

Review article

Speech to speech translation with translatotron: A state of the art review [☆]

Jules R. Kala ^{a, , *}, Emmanuel Adetiba ^{b,c,f}, Abdultaofeek Abayomi ^{d,e}, Oluwatobi E. Dare ^{b,c}, Ayodele H. Ifijeh ^{b,c}

^a International University of Grand-Bassam, Grand-Bassam, Côte d'Ivoire

^b Department of Electrical and Information Engineering, Covenant University, Ota, Nigeria

^c Covenant Applied Informatics & Communication African Center of Excellence (CApIC-ACE), Covenant University, Ota, Nigeria

^d Innovation and Advanced Science Research Group (IASRG), Summit University, PMB 4412, Offa, 250101, Kwara State, Nigeria

^e Research Associate, Faculty of Engineering, Built Environment and Information Technology, Water Sisulu University, Buffalo City Campus, East London 5200, South Africa

^f HRA, Institute for Systems Science, Durban University of Technology, Durban, South Africa

ARTICLE INFO

Keywords:

Translatotron

BLEU

Cascade

Speech-to-speech

ABSTRACT

A speech-to-speech translation using cascade-based methods has been considered a benchmark for a very long time. Still, it is plagued by many issues, like the time to translate a speech from one language to another and compound errors. These issues are because cascade-based methods use a combination of other methods, such as speech recognition, speech-to-text transcription, text-to-text translation, and finally, text-to-speech transcription. Google proposed Translatotron, a sequence-to-sequence direct speech-to-speech translation model that was designed to address the issues of compound errors associated with cascade-based models. Today, there are 3 versions of the Translatotron model: Translatotron 1, Translatotron 2, and Translatotron 3. Translatotron 1 is a proof of concept to demonstrate direct speech-to-speech translation. This first approach was found to be less effective than the cascade model, but it was producing promising results. Translatotron 2 was an improved version of Translatotron 1 with results similar to the cascade-based model. Translatotron 3, the latest version of the model, significantly improves the translation and is better than the cascade model at some points. This paper presents a complete review of speech-to-speech translation using Translatotron models. We will also show that Translatotron is the best model to bridge the language gap between African Languages and other well-formalized languages.

Contents

1. Introduction	2
2. Speech to speech translation	2
3. Translatotron 1	3
4. Translatotron 2	3
5. Translatotron 3	4
6. Existing corpora for direct speech-to-speech translation	5
7. Comparative study	5
8. Deployment roadmap for translatotron	6
9. Theoretical and practical implications of this review	7
10. Limitations and open challenges	7
11. Conclusion	7
CRedit authorship contribution statement	8
Declaration of competing interest	8
Acknowledgement	8

[☆] This document is the results of the research project funded by Google.

* Corresponding author.

E-mail address: kala.j@iugb.edu.ci (J.R. Kala).

<https://doi.org/10.1016/j.rineng.2025.107780>

Received 20 April 2025; Received in revised form 6 October 2025; Accepted 16 October 2025

Data availability	8
References	8

1. Introduction

In the World, we can identify approximately 7,151 languages [8]. Some are formalized and others are not (mostly African languages). The languages are sometimes not properly formalized or still going through a formalization process. To bridge the communication barrier, it is important to design artificial intelligence (AI) models that are able to translate from one language to another. The translation model could help preserve the language. Speech-to-speech translation is certainly going to allow other languages, such as Yoruba (or any other non-formalized African language), to be used for the design of Yoruba image captioning or for the design of models for Yoruba pronunciation correction as presented in [28,23].

Many AI models have been designed to automate the translation process, we can identify cascade-based models and speech-to-speech (S2ST) (or direct speech translation) models. Cascade-based translation models use a combination of many methods to achieve speech-to-speech translation [4]. On the other hand, the speech-to-speech model translates languages into another directly, hence avoiding compound errors observed when using the cascade approach [19]. The goal of this paper is to review the speech-to-speech translation models. A particular focus will be on the translatotron models designed by Google.

For an in-depth understanding of this literature review, it is important to define some key terms. Automatic speech recognition is a technology that converts spoken language into written text [39]. On the other hand, machine translation is a software used to translate text or speech from one language to another [21]. While, text-to-speech synthesis is a technology used to convert written text into spoken language [11]. End-to-end speech translation is a system that directly translates speech from one language into another without an intermediate text representation [36]. Neural machine translation: is an artificial neural network-based translation model that uses neural networks to predict the likelihood of a sequence of words in a given target language [33]. Real-time translation: it is a speech translation done in real-time or with relatively low delays [1]. The multimodal translation is a translation system that includes many input modes: text, speech, and visual cues [34]. Low-resource languages are languages with limited resources available to train AI models [7]. Zero-shot translation is the ability of a translation model to translate between two language pairs without being explicitly trained [20]. Paralinguistic features are nonlexical language elements such as tone, pitch, intonation, and emotion [10]. There are also translation model performance evaluations or metrics that are critical when comparing translation models [31]. The first metric is BLEU (Bilingual Evaluation Under Study) it is a value comparing the machine translation output to a reference translation. TER (Translation Edit Rate) represents the number of edits required to transform the machine translation into the reference translation. The MCD (Mel-Cepstral Distortion) is a metric used to evaluate the quality of a synthesized speech.

This review work will provide a clear answer to the following question: what are the differences between all translatotron models, and which model is adequate for the design and implementation of a S2ST model?

The contributions of this work are:

- A review of S2ST models.
- A comparison of translatotron's models.

In this literature review work, we will be exploring how translatotron has impacted the design of direct speech-to-speech translation models, with a particular focus on the contribution of this model in language

processing, language preservation, speaker identity preservation, and cross-lingual communication.

In summary, this work will provide an overview of the state of the art of S2ST models with a particular focus on Translatotron models. For translatotron models, a critical analysis of their evolution, technical innovations, and performance tradeoffs will be explored. The goal is to show how these models evolved over the versions to overcome limitations and even outperform the cascade base model.

The rest of this paper is organized as follows: Section one: a review of speech-to-speech translation models, followed by an in-depth description of the translatotron models from sections two to four, and, in section five, some corpora used for S2ST will be presented. In section six, a comparative study of the translatotron model will be presented, followed by the conclusion.

2. Speech to speech translation

Translatotron wasn't the first and not the only model to implement speech-to-speech translation. In this section, we will be exploring some models and approaches used for the implementation of speech-to-speech translation.

Cheng et al., [5], developed a speech-to-speech translation model (S2ST) for real-world unwritten language. Their focus was on the translation of languages that don't have a standard text-writing system. Translation from English to Taiwanese Hokkien was used as a case study. Data used to build the model were obtained from human-annotated data, mining of speech datasets, and the use of pseudo-labeling techniques. They took advantage of the advances in applying a self-supervised discrete representation model and improved their speech-to-speech model.

It is shown that S2ST models benefited from the advances made in the field of speech representation but are still facing issues such as acoustic multi modality and high latency. Huang et al., [16], also agreed with these observations and also stated that current S2ST uses autoregressive models that predict each unit based on the previously generated unit, and failed to take advantage of parallel computation. Thus the authors, designed TranSpeech, an S2ST with bilateral perturbation to help solve the acoustic modal issue. The proposed model learns only the linguistic information from speech to generate a more deterministic representation. TranSpeech is considered the first S2ST model that doesn't use autoregression. Overall the proposed model yields an improvement of 2.9 BLEU on average compared to the baseline. The parallel decoding nature significantly reduced the latency.

An early attempt into S2ST was proposed by Vidal, [35]. The author designed a finite state S2ST which is a fully integrated approach to translating speech input language into an output language. The proposed model mapped the input to the output language in terms of the finite state translation model which is learned from input and output sentences. The model was also integrated with the standard acoustic phonetics model of the input language. The resulting global model supplied through Viterbi provides an optimal output sentence for each input. The model was used for an hotel front desk speech services and displays a 700-words dictionary.

Nakamura et al., [26], designed ATR(Advanced Telecommunications Research Institute International) Multilingual Speech-to-Speech Translation System. The purpose of the proposed system was to translate from English to Asian languages (Chinese and Japanese). The system is composed of a large vocabulary, a continuous speech recognition machine, text-to-text translation, and a text-to-speech synthesizer. All the components are multilingual. A statistical machine learning model trained on

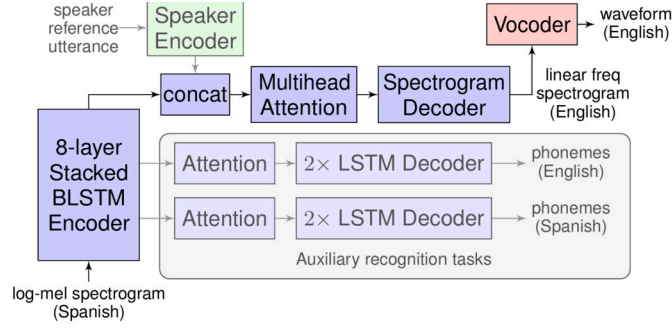


Fig. 1. Translatotron 1 architecture [19].

the corpus forms the basis of the proposed model. The dataset used to train the model is composed of 600000 sentences. The system was able to achieve the level of a person with a score of 750 on the TOEIC (Test of English for International Communication) test.

Data scarcity is one of the drawbacks of S2ST model's design, which is due to missing parallel speeches. Previous S2ST models utilize TTS models to generate sample speeches and augment the dataset to increase its size. There are no study investigating the impact of the synthesized speeches on the quality of the model. Shi et al., [30] show that combining speeches from different TTS models potentially improves the quality of the resulting S2ST model. They proposed a multitask framework with targets from multiple TTS models that optimizes S2ST systems. They report a 2.8 BLEU improvement over the Fisher Spanish dataset.

Designing systems to produce translation from languages without a standard writing system is the focus of Chen et al., [6]. They demonstrated the efficiency of the proposed system on a use case based on the translation from English to Hokkien (a language similar to Mandarin). They also proposed an end-to-end solution from data collection to training. As a result, they released a dataset that can be used to train other S2ST models.

Lee et al., [24], proposed a direct S2ST model with discrete units. The translation model translates from one language to another without relying on intermediate text generation. The proposed model is based on a self-supervised discrete speech encoder on a target speech and a training of a sequence-to-sequence speech-to-speech unit translation to predict the representation of the target speech. The system was able to generate dual modality output (speech and text) in a single inference when target text transcription was available. Experiments on the Fisher Spanish to English dataset yield an improvement of 6.7 BLEU compared to the baseline. When the model is trained without the text transcription, it yields results comparable to the text-based model baseline.

3. Translatotron 1

Translatotron 1 developed by Jia et al., [19]. It is the first attempt at the design of an attention-based sequence to sequence neural network that was able to directly translate speech from one language into another language without intermediate text representation. The proposed model is trained by mapping speech spectrograms of the input language to the output language spectrograms corresponding to the translation. The author demonstrates the ability of the model to synthesize the speaker's voice during translation. The tests conducted on two Spanish-to-English speech translation datasets were not good but promising because the model underperformed compared to the baseline cascade. The main goal of the model was to demonstrate that direct speech-to-speech translation could be achieved.

Fig. 1, describes translatotron 1 Architecture. The architecture is composed of the following elements:

- A sequence-to-sequence attention-based neural network is used to generate a target spectrogram.

Table 1
Comparison of Conversational and Fisher datasets [19].

	Conversational	Fisher
Num train examples	979k	120k
Input / output sample rate (Hz)	16k / 24k	8k / 24k
Learning rate	0.002	0.006
Encoder BLSTM	8×1024	8×256
Decoder LSTM	6×1024	4×1024
Auxiliary decoder LSTM	2×256	2×256
Source / target input layer	8 / 8	4 / 6
Dropout prob	0.2	0.3
Loss decay	constant 1.0	$0.3 \rightarrow 0.001$ at 160k steps
Gaussian weight noise stddev	none	0.05

- A vocoder to transform the target spectrogram to a time domain waveform.
- An optional speaker encoder that is pre-trained and can be used to identify speakers and enable cross-language voice conversation with simultaneous translation.

In Fig. 1, the direct speech-to-speech translation (using $t = \text{translatotron 1}$) between English and Spanish is taken as an example. The architecture takes as input a log-mel spectrogram of a speech in Spanish (this is the acoustic representation). The acoustic representation is fed into an 8-layer Stacked BLSTM (Bidirectional Long-term Short-term Memory) which processes the spectrogram into latent features that capture acoustic and linguistic properties in the spectrogram. The speaker conditioning part is divided into two parts: The speaker encoder (capturing speaker voice characteristics) and the speaker embedding, which is concatenated with the encoder output to preserve speaker identity. The part combining the attention and decoder part helps produce context-aware representations using multihead attention and the spectrogram decoder to generate English from Spanish. The vocoder helps convert the predicted English spectrogram into a waveform. The Attention + 2x LSTM Decoders are used for phoneme recognition, both in Spanish and English. This section serves as a regularizer to improve the robustness and alignment of the prediction.

Table 1 describes the data used to train the model at each level of the architecture.

4. Translatotron 2

Jia et al., [18] designed Translatotron 2, which is a direct speech-to-speech translation model that can be trained end-to-end. The proposed model is composed of a speech encoder, a linguistic decoder, an acoustic synthesizer, and a single attention module to connect all these components. Fig. 2 describes the translatotron 2 architecture. The overall process of translatotron 2 can be summarized with the following equation (1):

$$S' = D(\mathcal{E}(S^s)), \quad (1)$$

Where S' is the predicted target spectrogram, S^s is the spectrogram sequence of the source speech, D is the decoder containing the attention module finally $\mathcal{E}$ is the encoder. Translatotron 2 model is trained in a supervised manner. L_1 and L_2 losses predicted between the target spectrogram S' and the real spectrogram S^t are combined to design the model loss function. The following equation (2) describes translatotron 2 loss function. The loss function is used to evaluate how close the model is to the actual prediction (transcription).

$$\mathcal{L}_{\text{spec}}(S', S^t) = \frac{1}{TK} \sum_{i=1}^T \sum_{j=1}^K \|S'_i - S^t_j\|_j^2, \quad (2)$$

With S'_i representing the i -th frame of S' , T is the number of frame in S' , K the number of frequency bins in S'_i , and $\|\cdot\|_j^2$ the distance. The du-

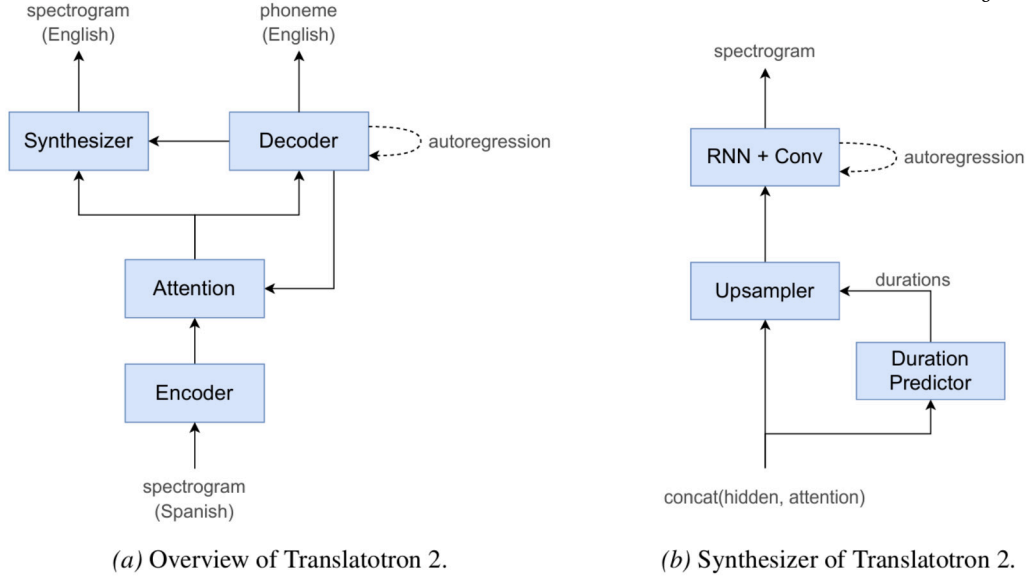


Fig. 2. Translatotron 2 Architecture and training method [18].

ration loss is the second loss linked to the total number of frames T and the total duration of phonemes coming from the acoustics synthesizer as described in equation (3).

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

With d_i equals i -th phonem predicted duration. The auxiliary phoneme loss uses the sequence of predicted probabilities over target phonemes. The sequence is described as $\tilde{P}^t = \{\tilde{P}_1^t, \dots, \tilde{P}_p^t\}$, and the ground as $P^t = \{P_1^t, \dots, P_p^t\}$, and the cross Entropy as $CE(.,.)$. Equation (4) describes the auxiliary phoneme loss.

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

Finally equation (5) presents the overall loss.

$$\mathcal{L} = \mathcal{L}_{\text{spec}}(S^t, S^t) + \lambda_{\text{dur}} \mathcal{L}_{\text{dur}} + \lambda_{\text{phn}} \mathcal{L}_{\text{phn}}(\tilde{P}^t, P^t). \quad (5)$$

Where λ_{dur} and λ_{phn} are the weights of each loss.

Experimental results show that translatotron 2 outperformed translatotron 1 with a large margin of approximately +15.5 BLEU and obtained results that are similar to the baseline cascade. The proposed model can preserve the voice of each speaker in the training dataset. The translatotron 2 architecture was designed to address 3 performance bottlenecks observed in translatotron 1 namely:

- Limitations of the auxiliary textual supervision during training, because it doesn't impact the S2ST directly.
- Modeling translation alignment between 2 very long spectrogram sequences poses a serious challenge when using the attention mechanism.
- Overgeneration and undergeneration are the robustness issues linked to attention-based speech generation.

The training of this model requires a minimum of around 500 hours of paired training data, representing the two languages in direct speech-to-speech translation. The model uses more than 100 million parameters, which requires 2 to 3 weeks on an NVIDIA V100 and approximately 3 to 5 days on a TPU v3 pod.

5. Translatotron 3

One of the techniques used to acquire multilingual word embeddings is Multilingual Unsupervised Embedding (MUSE) described in Conneau et al., [9]. The embedding can be represented as two matrices X and Y if it is assumed that the learning process starts from two sets of word embedding trained on monolingual corpora. The two matrices $X \in \mathbb{R}^{c \times M}; Y \in \mathbb{R}^{c \times M}$ can be used to represent words, where M and c are the embedding parameters. M is the number of the most frequent words used for the computation of embedding with a dimension c . Training a MUSE model is about mapping W^* , knowing that

$$W^* = \arg \min_{W \in \mathbb{R}^{c \times c}} \|WX - Y\|_F = UV^T, \text{ with } U\Sigma V^T = \text{SVD}(YX^T). \quad (6)$$

Where $\|\cdot\|_F$ is the Frobinus norm.

Equation (6) is solved using an adversarial method. The elements randomly sampled from W , X and Y are classified using a discriminator. To avoid the discriminator from inferring the original input W is updated. The discriminator is a classifier with a loss defined using equation (7).

$$\mathcal{L}_w = -\frac{1}{M} \sum_{i=1}^M \{ \log P_D(\text{source} = 0 | Wx_i) + \log P_D(\text{source} = 1 | y_i) \} \quad (7)$$

With $P_D(\cdot)$ represents the discriminator.

Translatotron 3 [25] is a recent approach used to design an unsupervised direct S2ST model that uses a shared encoder and a separated decoder for the source and target languages. A reconstruction loss, a MUSE embedding loss, and a S2S back translation loss are used to train the proposed model. The proposed model has two decoders one for the target D_t and the other for the source D_s . Each decoder contains an acoustic synthesizer, a linguistic decoder, and a singular attention model. It also has an encoder $\mathcal{E}$, to encode both target and source languages. The encoder used in translatotron 3 has the same architecture as the one used in translatotron 2 [18]. Equation (8) describes the encoding process.

$$\mathcal{E}(S^{in}) = [\mathcal{E}_m(S^{in}), \mathcal{E}_o(S^{in})] \quad (8)$$

With S^{in} the source or the target language. $\mathcal{E}_m(S^{in})$ the first half of the output is trained to a MUSE, it uses the MUSE loss. $\mathcal{E}_o(S^{in})$ does not need the MUSE loss to be updated. The decoder is described in equation (9).

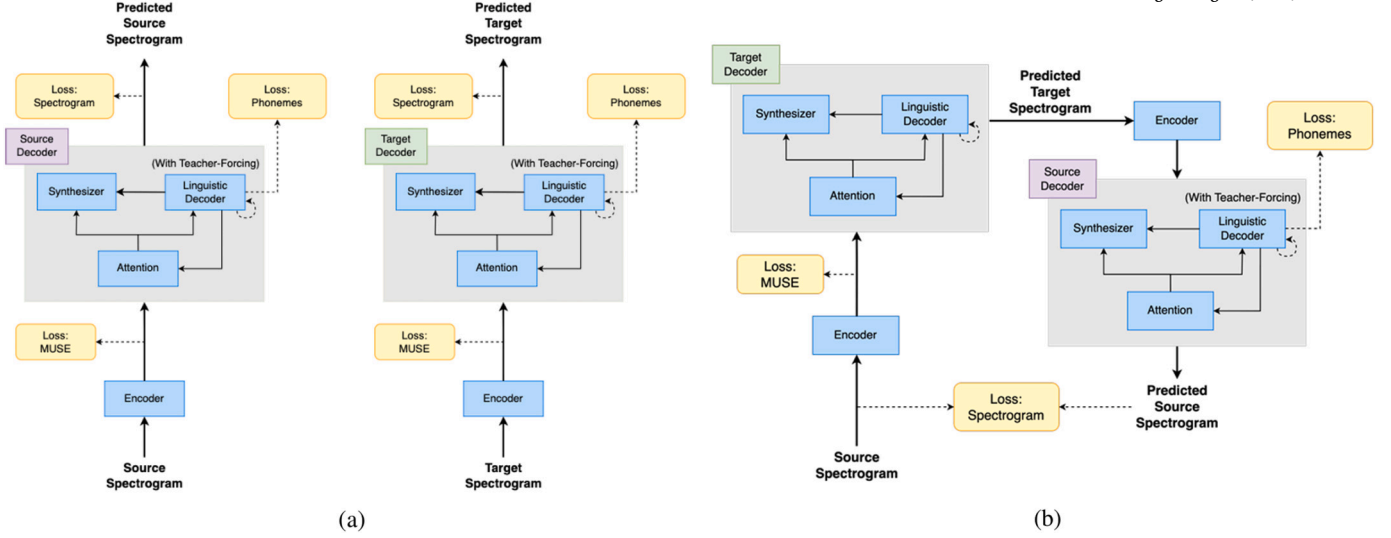


Fig. 3. Translatotron 3 Architecture and Training Process [25].

Table 2
Speech-to-Speech Translation Datasets.

Dataset	Description	Languages	Size
Fisher and Callhome Spanish-English [38]	Spanish to English translation Using a Speech to Speech model.	Spanish, English	Over 100 hours
GlobalPhone [40]	Multilingual speech dataset adaptable for S2ST.	20+ languages	Varies by language
VoxPopuli [37]	Multilingual dataset from European Parliament events.	15 European languages	Over 1,800 hours

$$S^{out} = D^{out}(\mathcal{E}(S^{in})), \quad (9)$$

Where S^{in} and S^{out} are the input and output of the spectrogram sequences.

The training process of the model focused on the reduction of the following losses:

- MUSE loss is described by equation (10).

$$\mathcal{L}_{MUSE}(S^{in}) = \frac{1}{n} \sum_{i=1}^n \|\mathcal{E}(S^{in})_i - E_i\|_2^2, \quad (10)$$

- Reconstruction loss in equation (11).

$$\begin{aligned} \mathcal{L}_{recon} = & \mathcal{L}_{spec}^{src}(S^s, S^s) + \mathcal{L}_{dur}^{src} + \mathcal{L}_{phn}^{src}(\tilde{P}^s, P^s) + \mathcal{L}_{spec}^{tgt}(S^t, S^t) + \mathcal{L}_{dur}^{tgt} \\ & + \mathcal{L}_{phn}^{tgt}(\tilde{P}^t, P^t), \end{aligned} \quad (11)$$

- A back translation loss in equation (12).

$$\mathcal{L}_{back-translation} = \mathcal{L}_{back-translation}^{src2tgt} + \mathcal{L}_{back-translation}^{tgt2src} \quad (12)$$

- The overall loss summarizing all loss is presented in equation (13).

$$\mathcal{L}_{BT-phase} = \mathcal{L}_{back-translation} + \mathcal{L}_{recon-phase}. \quad (13)$$

Experimental results in Spanish and English show that translatotron 3 outperformed the baseline cascade model with a margin of +18.4 BLEU, compared to the supervised approach that needs real data pairs that are sometimes unavailable. The unsupervised Nature of Translatotron 3 helps solve the issue by retaining para/non-linguistic elements such as pauses, speaking rates, and speaker identity. Fig. 3 describes the architecture and training process of translatotron 3.

6. Existing corpora for direct speech-2-speech translation

S2ST corpora available online are mainly used for the cascade approach. Very few datasets or corpora are specifically designed or compatible with direct S2ST. Table 2 shows the commonly used corpora for S2ST. When considering African languages, there is no corpus specifically designed for direct S2ST of African languages.

7. Comparative study

The comparative study of the three versions of the Translatotron model is summarized in the Table 3.

The first line of Table 3 lists all the models and the features used for the comparison. On line 2, the feature considered is the architecture with Basic Seq to Seq for Translatotron 1, improved Seq to Seq for Translatotron 2, and Advanced Transformer for Translatotron 3. On the last line, the feature considered is the sampling rate: with 16 kHz for translatotron 1 and 2, and a range of values from 16 kHz to 48 kHz for translatotron 3. This table clearly shows that Translatotron 3 is the best option when designing a sequence-to-sequence S2ST model. When designing an English-to-Yoruba translation model with a health environment as a context, we are of the opinion that Translatotron 3 would be the best option.

Fig. 4 shows the evolution of the speech-to-speech translation models. The tree shows that everything started in 1990 with the cascade models. The first sequence-to-sequence model appeared between 2012 and 2014. The latest version of the speech-to-speech translation, the Translatotron 3 model, was released in 2024.

This comparison section will be completed by introducing other S2ST models that were recently developed. The following table presents some newly designed models and compares them with all the translatotron versions. The comparison focused on the quality of the translation, naturalness & voice, and the benchmarks used to evaluate each model.

Table 3
Comparison of Translatotron 1, Translatotron 2, and Translatotron 3.

Features	Translatotron 1	Translatotron 2	Translatotron 3
Architecture	Basic seq2seq	Improved seq2seq	Advanced transformer
Direct Speech-to-Speech	Yes	Yes	Yes
Back-Translation	No	Yes	Yes
Reconstruction Loss	No	Yes	Yes
Phoneme-Level Supervision	No	Partial	Full
Speaker Adaptation	Limited	Improved	Advanced
Latency	High	Medium	Low
Translation Quality	Moderate	Good	Excellent
Training Data Size	100 hours	500 hours	1000+ hours
Data Diversity	Limited (single domain)	Moderate (multiple domains)	High (cross-domain)
Data Augmentation	None	Basic	Advanced
Parallel Data Required	Yes	Yes	Reduced (semi-supervised)
Sampling Rate	16 khz	16 khz	16 khz, 24 khz, 48 Khz
Strengths & Weaknesses	Simple, easy to train Low quality, high latency, weak speaker adaptation	Better translation, improved training Needs large parallel data, moderate latency	Excellent quality, low latency, strong speaker preservation Requires massive data & compute

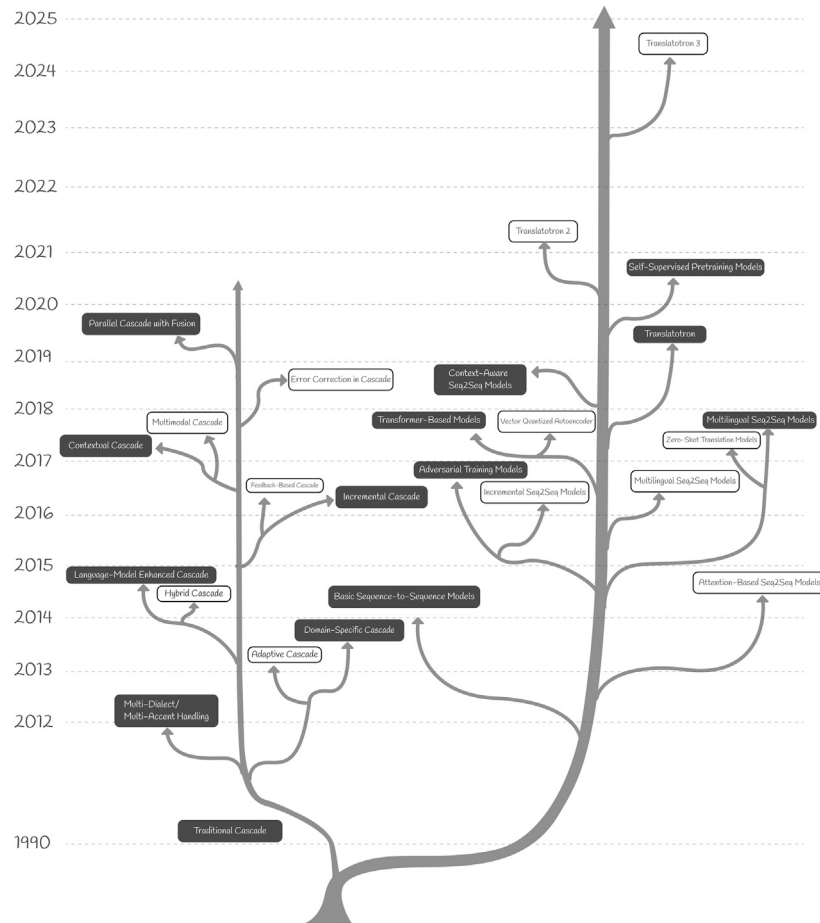


Fig. 4. Speech-to-Speech translation Evolution tree.

In Table 4, we can see that Translatotron 3 is one of the S2ST models that provide results comparable to Cascade. We also noticed that there are other models, such as Hibiki, that were introduced with new metrics ASR-BLEU, which is specific to S2ST models. The models presented in the table can also be differentiated using metrics such as computational efficiency, scalability, and multilingual support. All the Translatotron models, UT, S2TU, and SpeechPainter, are bilingual models, but SeamlessM4T and Unity are multilingual models. In terms of efficiency, Translatotron 1 has the lowest efficiency; the other models have a low-moderate or moderate efficiency, except UT, which has a high efficiency. Translatotron 1 has limited scalability; the scalability

was improved with the newer models, ranging from moderate to high scalability with Translatotron 2 and 3. Other models, like SpeechPainter, have low scalability, and SeamlessM4T has a very high scalability.

8. Deployment roadmap for translatotron

The deployment of Translatotron for low-resource languages like Yoruba should use the following steps:

Data acquisition and preparation

Since Yoruba is a low-resource language, we mostly rely on the community for data collection. This process will involve interviewing people

Table 4
Comparison of recent S2ST models.

Model	Translation Quality	Naturalness & Voice	Benchmarks
Translatotron v1 (2019)	BLEU $\approx$ 14–20 (Fisher/Callhome)	MOS $\approx$ 3.7–4.0	Fisher/Callhome
Translatotron 2 (2022)	+15.5 BLEU over v1; COMET $\approx$ cascades	MOS $\approx$ 4.0; strong Sim-MOS	CVSS, CoVoST, Fisher
Translatotron 3 (2023)	BLEU $\approx$ 14–18 (unsupervised CVSS)	MOS $\approx$ 3.8–4.0	CVSS (unsupervised)
UnitY (2022 [17])	Higher BLEU than TT2; better COMET	MOS $\approx$ 4.0	Fisher/Callhome, CVSS
DASpeech (2023) [13]	BLEU $\geq$ TT2; strong BLASER similarity	MOS $\approx$ 4.1; Sim-MOS $\approx$ 0.90	Fisher, CVSS
TranSpeech (2023) [12]	BLEU $\approx$ 18–20 (CVSS); strong chrF/COMET	MOS $\approx$ 4.0	CVSS
ComSpeech (2024) [14]	Highest BLEU on CVSS/Fisher; strong COMET	MOS $\approx$ 4.2; Sim-MOS $\approx$ 0.92	CVSS, Fisher
StyleS2ST (2023) [32]	BLEU comparable to TT2; strong BLASER style	More expressive MOS; style-aware Sim-MOS	CVSS + custom
StreamSpeech (2024) [14]	BLEU $\approx$ TT2; better COMET	MOS $\approx$ 4.0	CVSS, Fisher, IWSLT
SimulTron (2024) [2]	BLEU $>$ TT2; IWSLT SimulS2ST	MOS $\approx$ 3.9–4.0	IWSLT shared tasks
Hibiki (2025) [22]	ASR-BLEU $\approx$ 30.5 (Fr $\rightarrow$ En)	MOS $\approx$ 3.73 vs 4.12 (human)	Custom (Fr–En)
S2ST-Omni (2025) [27]	Competitive BLEU/COMET; strong BLASER	MOS $>$ 4.0	CVSS, Fisher, IWSLT
SeamlessM4T (2023) [3]	BLEU $\approx$ 20–25 (CVSS); high COMET	MOS $\approx$ 3.8–4.0; weaker in low-res.	CVSS, Fisher, IWSLT

to create a database of transcriptions from Yoruba to English, along with the associated audio. During that process, augmentation will be used to increase the size of the dataset. We may also utilize the existing Yoruba corpus. For the early version of the model implementation, the focus will be on the medical domain.

Quality assurance

The collection of the data will have to be validated by a linguist to make sure it satisfies language requirements. All transcriptions will have to go through a native speaker validation and a domain expert.

Model deployment

The deployment of the model required optimizing it for low-resource devices like cellphones. A cloud deployment will also be used for a web-based application.

Ethical and practical perspectives

The deployment of the proposed direct speech-to-speech model will require a large volume of voice data that should be anonymized and secured. The training options that could be considered are federated learning or on-device learning to avoid centralizing raw data. It might also be necessary to consider regulatory compliance, such as GDPR and HIPAA. Bias mitigation is important because speech-to-speech translation models are vulnerable to gender, accent, and dialect bias, which might result in poor translation accuracy. To mitigate bias, it is important to make sure the training data includes diverse accents, use fairness-aware training, and adversarial debiasing. The sensitive nature of speech-to-speech translation might require human-in-the-loop validation; hence, a human reviewer will have to validate a sensitive translation.

9. Theoretical and practical implications of this review

In the previous sections, the design and comparison of the three different attempts of direct speech-to-speech translation using Translatotron were presented. Translatotron significantly contributes to the body of knowledge of speech-to-speech translation without intermediate text. It also shows that a speech-to-speech model can capture speech nuances such as intonation and prosody. Translatotron can also contribute to the preservation of speaker voice characteristics during the translation. This particular property is the contribution of translatotron to the theory of speaker identity modeling in multilingual contexts [29], and hence advancing the understanding of voice characteristics transfer across languages. The current literature shows the ability of Translatotron to outperform the traditional translation methods. In the context of low-resource languages, Translatotron, just like many other machine learning translation-based models, can significantly contribute to the preservation of languages as presented in [15].

10. Limitations and open challenges

Translatotron systems have evolved from version 1 to version 3. During these evolutions, many limitations have been addressed. Translatotron 1 was weaker than the cascade model with a lower BLEU. Part of the limitation was solved with the newer versions. Robustness issues like long pauses, babbling, remain even with newer versions [19,18]. There is a constant need to align target speech. Emotion prosody and speaker rate preservation remain inconsistent [19,18]. Latency is also an issue with translatotron models, with a value that can be about 1 second for the early versions [25], but significantly improves with version 3. Voice spoofing was possible with version 1; the issue was addressed with the early versions by implementing restriction measures, since detection mechanisms are not yet available. Hence, the model could be used to copy a person's voice and use it for malicious purposes, since speaker voice preservation was integrated in the model architecture from version 1. Zero-shot with translatotron remains an open challenge [19,18,25]. The newer version of Translatotron uses unsupervised training but is applicable to monolingual data. The deployment of the translatotron model requires ethical consent and the setup of a synthetic detection system to avoid voice cloning. As a future direction, it will be interesting to develop a hybrid architecture that will be able to benefit from the cascade model and the translatotron S2ST Architecture. Improve style transfer by separating encoders for linguistic or acoustic features.

11. Conclusion

This literature review presents past and recent developments in the field of S2ST. A particular focus was on direct translation models: Translatotron. Translatotron evolved from a simple proof of concept to an unsupervised model that can produce results comparable to the baseline and even better than the unsupervised baseline model. Translatotron-based S2ST systems remain difficult to use with low-resource languages, and they require a substantial data curation process for the targeted parallel languages. Even if the evolution of translatotron addresses voice preservation and other security issues, such as latency and robustness, it remains a challenge (latency greater than 1s for some versions). Using other metrics, such as MOS (mean opinion score), provides more insight into how the model behaves in real life. Translatotron 3 leads the field of S2ST models with a BLEU of 28.5, a MOS of 4.1, and speaker voice preservation (0.86 cosine similarity), which shows that the qualities of translation and speech generated by the model are good. On the other hand, Simultron and Transpeech lead in the low-latency area, with 320 ms for Simultron. Translatotron 3 achieves a reduced training data need because of the unsupervised back translation (only available for monolingual models). To conduct a complete S2ST models comparison, it is necessary to use a common baseline. This work is a preliminary step toward designing an S2ST model that can translate between English and Yoruba in the medical environment. The choice of Translatotron 3 to design a translation model between English and Yoruba is based on the

properties displayed and the fact that not all variants of Yoruba are actually formalized and can be easily transcribed. Translatotron 3 is a large neural network that requires a significant amount of computation and storage resources. Cloud-based deployment is a valid option, but a reliable internet connection is necessary. For on-premises deployment, a significant investment in infrastructure is needed, or the model will require some optimizations that might reduce its capabilities. To use Translatotron 3 for a translation from Yoruba to English in the medical domain, it will be necessary to fine-tune an existing English-centric model, address the resource issues, incorporate medical knowledge, and speaker prosody. The hardware deployment requirement for Translatotron 3 can be a computer with 8Gb RAM, with a capable CPU or GPU, and a storage of 100Gb. Possible future directions for S2ST models include the capability to integrate visual, contextual, and textual modalities with the speech to enhance the translation by improving the comprehension and naturalness of the outputs. Combining these modalities also enables a more robust disambiguation. Merging S2ST models with reinforcement learning is another direction that could be considered. Reinforcement learning will help with the refinement of the translation. Reinforcement learning with human feedback is a viable option for reinforcement learning integration. Using LLMs such as GPT could enable semantically aware translation that is able to enable context-aware, cultural nuances, and discourse understanding.

CRedit authorship contribution statement

Jules R. Kala: Supervision, Resources, Formal analysis, Conceptualization. **Emmanuel Adetiba:** Validation, Resources, Funding acquisition, Conceptualization. **Abdultaofeek Abayomi:** Validation, Methodology, Formal analysis. **Oluwatobi E. Dare** and **Ayodele H. Ifijeh** worked on editing and proof reading of the article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jules Raymond Kala reports financial support was provided by Google Research. Emmanuel Adetiba reports a relationship with Google Garo-Covenant University that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors are grateful to Google for funding this work through the Google Academic Research Award to EA. Covenant Applied Informatics and Communication Africa Centre of Excellence (CApIC-ACE), Covenant University, is also acknowledged for the FEDGEN HPC infrastructure through a World Bank ACE Impact grant administered by the Nigerian National University Commission, and for also providing support for funding of this publication.

Data availability

Data will be made available on request.

References

- [1] E. Abraham, A. Nayak, A. Iqbal, Real-time translation of Indian sign language using lstm, in: 2019 Global Conference for Advancement in Technology (GCAT), IEEE, 2019, pp. 1–5.
- [2] A. Agranovich, E. Nachmani, O. Rybakov, Y. Ding, Y. Jia, N. Bar, H. Zen, M.T. Ramanovich, Simultron: on-device simultaneous speech to speech translation, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2025, pp. 1–5.
- [3] L. Barrault, Y.A. Chung, M.C. Meglioli, D. Dale, N. Dong, P.A. Duquenne, H. Elshahar, H. Gong, K. Heffernan, J. Hoffman, et al., Seamlessm4t: massively multilingual & multimodal machine translation, arXiv preprint arXiv:2308.11596, 2023.
- [4] L. Bentivogli, M. Cettolo, M. Gaido, A. Karakanta, A. Martinelli, M. Negri, M. Turchi, Cascade versus direct speech translation: do the differences still make a difference?, arXiv preprint arXiv:2106.01045, 2021.
- [5] P.J. Chen, K. Tran, Y. Yang, J. Du, J. Kao, Y.A. Chung, P. Tomasello, P.A. Duquenne, H. Schwenk, H. Gong, H. Inaguma, Speech-to-speech translation for a real-world unwritten language, arXiv preprint arXiv:2211.06474, 2022.
- [6] P.J. Chen, K. Tran, Y. Yang, J. Du, J. Kao, Y.A. Chung, P. Tomasello, P.A. Duquenne, H. Schwenk, H. Gong, et al., Speech-to-speech translation for a real-world unwritten language, arXiv preprint arXiv:2211.06474, 2022.
- [7] C. Cieri, M. Maxwell, S. Strassel, J. Tracey, Selection criteria for low resource language programs, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 4543–4549.
- [8] R.O. Collin, Ethnologue. Ethnopolitics, Phys. Rev. E 425 (2010) 32.
- [9] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, H. Jégou, Word translation without parallel data, arXiv preprint arXiv:1710.04087, 2017.
- [10] D. Crystal, R. Quirk, Systems of prosodic and paralinguistic features in English, Walter de Gruyter GmbH & Co KG 20 (2021) 22.
- [11] T. Dutoit, An Introduction to Text-to-Speech Synthesis, vol. 3, Springer Science & Business Media, 1997.
- [12] Q. Fang, S. Zhang, Z. Ma, M. Zhang, Y. Feng, Can we achieve high-quality direct speech-to-speech translation without parallel speech data?, arXiv preprint arXiv:2406.07289, 2024.
- [13] Q. Fang, Y. Zhou, Y. Feng, Daspeech: directed acyclic transformer for fast and high-quality speech-to-speech translation, Adv. Neural Inf. Process. Syst. 36 (2023) 72604–72623.
- [14] S. Guo, X. Li, M. Liu, W. Chen, Y. Feng, Streamuni: achieving streaming speech translation with a unified large speech-language model, arXiv preprint arXiv:2507.07803, 2025.
- [15] C.J. Horna-Saldaña, J.E. Perez Perez, M.L. Toro Galeano, Artificial intelligence in the preservation of native languages and bridging the information access gap for indigenous peoples, J. Enabl. Technol. (2025).
- [16] R. Huang, J. Liu, H. Liu, Y. Ren, L. Zhang, J. He, Z. Zhao, Transpeech: speech-to-speech translation with bilateral perturbation, arXiv preprint arXiv:2205.12523, 2022.
- [17] H. Inaguma, S. Popuri, I. Kulikov, P.J. Chen, C. Wang, Y.A. Chung, Y. Tang, A. Lee, S. Watanabe, J. Pino, Unity: two-pass direct speech-to-speech translation with discrete units, arXiv preprint arXiv:2212.08055, 2022.
- [18] Y. Jia, M.T. Ramanovich, T. Remez, R. Pomerantz, Translatotron 2: high-quality direct speech-to-speech translation with voice preservation, in: International Conference on Machine Learning, 2022, pp. 10120–10134.
- [19] Y. Jia, R.J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, Y. Wu, Direct speech-to-speech translation with a sequence-to-sequence model, arXiv preprint arXiv:1904.06037, 2019.
- [20] M. Johnson, M. Schuster, Q.V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, Google's multilingual neural machine translation system: enabling zero-shot translation, Trans. Assoc. Comput. Linguist. 5 (2017) 339–351.
- [21] P. Koehn, Statistical Machine Translation, Cambridge University Press, 2009.
- [22] T. Labiaus, L. Mazaré, E. Grave, P. Pérez, A. Défossez, N. Zeghidour, High-fidelity simultaneous speech-to-speech translation, arXiv preprint arXiv:2502.03382, 2025.
- [23] R. Lakshminarayanan, A. Shaik, A. Balasundaram, Automated speech therapy through personalized pronunciation correction using reinforcement learning and large language models, Results Eng. 25 (2025) 103943.
- [24] A. Lee, P.J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang, J. Pino, Direct speech-to-speech translation with discrete units, arXiv preprint arXiv:2107.05604, 2021.
- [25] E. Nachmani, A. Levkovitch, Y. Ding, C. Asawaroengchai, H. Zen, M.T. Ramanovich, Translatotron 3: speech-to-speech translation with monolingual data, in: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 10686–10690.
- [26] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J.S. Zhang, S. Yamamoto, E. Sumita, S. Yamamoto, The atr multilingual speech-to-speech translation system, in: IEEE Transactions on Audio, Speech, and Language Processing, 2006, pp. 365–376.
- [27] Y. Pan, Y. Yang, Y. Hu, J. Ye, X. Zhang, H. Zhou, L. Ma, J. Zhao, S2st-omni: an efficient and scalable multilingual speech-to-speech translation framework via seamlessly speech-text alignment and streaming speech decoder, arXiv preprint arXiv:2506.11160, 2025.
- [28] A.M. Rinaldi, C. Russo, C. Tommasino, Automatic image captioning combining natural language processing and deep neural networks, Results Eng. 18 (2023) 101107.
- [29] S. Shaumiwat, The virtual multilingualism phenomenon, the impact of social media on language structure and global linguistic identity, J. Amer. Inst. 2 (2025) 318–328.
- [30] J. Shi, Y. Tang, A. Lee, H. Inaguma, C. Wang, J. Pino, S. Watanabe, Enhancing speech-to-speech translation with multiple tts targets, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.
- [31] J. Son, B. Kim, Benchmark graphs for testing community detection algorithms, Information 14 (2023) 574.

- [32] K. Song, Y. Ren, Y. Lei, C. Wang, K. Wei, L. Xie, X. Yin, Z. Ma, Styles2st: zero-shot style transfer for direct speech-to-speech translation, arXiv preprint arXiv:2305.17732, 2023.
- [33] F. Stahlberg, Neural machine translation: a review, J. Artif. Intell. Res. 69 (2020) 343–418.
- [34] C. Taylor, The multimodal approach in audiovisual translation, Target 28 (2016) 222–236.
- [35] E. Vidal, Finite-state speech-to-speech translation, in: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997, pp. 111–114.
- [36] L.C. Vila, C. Escolano, J.A. Fonollosa, M.R. Costa-Jussà, End-to-end speech translation with the transformer, in: IberSPEECH, 2018, pp. 60–63.
- [37] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, E. Dupoux, Voxpopuli: a large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation, arXiv preprint arXiv:2101.00390, 2021.
- [38] R.J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, Z. Chen, Sequence-to-sequence models can directly translate foreign speech, arXiv preprint arXiv:1703.08581, 2017.
- [39] D. Yu, L. Deng, Automatic Speech Recognition., vol. 1, Springer, 2016.
- [40] B. Ziółko, P. Żelasko, I. Gawlik, T. Pędzimaż, T. Jadczyk, An application for building a Polish telephone speech corpus, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.