International Journal of Research in Advent Technology, Vol.13, No.1, March 2025
E-ISSN: 2321-9637
Available online at www.ijrat.org

Survey On Monolingual Speech-to-Speech Translation

Yash Vachhani and Prashant Viradiya

Abstract— Direct Speech-to-Speech (S2S) translation represents a significant goal in facilitating seamless cross-lingual communication, aiming to overcome the latency and error propagation issues inherent in traditional cascaded systems (ASRMT-TTS). However, the development of high-performing direct S2S models has been critically constrained by the scarcity of large-scale parallel S2S corpora. This survey details Translatotron 3, a pivotal direct S2S system introduced by Google Research in 2023. Translatotron 3 fundamentally shifts the paradigm by demonstrating, for the first time, the feasibility of training a high-quality, end-toend S2S model exclusively using readily available monolingual data resources: source/target speech and source/target text. Leveraging innovative techniques such as unsupervised utterance splitting, phoneme-based intermediate representations, speech-adapted back-translation, and a non auto regressive decoder for rapid inference, Translatotron 3 achieves strong translation quality and remarkable speaker voice preservation without requiring any parallel S2S examples. We critically review the technological context, dissect the model's architecture and training methodology, analyze its reported performance benchmarks, and discuss its profound implications for advancing S2S translation, particularly for the vast number of low-resource languages previously underserved by data-hungry models.

Index Terms— Direct Speech To Speech Translation, Monolingual Translation, Machine Translation, Unsupervised Speech To Speech Translation,

I. INTRODUCTION

Translatotron 3 is situated within the broader context of Speech-to-Speech Translation (S2ST) research, building on previous efforts while addressing their key limitations.

Traditionally, the S2ST systems [1,2,3] employed a cascade approach, sequentially linking three separate components: Automatic Speech Recognition (ASR) to transcribe source speech to text, Machine Translation (MT) to translate source text to target text, and Text-To-Speech (TTS) synthesis to generate target speech from translated text. While functional, this pipeline approach often suffers from the propagation of errors from one stage to the next and tends to lose the essential para-linguistic and non-linguistic characteristics (like emotion, speaking style, pauses, speaker identity) present in the original source speech.

To overcome these limitations and potentially preserve more speech nuances, researchers developed direct S2ST models. Jia et al. [4] introduced the first end-to-end model, "Translatotron", which directly mapped source speech spectrograms to target speech spectrograms. "Translatotron 2"
[5] followed as an improvement, offering better performance and controllability. However, these pioneering direct models, along with subsequent related work exploring different techniques such as discrete speech representations [6, 7, 8] or two-pass architectures [9], primarily relied on supervised learning. This requires large-scale, parallel bilingual speech datasets (source speech paired with corresponding target language speech), which are expensive and difficult to create, especially for low-resource languages. Furthermore, these datasets often lack the corresponding para-/non-linguistic labels, hindering the explicit transfer of such features.

Other research lines have explored ways to reduce the reliance on fully parallel speech data or leverage different types of data. Some approaches use self-supervision techniques with untranscribed speech and unpaired text data [10], generate pseudolabels from cascaded systems [11], combine teacher models with pseudolabeling for unlabeled data [9], or jointly pre-train models using monolingual speech alongside bilingual text datasets [12]. Another distinct approach involves using discrete speech tokens (learned representations from models such as SoundStream [13], w2v-BERT [14], EnCodec [15], Hubert [16]) combined with language models (LM) for S2ST or speech-to-text translation.

Crucially, the development of Translatotron 3 draws significant inspiration from the field of Unsupervised Machine Translation (UMT) [17, 18]. UMT demonstrated that text translation is possible using only large monolingual corporain each language, without any parallel sentences. Key techniques enabling UMT include back-translation [19] (where a model translates target text back to source and uses the original target text as pseudoparallel data) and

Manuscript revised on April 22, 2025 and published on May 2, 2025
Yash Vachhani, *Computer Engineering Department, Gyanmanjari Innovative University, Bhavnagar, India.*
Prashant Viradiya*, Computer Engineering Department, Gyanmanjari Innovative University, Bhavnagar, India*

doi: 10.32622/ijrat.131202513

unsupervised cross-lingual embedding mapping [18, 20] (learning shared representations for words across languages without bilingual dictionaries). Although some work explored unsupervised speech-to-text translation [21], Translatotron 3 specifically aims to apply these unsupervised principles (backtranslation, embedding mapping) directly to the end-to-end speech-to-speech translation task, thus eliminating the need for parallel speech data.

II. RELATED WORK

The field of speech-to-speech translation (S2ST) has evolved from modular, cascaded systems to fully end-toend models. This section details the foundational and recent works that led to the development of Translatotron 3.

a) Early Cascaded Systems

JANUS-III [1] was one of the first multilingual S2ST systems, developed in the late 1990s. It used a traditional cascade pipeline composed of ASR, MT, and TTS components. While a technical achievement for its time, JANUS-III suffered from high latency and error accumulation between stages. The model was rule-based and required languagespecific tuning, limiting its scalability.

Verbmobil [2], a German-funded project, aimed to translate spontaneous speech between German, English, and Japanese. It also followed a cascade approach and included domain-specific dialogue management. Verbmobil showed the feasibility of real-time S2ST but struggled with spontaneous speech variability and relied on handcrafted rules.

The ATR Multilingual S2ST System [3] represented a more structured approach by combining statis tical methods with traditional modules. It emphasized language-independent architecture and incorporated probabilistic models to improve robustness. However, like others in its class, it still lacked integration and suffered from component dependencies.

b) Direct Speech-to-Speech Models

Translatotron 1 [4] introduced the first end-to-end direct S2ST model. Unlike cascade systems, it bypassed intermediate textual representations, instead mapping input speech directly to output speech using a sequence-tosequence architecture. The model consisted of a speech encoder, a decoder that generated a spectrogram, and a vocoder to synthesize waveform. While groundbreaking, it underperformed in translation accuracy and could not preserve speaker characteristics well.

Translatotron 2 [22] addressed key limitations of its predecessor. It introduced:

A shared attention mechanism to improve alignment between input and output.

Explicit speaker embedding modules to retain speaker identity in the output.

Improved prosody transfer and smoother waveform synthesis.

Despite its improvements, the model still required parallel speech-to-speech data, making it impractical for many low-resource languages.

TranSpeech [8] proposed a novel method using bilateral perturbation to make the model more robust to noise and variability in speech. The bilateral training technique helped improve generalization, but like previous models, TranSpeech still depended on supervised data.

Translatotron 3 [29] is a pioneering end-to-end unsupervised model for direct speech-to-speech translation (S2ST). Break away from the conventional dependency on bilingual speech-text corpora and instead learns to translate using only monolingual speech-text datasets. This approach allows the model to scale better to low-resource languages and realworld scenarios where parallel datasets are scarce or infeasible to collect.

Unlike earlier systems that rely on cascaded pipelines of ASR (Automatic Speech Recognition), MT (Machine Translation), and TTS (Text-to-Speech), Translatotron 3 enables a fully integrated architecture capable of capturing linguistic and nonlinguistic information in a single unified model. Importantly, it is designed to preserve paralinguistic cues such as speaker identity, speaking rate, intonation, and pauses, elements often lost in conventional systems.

c) S2ST for Unwritten or Low-Resource Languages

UWSpeech [6] tackled S2ST for unwritten languages by leveraging auxiliary modalities like images or semantic grounding. This model used unsupervised techniques to align concepts across languages, though it often required multimodal supervision, such as co-occurring visual information.

Textless Speech-to-Speech Translation [7, 23] explored translation without any written text. These models used discrete acoustic units (e.g., quantized speech tokens) as both input and output representations. This approach bypassed the need for text corpora but required accurate unit discovery and often failed to capture complex semantics.

d) Unsupervised Machine Translation (Text and Speech)

Unsupervised Neural Machine Translation (UNMT) [18, 17] demonstrated that it is possible to train text-based machine translation systems using only monolingual corpora. These models relied on:

Denoising autoencoders to learn language structure. Back-translation to create pseudo-parallel training data.

Shared embedding spaces between languages.

These ideas significantly influenced the training methods adopted in Translatotron 3.

doi: 10.32622/ijrat.131202513

Chung et al. [**23**] proposed an unsupervised approach for speech-to-text translation. Their method utilized self supervised representations and pseudo-labeled text from ASR to generate parallel data. Though focused on ST, their back-translation approach inspired extensions to S2ST.

CycleGAN [24] proposed unpaired image-to-image translation using cycle-consistency loss. This concept of enforcing consistency between forward and backward transformations is foundational in Translatotron 3's back-translation phase.

e) Pretraining and Representation learning

wav2vec 2.0 [25], **HuBERT** [16], and **W2V-BERT** [14] enabled high-quality, self-supervised pretraining for speech representations. These models achieved state-of-the-art performance on downstream ASR and ST tasks by learning from raw audio without labels.

SpeechT5 [26] presented a unified encoder-decoder pretraining scheme for various spoken language tasks, supporting ASR, TTS, and S2ST via task-specific heads. Its generalizability is a strength, but it still requires fine-tuning with supervised datasets.

AudioPaLM [27] extended the capabilities of language models by enabling them to "speak and listen." It trained on paired audio-text datasets and supported multi-modal tasks, but was extremely data and resource intensive.

f) Multilingual Embeddings and Alignment

MUSE [28] introduced unsupervised alignment of word embeddings across languages. It enabled learning of crosslingual mappings with no parallel data and became instrumental in Translatotron 3, where part of the encoder is trained to align with these multilingual word embeddings.

Masked Autoencoders (MAE) [**24**] contributed to the rise of self-supervised learning by proposing masked reconstruction of visual data. Translatotron 3 applies this concept to spectrogram inputs, improving generalization and robustness.

III. Translatotron 3: Architecture and Training Methodology

Translatotron 3 is designed as a fully unsupervised, direct speech-to-speech translation model capable of learning without parallel bilingual data. The model architecture and training process are carefully crafted to enable effective crosslingual speech translation while preserving the speaker's identity and prosodic features. This section elaborates on the architecture and training objectives used in Translatotron 3.

A. Model Architecture

The architecture of Translatotron 3 follows a classic encoder-decoder paradigm with a shared encoder and two distinct decoders one for the source language and another for the target language. The encoder is responsible for transforming input speech (in the form of spectrograms) into a latent representation, which is later processed by the appropriate decoder based on the direction of translation.

Shared Encoder

The encoder (E) architecture is adapted from Translatotron 2, and it processes input spectrograms regardless of language. Its output is split into two parts:

The first half, $E_m(S_{in})$, is optimized to match multilingual word embeddings from the MUSE framework.
The second half, $E_o(S_{in})$, serves as a free latent representation not explicitly aligned to external embeddings.

By training the encoder to produce language-invariant features, the model constructs a shared multilingual latent space.

Dual Decoders

Each decoder consists of three modules: a linguistic decoder, an attention mechanism, and an acoustic synthesizer. The source decoder (D_s) and target decoder (D_t) are functionally similar but operate on different language outputs. During inference, the encoder processes the input speech, and the relevant decoder generates the translated spectrogram output.

B. Training Objectives

The training of Translatotron 3 is executed in two phases, each designed to build the multilingual embedding space and ensure translation accuracy. These are:

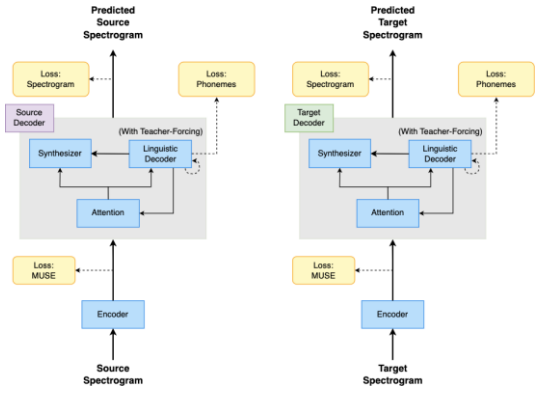

Fig. 1: Phase 1 uses the reconstruction loss via the auto-encoding path [29]

doi: 10.32622/ijrat.131202513

Phase 1: Masked Autoencoding with MUSE Loss
In the initial phase, the model is trained as a masked auto encoder. Input spectrograms are corrupted using SpecAugment, and the model is trained to reconstruct them. Additionally, the encoder is constrained to produce outputs that match pre-trained MUSE word embeddings.

MUSE Loss The MUSE loss aligns the encoder output with multilingual word embeddings using the following objective:

$$L_{\text{MUSE}}(S_{\text{in}}) = \frac{1}{n}\sum_{i=1}^{n} ||E(S_{\text{in}})_i - E_i||_2^2 \quad (1)\ 1$$

Here, S_{in} denotes the input spectrogram (either source or target language), and E_i is the MUSE embedding for the i-th word. This loss encourages the encoder to produce languageinvariant embeddings.

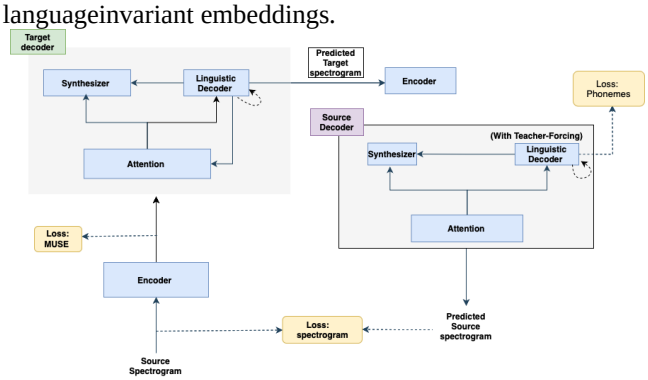

Fig. 2: Phase 2 employs the reconstruction loss via back-translation [29]

PHASE 2: BACK-TRANSLATION WITH RECONSTRUCTION LOSS
In the second phase, back-translation is introduced to simulate translation in an unsupervised setting. The model generates a pseudo-translation from a source utterance, re-encodes it, and reconstructs the original source from it. This cycle consistency ensures translation fidelity.

Reconstruction Loss The reconstruction loss ensures that the model can regenerate input spectrograms from their latent embeddings. It includes three components:
Spectrogram loss (L_{spec}): Frame-wise distance between predicted and ground-truth spectrograms.
Duration loss (L_{dur}): Discrepancy between predicted and actual phoneme durations.
Phoneme loss (L_{phn}): Cross-entropy loss on predicted vs. true phoneme sequences.
The overall reconstruction loss is:

$$\mathrm{L}_{\text{recon}} = \mathrm{L}_{\text{spec}} + \mathrm{L}_{\text{dur}} + \mathrm{L}_{\text{phn}} \quad (2)$$

Back-Translation Loss This loss uses the cycle consistency framework. The source spectrogram is translated to the target and then back to the source, enforcing the original spectrogram to be reconstructed accurately:

$$L_{\text{back-trans}} = L_{\text{spec}}(\hat{S}_s, S_s) + L_{\text{dur}} + L_{\text{phn}} \quad (3)$$

The process is symmetric, and a similar loss is computed for target-to-source direction as well.

C. Total Loss Function
The final loss used for training during each phase is:

$$L_{\text{recon-phase}} = L_{\text{recon}} + L_{\text{MUSE}}(S_s) + L_{\text{MUSE}}(S_t) \quad (4)$$

$$L_{\text{BT-phase}} = L_{\text{back-trans}} + L_{\text{recon-phase}} \quad (5)$$

These objectives jointly ensure that the model learns a highquality multilingual latent space, effective speech reconstruction, and accurate translation without any direct supervision.

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup
Translatotron 3 was evaluated on Spanish-English speech-to-speech translation tasks. The model was trained using only monolingual speech-text datasets, emphasizing its unsupervised training capability. The training utilized 64 TPUv4 devices over one week.

The experiments utilized the following datasets:

Unpaired Conversational Dataset (UC): A synthesized dataset comprising approximately 371 hours of English and 350 hours of Spanish speech, created by crowd-sourcing humans to read Spanish-English machine translation datasets. Both source and target speech were synthesized using a Phoneme-and-Grapheme NonAttentive Tacotron TTS model and a WaveRNN-based neural vocoder.

Common Voice 11 (CV11): A publicly available corpus containing a diverse set of speech recordings in multiple languages, including Spanish and English. The dataset was used in both its original and synthesized forms to assess the model's performance across different speech styles and recording conditions.

doi: 10.32622/ijrat.131202513

CoVoST 2 (CVE): A subset of Common Voice 11, used for evaluating Spanish-English real speech translation with verified translations.

B. Evaluation Metrics

The primary metrics for assessing translation quality were:
BLEU (Bilingual Evaluation Understudy): Measures the correspondence between machine-generated translations and reference translations.
MOS (Mean Opinion Score): Assesses the naturalness of synthesized speech on a scale from 1 to 5.
CS (Cosine Similarity): Evaluates speaker similarity between input and output speech.

C. Results

Table 1: Performance Comparison of Translatotron 3 with Baseline Cascade System

Dataset	Model	BLEU	MOS	CS
UC (Es-En)	Cascade	6.13	3.24	0.16
	Translatotron 3	24.27	4.20	0.65
UC (En-Es)	Cascade	5.58	3.48	0.15
	Translatotron 3	18.85	3.70	0.62
CV11 (Es-En)	Cascade	10.48	3.36	0.17
	Translatotron 3	14.25	4.13	0.56
CV11 (En-Es)	Cascade	9.46	3.64	0.16
	Translatotron 3	13.45	3.78	0.63
CVE (Es-En)	Cascade	9.92	3.30	0.17
	Translatotron 3	10.76	4.21	0.34

[29]

Translatotron 3 demonstrated significant improvements over the baseline cascade system across all datasets. Notably, on the synthesized Unpaired Conversational dataset, it achieved an improvement of 18.14 BLEU points over the base-line [29]. Additionally, the model effectively preserved paralinguistic features, including speaker identity and prosody, without explicit modeling or supervision, as evidenced by higher CS scores.

D. Ablation Study

An ablation study was conducted to assess the impact of various components:

The study revealed that removing the reconstruction loss or back-translation loss significantly degraded performance, highlighting their critical roles in the model's success.

These examples underscore Translatotron 3's ability to generate natural-sounding translations while maintaining speaker-specific features.

Table 2: Ablation Study Results on BLEU Scores

Model Variant	En-Es BLEU	Es-En BLEU
Full Model	18.85	24.27
Reconstruction Loss	0.41	2.99
Back-Translation Loss	1.91	3.62
MUSE Embedding Loss	5.44	6.22
SpecAugment	9.23	12.88

[29]

V. CONCLUSION

Translatotron 3 marks a significant advancement in the field of direct speech-to-speech translation (S2ST) by demonstrating that high-quality translation can be achieved without the use of parallel bilingual data. Unlike traditional cascade systems and earlier direct models such as Translatotron 1 and 2, which depended heavily on supervised datasets, Translatotron 3 leverages monolingual speech-text corpora to learn a multilingual latent space capable of producing fluent and expressive speech translations.
Beyond its impressive results, Translatotron 3 presents a paradigm shift in how we approach low-resource and underrepresented languages in S2ST. By eliminating the dependency on parallel corpora, it makes speech translation more inclusive, scalable, and adaptable to languages without standardized written forms.
In conclusion, Translatotron 3 sets a new benchmark in unsupervised S2ST research and lays a strong foundation for the development of multilingual, expressive, and privacy-preserving translation systems that are truly end-to-end and data-efficient.

REFERENCES

[1] Lavie, A., et al., 1997. "Janus-iii: Speech-to-speech translation in multiple languages". In ICASSP, Vol. 1, pp. 99–102.
[2] Wahlster, W., 2013. *Verbmobil: Foundations of speechto-speech translation*. Springer.
[3] Nakamura, S., et al., 2006. "The atr multilingual speech-to-speech translation system". *IEEE Transactions on Audio, Speech, and Language Processing*, **14**(2), pp. 365–376.
[4] Jia, Y., et al., 2019. "Direct speech-to-speech translation with a sequence-to-sequence model". In Interspeech, pp. 1123–1127.
[5] Jia, Y., et al., 2022. "Translatotron 2: High-quality direct speech-to-

doi: 10.32622/ijrat.131202513

speech translation with voice preservation". In Proceedings of the International Conference on Machine Learning (ICML), pp. 10120–10134.

[6] Zhang, C., et al., 2021. "Uwspeech: Speech to speech translation for unwritten languages". In AAAI, Vol. 35, pp. 14319–14327.

[7] Lee, A., et al., 2021. "Textless speech-to-speech translation on real data". *arXiv preprint arXiv:2112.08352*.

[8] Huang, R., et al., 2022. "Transpeech: Speech-to-speech translation with bilateral perturbation". *arXiv preprint arXiv:2205.12523*.

[9] Wang, C., et al., 2022. "Simple and effective unsupervised speech translation". *arXiv preprint arXiv:2210.10191*.

[10] Tang, Y., et al., 2022. "Unified speech-text pretraining for speech translation and recognition". In ACL, pp. 1488–1499.

[11] Dong, Q., et al., 2022. "Leveraging pseudo-labeled data to improve direct speech-to-speech translation". *arXiv preprint arXiv:2205.08993*.

[12] Wei, K., et al., 2022. "Joint pre-training with speech and bilingual text for direct speech-to-speech translation". *arXiv preprint arXiv:2210.17027*.

[13] Zeghidour, N., et al., 2022. "Soundstream: An end-to-end neural audio codec". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **30**, pp. 495– 507.

[14] Chung, Y.-A., et al., 2021. "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training". In IEEE ASRU,
pp. 244–250.

[15] Defossez, A., et al., 2022. "High fidelity neural audio compression". *arXiv preprint arXiv:2210.13438*.

[16] Hsu, W.-N., et al., 2021. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, pp. 3451–3460.

[17] Artetxe, M., et al., 2018. "Unsupervised neural machine translation". In ICLR.

[18] Lample, G., et al., 2018. "Unsupervised machine translation using monolingual corpora only". In ICLR.

[19] Sennrich, R., et al., 2015. "Improving neural machine translation models with monolingual data". *arXiv preprint arXiv:1511.06709*.

[20] Artetxe, M., et al., 2018. "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings". In ACL.

[21] Chung, Y.-A., et al., 2019. "Towards unsupervised speech-to-text translation". In ICASSP, pp. 7170–7174.

[22] Jia, Y., Ramanovich, M. T., Remez, T., and Pomerantz, R., 2022. "Translatotron 2: High-quality direct speechto-speech translation with voice preservation". In International Conference on Machine Learning, PMLR, pp. 10120–10134.

[23] Lee, A., et al., 2022. "Direct speech-to-speech translation with discrete units". In ACL, pp. 3327–3339.

[24] Zhu, J.-Y., et al., 2017. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In ICCV, pp. 2223–2232.

[25] Baevski, A., et al., 2020. "wav2vec 2.0: A framework for self-supervised learning of speech representations". *NeurIPS*, **33**, pp. 12449–12460.

[26] Ao, J., et al., 2021. "Speecht5: Unified-modal encoder decoder pre-training for spoken language processing". *arXiv preprint arXiv:2110.07205*.

[27] Rubenstein, P. K., et al., 2023. "Audiopalm: A large language model that can speak and listen". *arXiv preprint arXiv:2306.12925*.

[28] Conneau, A., et al., 2017. "Word translation without parallel data". *arXiv preprint arXiv:1710.04087*.

[29] Nachmani, E., Levkovitch, A., Ding, Y., Asawaroengchai, C., Zen, H., and Tadmor Ramanovich, M., 2023. "Translatotron 3: Speech to speech translation with monolingual data". *arXiv preprint arXiv:2305.17547*.

AUTHORS PROFILE

Yash Vachhani, I am student of Gyanmanjari Innovative University computer engineering department. I am pursuing Masters in compute engineering. I completed my Bachelors from government engineering college, Bhavnagar. My research focuses on Conversational AI, with an emphasis on natural language understanding, dialogue systems, and human-computer interaction.

Prof. Prashant Viradiya, assistant professor in computer department at Gyanmanjari Innovative University. I have 12 years of teaching experience. I completed my masters in computer engineering and pursuing my Ph. D in computer engineering. I am passionate about Web Design & Development.

doi: 10.32622/ijrat.131202513