	1.82 ± 0.07	$\mathbf{1.76 \pm 0.07}$	$\mathbf{2.51 \pm 0.08}$
Translatotron	1.87 ± 0.06	1.90 ± 0.07	2.06 ± 0.07	2.05 ± 0.07
+ ConcatAug	2.18 ± 0.07	$\mathbf{1.71 \pm 0.06}$	1.93 ± 0.07	2.35 ± 0.07
Reference (synthetic)	2.58 ± 0.08	1.62 ± 0.06	1.83 ± 0.07	2.44 ± 0.07

5-point subjective MOS on both naturalness and speaker similarity was evaluated with 1,000 random samples or pairs of samples from the test set, reported with 95% confidence interval. As Table 3 shows, when the proposed approach for voice preservation was used, both Translatotron 2 and Translatotron obtained about the same speaker similarity MOS as the original Translatotron but significantly better translation quality. Translatotron 2 further outperformed Translatotron in terms of translation quality and speech naturalness, which is consistent with the experimental results for translating in a single speaker's voice (Sec. 5.1, 5.2). It is worth to note that the speaker similarity from S2ST models is capped by the same of the training targets, which by itself is limited. This can be partially due to the performance of the voice-transferring TTS model used for synthesizing the training targets, and partially due to the fact that cross-lingual speaker similarity evaluation is more challenging to raters (some rating comments are purely based on language difference), as also observed in Zhang et al. (2019). Obtaining better quality training targets, such as human recordings instead of synthesized speech, may further improve the performance of voice preservation with the proposed approach.

5.4.1. Speaker turns

Speaker similarity evaluation with speaker turns on entire translation speech is challenging because it would require speaker separation on both source and target speeches. The content re-ordering during translation and translation errors would also add extra difficulty. We approximated by considering the leading/trailing short segments in the translation speech as corresponding to each of the two speakers in the source speech with a single speaker turn.

We trained Translatotron 2 and Translatotron on the dataset described in Sec. 5.4, with half of the training examples augmented by ConcatAug. The evaluation set was artificially constructed in a similar way by applying ConcatAug, so that each utterance contains two speakers' voices. We evaluated subjective speaker similarity MOS between the two entire source utterances before ConcatAug and the leading/trailing 1.6-second segments from the translation speech. Evaluation examples with target speech shorter than 2 seconds before ConcatAug were discarded.

As can be seen from Table 4, the impact of ConcatAug is consistent on Translatotron 2 and Translatotron. When ConcatAug was not used during training, for each source speaker, the similarity compared to the leading/trailing segment from the translation speech was about the same; and for each segment in the translation speech, the speaker similarity compared to the first/second source speaker was also close. This suggests that the translation speech imitated both source speakers at the same time regardless of the speaker turn. When ConcatAug was used, both models obtained significantly higher speaker similarity on matched pairs than mismatched pairs, indicating that the models successfully separated two speakers and preserved voices for each of them respectively. It can also be seen that Translatotron 2 obtained significantly higher speaker similarity than Translatotron on matched pairs, indicating the effectiveness of Translatotron 2.

Such quantitative evaluation cannot reflect how the predicted translation speech transits from one speaker's voice to another speaker's. Listening to audio samples (available online) verified that the voice changed instantly on sentence boundaries without blurry, rather than a smoothed transition. A sample of S2ST on such a speaker turn from Translatotron 2 is visualized in Figure 2.

While ConcatAug enables S2ST models to preserve speakers' voices on speaker turns and improves translation quality on small datasets, it may negatively impact the speech naturalness and speaker similarity on models with strong performance, as shown in Table 2 and Table 3. It may be explained by the fact that the augmented utterances sound less natural and may involve abrupt change in volume and background noise on the artificial speaker turns. This suggests headroom for improvement.

5.5. Multilingual S2ST

We also conducted experiments to evaluate the performance of multilingual X→En S2ST. We trained Translatotron 2 and Translatotron on the 4 high-resource language pairs from the CoVoST 2 corpus (Wang et al., 2021a), using

Table 5: Ablation studies of multilingual X→En S2ST on 4 high-resource language pairs from CoVoST 2, measured by BLEU on ASR transcription from the translation speech. + / − indicates using or replacing a component (see Sec. 5.5.1).

	fr	de	es	ca
Translatotron 2	**27.0**	**18.8**	**27.7**	**22.5**
− SpecAugment	25.9	17.9	25.9	21.8
− Conformer encoder	26.4	18.1	26.4	21.8
− NAT synthesizer	26.9	18.3	27.0	22.0
Translatotron (w/ SpecAugment)	17.7	9.9	17.7	13.1
+ Conformer encoder	18.9	10.8	18.8	13.9
+ NAT synthesizer	4.0	2.1	3.5	2.5
ST (Wang et al., 2021a)	27.0	18.9	28.0	23.9
Reference (synthetic)	82.1	86.0	85.1	89.3

TTS synthesized target speech in a single female speaker's voice.[4] The original Common Voice (Ardila et al., 2020) data split instead of the CoVoST 2 data split was followed. The models were not explicitly conditioned on languages. For a fair comparison, both models used SpecAugment, but did not use auxiliary supervision from the source phonemes.

The translation quality as measured by BLEU on ASR transcription from the translation speech is shown in the first rows of each block in Table 5. Translatotron 2 outperformed Translatotron by +9.4 BLEU on average on the 4 language pairs. Although the BLEU scores are not directly comparable between S2ST and ST (because of ASR transcription and BLEU calculation difference), the close numbers suggest that Translatotron 2 obtained translation quality comparable to the baseline ST model.

5.5.1. Ablation studies

To understand the importance of each component in Translatotron 2, we conducted ablation studies on this multilingual X→En dataset. All models in the ablation used the same input and output features, SpecAugment settings, and learning rate schedules (detailed in Appendix A). No auxiliary supervision from source text was used. For models not using a Conformer encoder, we first applied the same 4× temporal subsampling as in the Conformer encoder, then used a 256×8 bidirectional LSTM stack to encoder the subsampled features. The number of parameters in this LSTM encoder is close to the same in the Conformer encoder. For the Translatotron model using a NAT synthesizer, the same hyperparameters as in Translatotron 2 were used. For Translatotron 2 not using a NAT synthesizer, a non-autoregressive Conformer synthesizer (Sec. 5.6) was used. All the rest hyperparameters followed Appendix A for Translatotron 2, and

[4]An expanded version of this dataset is released as the CVSS corpus (Jia et al., 2022). However, the results are not directly comparable because of different data splits and reference texts.

Table 6: Ablation studies on Conversational dataset using an autoregressive RNN + Conv synthesizer and a non-autoregressive Conformer synthesizer.

Synthesizer	BLEU	Naturalness
RNN + Conv	55.6	4.21 ± 0.06
Conformer	54.5	3.61 ± 0.09

followed the Conversational model from Jia et al. (2019b) for Translatotron. All models were trained for 200K steps with a batch size of 768. The checkpoints for evaluation were picked by the best average BLEU on 4 language pairs on the validation set.

The results are shown in Table 5. As can be seen, while the use of Conformer, SpecAugment, and NAT synthesizer helps the performance of Translatotron 2, replacing them with alternative architectural choices or removing SpecAugment only reduced the performance by a small degree (<2 BLEU). Similarly, directly using these components in Translatotron does not bring its performance close to Translatotron 2. These results suggest that the improvements of Translatotron 2 primarily comes from the high-level architectural design which addressed the performance bottlenecks existing in Translatotron (Sec. 3), rather than the choices of each individual component.

5.6. Non-autoregressive synthesizer

It is tempting to use a non-autoregressive architecture for the synthesizer of Translatotron 2, which may significantly reduce its inference latency, similar to recent works on non-autoregressive TTS (Ren et al., 2019; 2021; Guo et al., 2021; Lee et al., 2021b; Elias et al., 2021b;a). We experimented with using a 6-layer Conformer synthesizer (Guo et al., 2021) with a dimension of 512 and 8 attention heads on both Conversational and CoVoST 2 datasets.

As can be seen from Table 5 and 6, using a Conformer-based non-autoregressive synthesizer obtained comparable translation quality to using an autoregressive NAT synthesizer (with BLEU on ASR transcription up to 1.1 BLEU lower). However, it caused a significant regression on the naturalness of the predicted translation speech, which is consistent with the observation in TTS in Shen et al. (2020); Peng et al. (2020); Hwang et al. (2021), etc., suggesting more exploration is needed on this direction.

6. Conclusion

We proposed *Translatotron 2*, a neural direct S2ST model that can be trained end-to-end. Experimental results on three datasets consistently suggest that Translatotron 2 outperforms the original Translatotron by a large margin on both translation quality (up to +15.5 BLEU) and speech

generation quality, and approaches cascade S2ST.

In addition, we proposed a simple method for preserving speakers' voices from the source speech to the translation speech in a different language. Unlike existing approaches, the proposed method is able to preserve each speaker's voice on speaker turns without requiring for speaker segmentation. Furthermore, compared to existing approaches, it better preserves speaker's privacy and mitigates potential misuse of voice cloning for creating spoofing audio artifacts.

Future works include extending Translatotron 2 to support simultaneous translation, cross-lingual prosody transfer, unwritten languages, and further quality improvement by utilizing self-supervised pre-training (Baevski et al., 2020; Wang et al., 2021b) and weakly supervised data (Jia et al., 2019a).

Acknowledgements

The authors would like to thank Chung-Cheng Chiu, Quan Wang, Heiga Zen, Ron J. Weiss, Wolfgang Macherey, Yu Zhang, Yonghui Wu, Hadar Shemtov, Ruoming Pang, Nadav Bar, Michael Hassid, and the rest of the Google Research team for helpful discussions and previous work on data preparation.

References

Anastasopoulos, A., Bojar, O., Bremerman, J., Cattoni, R., Elbayad, M., Federico, M., Ma, X., Nakamura, S., Negri, M., Niehues, J., et al. Findings of the IWSLT 2021 evaluation campaign. In *International Conference on Spoken Language Translation (IWSLT)*, 2021.

Ansari, E., Axelrod, A., Bach, N., Bojar, O., Cattoni, R., Dalvi, F., Durrani, N., Federico, M., Federmann, C., Gu, J., et al. Findings of the IWSLT 2020 evaluation campaign. In *International Conference on Spoken Language Translation (IWSLT)*, 2020.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. Common Voice: A massively-multilingual speech corpus. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2020.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Battenberg, E., Skerry-Ryan, R., Mariooryad, S., Stanton, D., Kao, D., Shannon, M., and Bagby, T. Location-relative attention mechanisms for robust long-form speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

Chen, M., Chen, M., Liang, S., Ma, J., Chen, L., Wang, S., and Xiao, J. Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding. In *Proc. Interspeech*, 2019.

Chiu, C.-C., Tripathi, A., Chou, K., Co, C., Jaitly, N., Jaunzeikare, D., Kannan, A., Nguyen, P., Sak, H., Sankar, A., Tansuwan, J., Wan, N., Wu, Y., and Zhang, X. Speech recognition for medical conversations. In *Proc. Interspeech*, 2018.

Di Gangi, M. A., Negri, M., and Turchi, M. One-to-many multilingual end-to-end speech translation. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

Elias, I., Zen, H., Shen, J., Zhang, Y., Jia, Y., Skerry-Ryan, R., and Wu, Y. Parallel Tacotron 2: A non-autoregressive neural TTS model with differentiable duration modeling. In *Proc. Interspeech*, 2021a.

Elias, I., Zen, H., Shen, J., Zhang, Y., Jia, Y., Weiss, R., and Wu, Y. Parallel Tacotron: Non-autoregressive and controllable TTS. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021b.

Griffin, D. and Lim, J. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech*, 2020.

Guo, P., Boyer, F., Chang, X., Hayashi, T., Higuchi, Y., Inaguma, H., Kamo, N., Li, C., Garcia-Romero, D., Shi, J., et al. Recent developments on espnet toolkit boosted by conformer. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5874–5878, 2021.

He, M., Deng, Y., and He, L. Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural TTS. In *Proc. Interspeech*, 2019.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hwang, M.-J., Yamamoto, R., Song, E., and Kim, J.-M. TTS-by-TTS: TTS-driven data augmentation for fast and high-quality speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6598–6602, 2021.

ITU. ITU-T F.745: Functional requirements for network-based speech-to-speech translation services, 2016. International Telecommunication Union.

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I. L., and Wu, Y. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Jia, Y., Johnson, M., Macherey, W., Weiss, R. J., Cao, Y., Chiu, C.-C., Ari, N., Laurenzo, S., and Wu, Y. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019a.

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

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

Jia, Y., Tadmor Ramanovich, M., Wang, Q., and Zen, H. CVSS corpus and massively multilingual speech-to-speech translation. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2022.

Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.-E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A., and Dupoux, E. Libri-light: A benchmark for ASR with limited or no supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A. v. d., Dieleman, S., and Kavukcuoglu, K. Efficient neural audio synthesis. In *Proceedings of International Conference on Machine Learning (ICML)*, 2018.

Kano, T., Sakti, S., and Nakamura, S. Transformer-based direct speech-to-speech translation with transcoder. In *IEEE Spoken Language Technology Workshop (SLT)*, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Krueger, D., Maharaj, T., Kramár, J., Pezeshki, M., Ballas, N., Ke, N. R., Goyal, A., Bengio, Y., Courville, A., and Pal, C. Zoneout: Regularizing RNNs by randomly preserving hidden activations. In *International Conference on Learning Representations (ICLR)*, 2017.

Lavie, A., Waibel, A., Levin, L., Finke, M., Gates, D., Gavalda, M., Zeppenfeld, T., and Zhan, P. JANUS-III: Speech-to-speech translation in multiple languages. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1997.

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

Lee, A., Chen, P.-J., Wang, C., Gu, J., Ma, X., Polyak, A., Adi, Y., He, Q., Tang, Y., Pino, J., et al. Direct speech-to-speech translation with discrete units. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Lee, Y., Shin, J., and Jung, K. Bidirectional variational inference for non-autoregressive text-to-speech. In *International Conference on Learning Representations (ICLR)*, 2021b.

Ma, X., Gong, H., Liu, D., Lee, A., Tang, Y., Chen, P.-J., Hsu, W.-N., Heafield, K., Koehn, P., and Pino, J. Direct simultaneous speech to speech translation. *arXiv preprint arXiv:2110.08250*, 2021.

McCarthy, A. D., Puzon, L., and Pino, J. SkinAugment: Auto-encoding speaker conversions for automatic speech translation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

Nakamura, S., Markov, K., Nakaiwa, H., Kikui, G., Kawai, H., Jitsuhiro, T., Zhang, J.-S., Yamamoto, H., Sumita, E., and Yamamoto, S. The ATR multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

Narayanan, A., Prabhavalkar, R., Chiu, C.-C., Rybach, D., Sainath, T. N., and Strohman, T. Recognizing long-form speech using streaming end-to-end models. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G., Lockhart, E., Cobo, L., Stimberg, F., et al. Parallel WaveNet: Fast high-fidelity speech synthesis. In *Proceedings of International Conference on Machine Learning (ICML)*, 2018.

Oord, A. v. d., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. LibriSpeech: an ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

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

Park, D. S., Zhang, Y., Jia, Y., Han, W., Chiu, C.-C., Li, B., Wu, Y., and Le, Q. V. Improved noisy student training for automatic speech recognition. In *Proc. Interspeech*, 2020.

Pathak, M. A. and Raj, B. Privacy-preserving speaker verification and identification using Gaussian mixture models. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2):397–406, 2012.

Peng, K., Ping, W., Song, Z., and Zhao, K. Non-autoregressive neural text-to-speech. In *Proceedings of International Conference on Machine Learning (ICML)*, 2020.

Post, M., Kumar, G., Lopez, A., Karakos, D., Callison-Burch, C., and Khudanpur, S. Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus. In *International Conference on Spoken Language Translation (IWSLT)*, 2013.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. FastSpeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Ren, Y., Hu, C., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. FastSpeech 2: Fast and high-quality end-to-end text-to-speech. In *International Conference on Learning Representations (ICLR)*, 2021.

Salesky, E., Mäder, J., and Klinger, S. Assessing evaluation metrics for speech-to-speech translation. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

Shen, J., Nguyen, P., Wu, Y., Chen, Z., et al. Lingvo: A modular and scalable framework for sequence-to-sequence modeling. *arXiv preprint arXiv:1902.08295*, 2019.

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

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Tjandra, A., Sakti, S., and Nakamura, S. Speech-to-speech translation between untranscribed unknown languages. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T., and Lee, K. A. ASVspoof 2019: Future horizons in spoofed and fake audio detection. In *Proc. Interspeech*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Eje Henter, G., Le Maguer, S., Malisz, Z., Székely, É., Tånnander, C., et al. Speech synthesis evaluation – state-of-the-art assessment and suggestion for a novel research program. In *Proceedings of the 10th Speech Synthesis Workshop (SSW10)*, 2019.

Wahlster, W. *Verbmobil: Foundations of speech-to-speech translation*. Springer, 2000.

Wan, L., Wang, Q., Papir, A., and Moreno, I. L. Generalized end-to-end loss for speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

Wang, C., Wu, A., and Pino, J. CoVoST 2: A massively multilingual speech-to-text translation corpus. In *Proc. Interspeech*, 2021a.

Wang, C., Wu, A., Pino, J., Baevski, A., Auli, M., and Conneau, A. Large-scale self-and semi-supervised learning for speech translation. In *Proc. Interspeech*, 2021b.

Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T., Lee, K. A., et al. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114, 2020.

Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y., and Chen, Z. Sequence-to-sequence models can directly translate foreign speech. In *Proc. Interspeech*, 2017.

Xin, D., Komatsu, T., Takamichi, S., and Saruwatari, H. Disentangled speaker and language representations using mutual information minimization and domain adaptation

for cross-lingual TTS. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

Yi, Z., Huang, W.-C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., Ling, Z., and Toda, T. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 2020.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., and Wu, Y. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. In *Proc. Interspeech*, 2019.

Zhang, C., Tan, X., Ren, Y., Qin, T., Zhang, K., and Liu, T.-Y. UWSpeech: Speech to speech translation for unwritten languages. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Chen, Z., Skerry-Ryan, R., Jia, Y., Rosenberg, A., and Ramabhadran, B. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. In *Proc. Interspeech*, 2019.

Zheng, Y., Wang, X., He, L., Pan, S., Soong, F. K., Wen, Z., and Tao, J. Forward-backward decoding for regularizing end-to-end TTS. In *Proc. Interspeech*, 2019.

A. Table of hyper-parameters

Table 7: Model hyper-parameters used in the experiments. ("$\times n$": n layers; †: 128-dim pre-net is used for translating in a single voice; 16-dim pre-net is used for voice preservation.)

	Fisher Es-En	CoVoST 2	Conversational
Input			
Sample rate (Hz)	8,000	48,000	16,000 – 48,000
Mel channels		80	
Mel lower band (Hz)		125	
Mel upper band (Hz)	3,800	7,600	7,600
Frame size (ms)		25.0	
Frame step (ms)		10.0	
Output			
Sample rate (Hz)		24,000	
Mel channels		128	
Mel lower band (Hz)		20	
Mel upper band (Hz)		12,000	
Frame size (ms)		50.0	
Frame step (ms)		12.5	
SpecAugment			
Freq blocks		2	
Time blocks		10	
Freq block max length ratio		0.33	
Time block max length ratio		0.05	
Encoder			
Conformer dims		144×16	
Attention heads		4	
Conv kernal size		32	
Subsample factor		4	
Attention			
Output dim	256	512	512
Hidden dim	512	512	512
Attention heads	4	8	8
Dropout prob	0.1	0.2	0.2
Decoder			
LSTM dims	256×4	512×6	512×4
Zoneout prob	0.1	0.1	0.1
Phoneme embedding dim	96	256	256
Label smoothing uncertainty	0.1	0.1	0.1
Loss weight	10.0	10.0	10.0
Duration predictor			
Bi-LSTM (dim $\times$ layers)	64×2	128×2	128×2
Loss weight	1.0	1.0	1.0
Synthesizer			
LSTM dims		$1{,}024 \times 2$	
LSTM zoneout prob		0.1	
Pre-net dims	128×2	128×2	128 / 16 † $\times 2$
Pre-net dropout prob		0.5	
Post-net (kernel, channels) $\times$ layers		$(5, 512) \times 4 + (5, 128)$	
Loss weight		1.0	
Training			
Optimizer		Adam (Kingma & Ba, 2015)	
Learning rate schedule		Vaswani et al. (2017)	
Learning rate (peak)	4.2×10^{-3}	2.2×10^{-3}	3.3×10^{-3}
Warm-up steps	10K	20K	10K
Batch size	1,024	768	768
L^2 regularization weight	10^{-6}	10^{-6}	10^{-6}

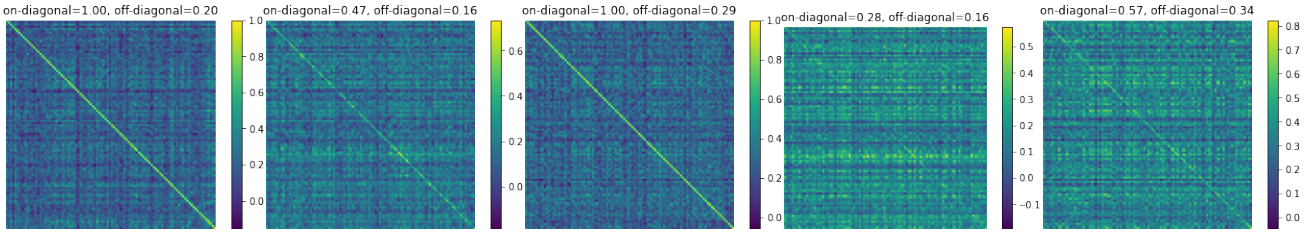


Figure 3: Affinity matrices of d-vector similarity among 100 random examples. ("s2st" refers to the predictions from Translatotron 2.)

Table 8: Objective d-vector similarity between the predicted translated speech (English) and the source human speech (Spanish) on speaker turns. The similarity between the leading/trailing 1.6-second segment from the predicted speech and the entire 1st/2nd source speaker's speech is measured. (↑ / ↓ means higher/lower values are better.)

	1st source speaker		2nd source speaker	
	Leading seg. ↑	Trailing seg. ↓	Leading seg. ↓	Trailing seg. ↑
Translatotron 2	0.21	0.19	0.21	0.19
+ ConcatAug	0.20	**0.14**	**0.14**	0.21
Translatotron	0.20	0.22	0.27	0.29
+ ConcatAug	**0.32**	0.16	**0.14**	**0.35**
Reference (synthetic)	0.48	0.17	0.15	0.48

B. Objective speaker similarity analysis

Subjective speaker similarity evaluation is costly and has a long turnaround. We explored alternative objective evaluation using separately trained speaker encoders, such as d-vector (Wan et al., 2018). We evaluated the voice retention performance using the cosine similarity of the d-vectors.

We first checked the scenario that each input contains a single speaker's recording. Figure 3 visualizes the affinity matrices of d-vector similarity among different input utterances for a Translatotron 2 model. The outstanding higher similarity values on the diagonals indicate that the model is able to preserve the source speaker's voice in the predicted translation speech.

We then conducted a detailed evaluation for the voice retention performance for speaker turns. The experiment setting up was identical to Section 5.4.1, except that the speaker similarity was measured by d-vector similarity instead of subjective MOS evaluation. The d-vectors for each source speaker were computed on the entire original utterance before concatenation; the d-vectors for each speaker in the prediction is approximated by computing on the leading/trailing 1.6 seconds of predicted speech.

The results are shown in Table 8. Consistent with the MOS evaluation results in Table 4, when the concatenation augmentation was not used, the d-vector similarity to each source speaker is about the same regardless if it was compared to the leading or trailing segments, indicating that the predicted speech was in a single speaker's voice and the model was unable to separate different speakers in the input, but rather optimized for both source speakers at the same time. When the concatenation augmentation was used, the d-vector similarity was significantly higher between matched pairs than between unmatched pairs, indicating that the models were able to separate different speakers in the input and preserve their voices in the predicted translation speech respectively.

However, when these similarities are compared among different models, it seems to suggest that Translatotron performed better than Translatotron 2, which is contradictory to the subjective evaluation results in Table 4. By carefully listening to the audio samples, we found that such discrepancy may be due to the fact that the d-vector model was also sensitive to non-voice related acoustic characteristics, such as reverb and channel noise in the audios. This is likely a consequence of the fact that in the large-scale training set for the d-vector model used in the evaluation, each speaker is typically associated with a particular recording condition, e.g. recording device and room. Because the encoder output from the Translatotron model was of significantly larger dimension than from the Translatotron 2 model (2048 vs 144), it was capable of carrying more non-voice acoustic information and thus obtained better d-vector similarity, which not necessarily indicating higher speaker similarity.

These results suggest that while such speaker encoder-based objective analysis reveals insightful indications about the performance of the S2ST models, it can be less reliable compared to subjective MOS evaluation. Such reliability also highly depends on the training details of the speaker encoder model being used, especially the training corpus.